More than Meets the Eye:

The Conceptual Essence of Intrinsic Memorability

Lorin Sweeney, B.Sc.

Supervised by Prof. Alan Smeaton & Dr. Graham Healy

DCU

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

A thesis presented for the degree of Doctor of Philosophy

School of Computing

Dublin City University

September 2023

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: ______________ ID No: 15707415 Date: 05/07/2023

Dedication

To Katie-Rose,

Through mirrored halls of memory and the winding maze of my mind, you have been my constant, a single, unwavering flame in a lantern of swirling shadows. As the echo of your support resounds through the canyons of memory explored within these pages, to you, my steadfast torchbearer, this work bows in dedication.

Acknowledgements

First and foremost, I would like to express heartfelt gratitude to my exceptional supervisors—Professor Alan Smeaton and Dr. Graham Healy—for their invaluable guidance, support, and mentorship. It was Alan who first identified the resonance between my innate curiosity, broad scientific interests, and technical skills, and the captivating world of memorability research. His unique ability to navigate the broader scholarly discourse, while consistently ensuring my research remained on a meaningful and productive track, has been invaluable. Graham complemented this journey with his exceptional depth of neuroscientific knowledge and his meticulous attention to methodological rigor. His hands-on guidance in navigating the technical challenges associated with neurophysiological studies has been of great value. His unwavering dedication to precision and detail served as a perfect counterpoint to Alan's broader perspective, fostering a comprehensive and balanced approach to my research. Their collective wisdom, imbued with their distinct yet complementary strengths, cultivated an intellectually nurturing environment that fostered my growth as a scholar. Their steadfast belief in my abilities, coupled with their patient mentorship, shaped not only this thesis but also influenced my personal evolution as a researcher. My gratitude towards them extends beyond mere words. The influence of their mentorship reverberates beyond the confines of this work and continues to inspire my intellectual (financial support from Science Foundation Ireland [12/RC/2289_P2] at Insight the SFI Research Centre for Data Analytics at DCU).

I am deeply thankful to my family, who provided a nurturing and supportive backdrop throughout this journey. Their unwavering faith in me and constant en-

courragement became a source of quiet strength, helping me to navigate the often tumultuous seas of academic research. My parents have always been my grounding force, their love and support acting as the foundation upon which I've been able to construct my scholarly pursuits. They have championed my endeavors, comforted my uncertainties, and celebrated my victories. I owe a significant part of my perseverance and resilience to their ceaseless encouragement.

06/09/2023

List of Publications

- R. S. Kiziltepe, M. G. Constantin, C. H. Demarty, *et al.*, "Overview of the mediaeval 2021 predicting media memorability task," in *MediaEval Multimedia Benchmark Workshop Working Notes*, 2021.
- L. Sweeney, M. G. Constantin, C. H. Demarty, *et al.*, "Overview of the mediaeval 2022 predicting video memorability task," in *MediaEval Multimedia Benchmark Workshop Working Notes*, 2022.
- L. Sweeney, G. Healy, and A. F. Smeaton, "The influence of audio on video memorability with an audio gestalt regulated video memorability system," in *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, IEEE, Jun. 2021, pp. 1–6.
- ——, "Diffusing surrogate dreams of video scenes to predict video memorability," in *MediaEval Multimedia Benchmark Workshop Working Notes*, 2023.
- A. García Seco de Herrera, M. G. Constantin, C. H. Demarty, *et al.*, "Experiences from the mediaeval predicting media memorability task," 2022.
- S. Cummins, L. Sweeney, and A. Smeaton, "Analysing the memorability of a procedural crime-drama tv series, CSI," in *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*, Sep. 2022, pp. 174–180.
- R. S. Kiziltepe, L. Sweeney, M. G. Constantin, *et al.*, "An annotated video dataset for computing video memorability," *Data in Brief*, vol. 39, p. 107 671, 2021.

- L. Sweeney, G. Healy, and A. F. Smeaton, "Memories in the making: Predicting video memorability with encoding phase EEG," in *2023 International Conference on Content-Based Multimedia Indexing (CBMI)*, Accepted, yet to be published, IEEE, Jun. 2023.
- V. Mudgal, Q. Wang, L. Sweeney, G. Healy, and A. F. Smeaton, "Using saliency and cropping to improve video memorability," in *2024 International Conference on Multimedia Modelling (MMM)*, Accepted, yet to be published, Springer, Feb. 2024.
- L. Sweeney, G. Healy, and A. F. Smeaton, "The conceptual essence of intrinsic memorability," In Progress, Feb. 2024.

List of Abbreviations

AMT	Amazon Mechanical Turk
ANT	Attention Network Test
ASGD	Averaged Stochastic Gradient Descent
ATL	Anterior Temporal Lobe
AUC	Area Under the Curve
AUG	Codon for Methionine or start codon in biology, could also be short for August
AWD	Average-Stochastic Gradient Descent Weight-Dropped
BA	Balanced Accuracy
BLIP	Bootstrapping Language-Image Pre-training
BRR	Bayesian Ridge Regression
CLIP	Contrastive Language–Image Pretraining
CNN	Convolutional Neural Network
CNP	Cortical Neural Pattern
CNS	Central Nervous System
CSS	Caption Specificity Score
DLPFC	Dorsolateral Pre-frontal Cortex

EEG	Electroencephalography
ERD	Event-Related Desynchronisation
ERP	Event-Related Potential
ERS	Event-Related Synchronisation
ERSP	Event-Related Spectral Perturbation
EVC	Extrastriate Visual Cortex
FFA	Fusiform Face Area
FIR	Finite Impulse Response
FPR	False Positive Rate
GCC	Google Conceptual Captions
HNP	Hippocampal Neural Pattern
ICA	Independent Component Analysis
IDF	Inverse Document Frequency
IFG	Inferior Frontal Gyrus
ISI	Inter-Stimulus Interval
IT	Inferior Temporal Cortex
LEC	Lateral Entorhinal Cortex
LOC	Lateral Occipital Complex
LPC	Late Positive Component
LSTM	Long Short-Term Memory
LTP	Long-Term Potentiation

MAR	Memory As Reinstatement
MEC	Medial Entorhinal Cortex
MED	Multimodal mixture of Encoder-Decoder
MEG	Magnetoencephalography
MTL	Medial Temporal Lobe
NEC	Named Entity Count
NER	Named Entity Recognition
NLP	Natural Language Processing
PANNs	Pre-trained Audio Neural Networks
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PFC	Pre-frontal Cortex
PHC	Parahippocampal Cortex
PPA	Parahippocampal Place Area
PRC	Perirhinal Cortex
ROC	Receiver Operating Characteristic
RSA	Representational Similarity Analysis
RSCV	Randomised Search Cross-Validation
SD	Subject Dependent
SI	Subject Independent
SID	Subject Identification

SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
STM	Short-Term Memory
TCM	Temporal Context Model
TF	Term Frequency
TPR	True Positive Rate
ULMFiT	Universal Language Model Fine-tuning
VAE	Variational Autoencoder
ViT	Vision Transformer
VVIQ	Vividness of Visual Imagery Questionnaire

Contents

List of Figures

List of Tables

More than Meets the Eye: The Conceptual Essence of Intrinsic Memorability

Lorin Sweeney

Abstract

In a world where sensory threads weave an endless tapestry of multi-modal data, the human brain stands as the masterful weaver of meaning. As we wade through this tempest of input, our brain spins these threads into an intelligible internal representation and holds on tight to what it deems important. But what, exactly, makes certain threads more important than others? And how can we predict their significance?

Memorability is the tensile strength of the threads that tie us to the world. It is a proxy for human importance, indicating which threads the human brain will curate and retain with exceptional fidelity. This research investigates these multisensory threads by exploring the influence of audio, visual, and textual modalities on predicting video memorability, and how the interplay between them can influence the overall memorability of a given piece of content. The findings suggest that, while visual data may dominate our sensory experience, it is the underlying conceptual essence that truly holds the key to memorability. This thesis leverages state-of-the-art image synthesis techniques to distill and examine this essence, creating surrogate dreams of video scenes to facilitate the disentanglement of conceptual and perceptual elements of memorability. The work also leverages human EEG data to explore the possibility of a moment of memorability—a moment of encoding that corresponds to a remembering moment—which we expect to exist due to the temporal nature of the world and the natural encoding limits of our brains. The previously murky relationship between the two core means of remembrance—recognition and recall—are reconciled by conducting a novel video memorability drawing task.

The research sheds new light on the nature of multi-modal memorability, providing a deeper understanding of how our brain processes and retains information in a complex sensory world. By uncovering the conceptual essence that lies at the heart of memorability, it opens up new avenues for predicting and curating more meaningful media content, and ultimately deepen our connection to the world around us.

Chapter 1

Introduction

"Memory is the treasury and guardian of all things."

— Cicero

Memories are the warp and weft of the rich tapestry we call life. They provide continuity to our existence, knitting together the disparate threads of sensory experiences into a cohesive narrative of self. They are the keepers of our continuity of self—without them, the very fabric of our being would fray. Each memory—a burst of laughter, a moment of sorrow, an encounter with wonder—acts as a unique thread, contributing to the intricate design of our personal and collective histories. Yet, we scarcely have any influence on what we will ultimately remember or forget. The brain presides over the mechanisms of our memory from an opaque glass office, exercising sole editorial influence over its edifice. In fact, our odds of guessing what we will remember are not much better than chance [1]. This lack of meta-cognitive insight, which prevents us from diving into our unconscious undercurrent, is what motivates and brings meaning to our investigation into the concept of memorability.

Memorability is akin to the tensile strength of our sensory threads—it speaks to their endurance. It serves as a gauge of human importance, marking the data our brain elects to retain with utmost fidelity. As such, the study of memorability—generally known as the likelihood that something will be remembered or forgotten—offers us a novel lens through which to examine our cognitive processes, and to potentially predict the enduring impact of experiences. In an age where information is abundant, it is not the acquisition but the preservation of knowledge that challenges us. We are constantly immersed in a deluge of multi-modal data, vying for our attention and retention. Despite this sensory onslaught, our brains demonstrate a remarkable ability to select, store, and recall information. But what makes some experiences indelible and others ephemeral? What, if any, are the guiding principles of this cognitive curation? What factors determine the memorability of an experience? These are some of the crucial questions that my thesis endeavours to shed light on.

I explore the influence of audio, visual, and textual modalities on video memorability, and the interplay between them. I further probe into the temporal nature of our lived experiences and the inherent limitations of our cognitive architecture. I investigate the notion of a "moment of memorability"—a precise instant of encoding that maps onto a subsequent moment of recollection. Such moments, I posit, are integral to our understanding of how memories are formed, stored, and recalled. Additionally, I take a closer look at the interplay between recognition and recall, the two primary modes of memory retrieval. This is accomplished by undertaking a novel video memorability drawing task, which ultimately seeks to reconcile these two constructs' previously nebulous relationship. Through this exploration, I hope to illuminate the complex dynamics of memory formation and retrieval, providing new insights into what makes our experiences memorable. In the final phase of my exploration, I seek to understand whether it is the underlying conceptual essence of stimuli that holds the key to memorability, rather than the stimuli's visual characteristics. To this end, I employ advanced image synthesis techniques to distil this essence, creating, what I term, "surrogate dreams of video scenes". This allows me to tease apart the intertwined threads of conceptual and perceptual elements and examine their distinct influences on memorability. My study on memorability, therefore, extends beyond a mere exploration of cognitive processes. It also encompasses a broader perspective of our being—of how we experience, interpret, remember, and ultimately, make sense of our surroundings.

1.1 Research Objectives and Questions

The term memorability can cause confusion since its meaning changes depending on the context of its usage. There are many different modalities and measurement paradigms for measuring remembrance, however, none of them properly capture the functional essence of real-world memory and the diversity of its contexts.

My research hypothesis addresses this issue as follows:

> The intrinsic memorability of a given stimulus in a real-world context is a multi-faceted construct, influenced significantly by a dynamic and synergistic interplay of various sensory modalities. Among these modalities, the visual component, owing to our inherent cortical bias, holds a paramount position in communicating the degree of memorability. Remembrance is polymorphic process—it can take the form of flashing familiarity, detailed recollection, or a chimeric mixture, where the former sets the stage for the latter—suggesting a discernible relationship between recognition memorability and recall memorability. This process unfolds within the constraints of biological storage limitations, suggesting that specific moments of representational compression should exist, and give rise to parallel specific moments of remembering. Lastly, if the ultimate function of perception is to facilitate conceptual understanding/representation, and memorability is a measure of human information utility, then memorability is thus not merely a perceptual property, but a conceptual one, and can accordingly be distilled down to the underlying conceptual essence of a stimulus.

The overall research hypothesis can be broken down into narrower sub-hypotheses:

1. Hypothesis 1 (H1): Real-world memory—as we experience it navigating through our daily environment—is highly dynamic in nature due to the complex multi-sensory data we parse as a means to functioning in the world. Memorability

should be equally multi-modal, so a measurable interaction of influence between the modalities in multi-modal data should exist.

2. Hypothesis 2 (H2): We are visual cortex dominant creatures, so visual sensory data should exert the greatest influence on memorability.

3. Hypothesis 3 (H3): Given that a feeling of familiarity (recognition) typically precedes the recollection of details (recall), there should be a measurable relationship between recognition memorability and recall memorability.

4. Hypothesis 4 (H4): Due to the temporal nature of the world, and the natural encoding limits of our brains, a moment at which a compressed representation is assigned memorability—a "moment of memorability"—should exist.

5. Hypothesis 5 (H5): In light of the considerable cognitive psychological evidence highlighting the role of perception in creating mental representations of our environment, which subsequently direct our interactions and decisions in response to that environment, it follows that if memorability serves as an index of information utility, the memorability of a stimulus ought to be traceable to its foundational conceptual representation.

1.1.1 Research Questions

To evaluate the aforementioned hypotheses, my research will focus on addressing the following questions:

1. RQ1: Do each of the constituent modalities in a multi-modal medium, such as video, contribute equally to overall memorability, and how do they interact?

2. RQ2: Is recognition memorability a precursor to recall memorability, and are they correlated?

3. RQ3: Is there a moment of memorability—that is, is there an encoding moment which corresponds to a remembering moment—and can we predict it?

4. RQ4: Is perception merely a means to conceptual understanding, and can the intrinsic memorability of visual content accordingly be distilled to its underlying concept/meaning?

1.2 Thesis Structure

The thesis is organised into the following chapters:

- **Chapter 2** is organised into three broad sections: Fundamentals of Neuroscience, Memory, and Remembrance: The Act of Remembering. The first section, "Fundamentals of Neuroscience", provides a comprehensive foundation for understanding the neural mechanisms underlying memory. It includes an overview of the brain, neurons, their structures and functions, and a brief explanation of Electroencephalography (EEG), a critical tool for studying brain activity. The second section, "Memory", delves deeper into the subject, presenting a broad view of what constitutes memory, followed by an in-depth analysis of short-term memory, its neural mechanisms, associated brain systems, and the role of the medial temporal lobe and hippocampus in explicit memory. This section also addresses the complexities of episodic memory, investigating the processes involved in reinstating memories, the oscillations during encoding, the significance of context, and the subjectivity of time in memory formation. The final section, "Remembrance: The Act of Remembering", focuses on the concept of remembrance, differentiating between the two main forms of memory retrieval: recognition and recall, highlighting the different neural and cognitive processes involved, and outlining common methodologies used for measuring recognition and recall. The intent of the chapter is to provide a thorough understanding of memory, what it means to remember, and how memorability, the likelihood of something being subsequently remembered, can be measured and understood.

- **Chapter 3** focuses on memorability, defined generally, within the context of this thesis, and exploring its origins. It then explores the concept of memorability as an inherent attribute of images, investigating the specific properties that influence it and how these might translate into other cognitive contexts. This chapter also delves into the neurological underpinnings of visual memorability, discussing spatial representations and the mechanisms that drive this process in the brain. Furthermore, it extends the discussion of memorability beyond the visual, considering how textual and auditory stimuli also contribute to the formation of multi-sensory memories.

- **Chapter 4** presents an examination of how various sensory modalities, including visual, textual, and auditory, influence video memorability. Using the TRECVid 2019 and Memento10k datasets, the chapter discusses the construction and performance of modality-specific models. The chapter introduces the concept of "Audio Gestalt" and presents an audio gestalt regulated multimodal video memorability prediction framework, which is used to reveal the complex interplay of sensory inputs in memorability prediction.

- **Chapter 5** introduces a novel integration of machine learning with electroencephalography (EEG) in the study of memory mechanisms and memorability. The chapter starts by detailing the EEG data acquisition process, and proceeds by outlining a novel pilot study on video memorability using EEG data. This concept is later expanded upon in the chapter, with an exploration into how encoding phase EEG signals during video presentation can be used to predict individual recognition upon subsequent viewing. It is hypothesised that these neural signals recorded at specific moments can differ depending on whether an event is remembered or forgotten. The chapter also compares different methodological approaches, such as subject-independent and subject-dependent training, and single electrode versus composite electrode data, with a specific focus on theta band activity over the right temporal lobe. This chapter, therefore, expands the research horizons by investigating the interplay between EEG features and video memorability, fostering a vibrant cross-pollination of ideas and methods between the fields of neurophysiology and computational research.

- **Chapter 6** delves into the dichotomy between the two main means of remembrance—recognition and recall—setting the stage for their reconciliation. It introduces a unique drawing-based video recall experiment, detailing the experimental design and discussing the various methods of quantifying recall. These methods include semantic similarity, drawing-based measures, textual measures, and other diverse measures, offering a more complete picture of the relationship between recall and recognition memorability.

- **Chapter 7** questions the very nature of video memorability by exploring a radically new perspective—that memorability is a conceptual, rather than perceptual feature. It details how cutting-edge image synthesis models can be leveraged to perform visual abstraction and conceptual distillation. The chapter introduces “ConceptualDream,” a video memorability prediction framework, which combines synthetic image generation with state-of-the-art memorability prediction techniques.

- **Chapter 8** concludes this thesis, succinctly encapsulating the key findings of the investigation and the implications thereof. It provides an overview of the interplay between synthetic data, conceptual representation, and memorability prediction, thereby delivering a nuanced perspective on video memorability. Additionally, this chapter contains a section dedicated to potential future research directions, highlighting areas that could be explored to build upon the findings of this thesis and contribute to a more comprehensive understanding of the subject matter.

Chapter 2

Background

This chapter aims to provide a knowledge basis from which to understand the overall thesis. It is divided into three core topics, starting with a basic overview of the brain, neurons, their functions, and an introduction to electroencephalography (EEG). The next section delves into memory, examining short-term memory, its neural bases, the role of the medial temporal lobe (MTL) and hippocampus in explicit memory, and the intricacies of episodic memory including context and time subjectivity. The final section delves into remembrance—the act of remembering—distinguishing between recognition and recall, discussing their neural and cognitive processes, and outlining methods for measuring them.

Fundamentals of Neuroscience

"The Brain—is wider than the Sky—
For—put them side by side—
The one the other will contain
With ease—and You—beside—

The Brain is deeper than the sea—
For—hold them—Blue to Blue—
The one the other will absorb—
As Sponges—Buckets—do—"

— Emily Dickinson

2.1 The Brain: An Overview

Standing centre stage, serving as the epicenter of human experience, the brain orchestrates a symphony of thoughts, emotions, and actions. This intricate and multifaceted organ, composed of approximately 86 billion neurons, weaves a tapestry of interconnected networks and circuits, enabling the seamless integration of sensory input, decision-making, and motor output [2]. As the conductor of the central nervous system (CNS), the brain not only governs the most basic of survival functions but also the most complex and intricate cognitive processes, making it the crown jewel of biological evolution. Its exquisite architecture, like a great cathedral of thought, is comprised of seven major structures: the spinal cord, medulla oblongata, cerebellum, pons, midbrain, diencephalon, and cerebrum (Figure 2.1, A). Each component, though individual, harmonises in a concert of functions that underpin our very essence.

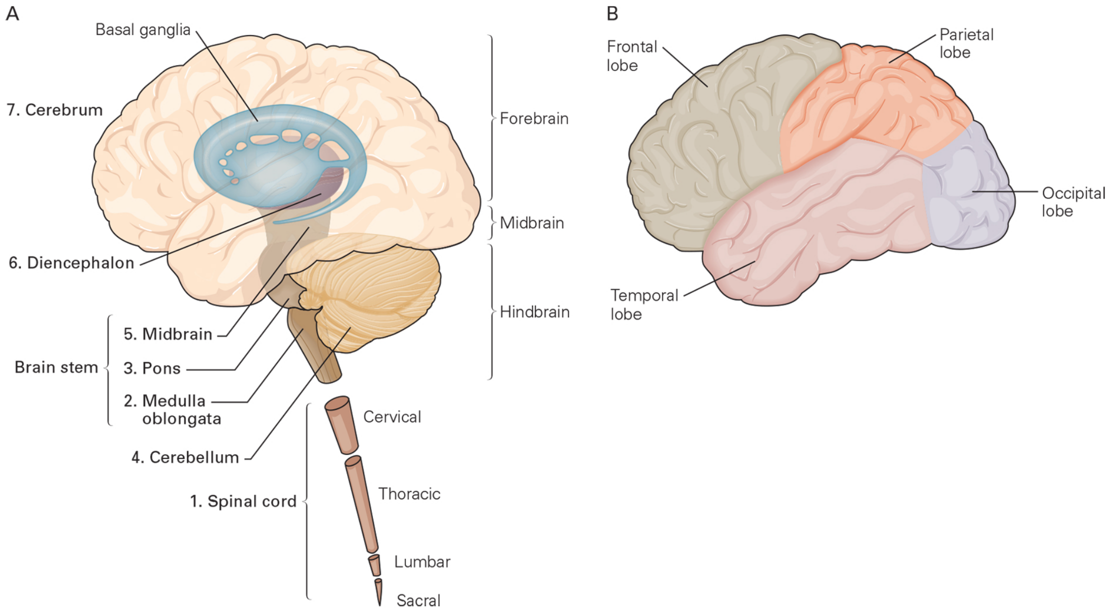


Figure 2.1: The central nervous system (CNS). **A.** Seven regions of CNS. **B.** Four lobes of the cerebral cortex. [3]

Spinal Cord

The spinal cord, a faithful courier, is the highway of neural information, ferrying messages between the brain and the body. It is the silent string player in the orchestra, continuously playing, seldom in the spotlight, but integral to the overall performance. It is the custodian of our reflexes, and in some cases our survival, as those rapid responses to environmental stimuli are crucial in make or break situations.

Medulla Oblongata

The medulla oblongata is the steady drummer who keeps time. It orchestrates the rhythm of life, regulating vital functions that flow beneath our river of consciousness, such as breathing, heart rate, and blood pressure [4]. It is the stalwart anchor of our existence. Nestled adjacent to the medulla oblongata lies the cerebellum, a diminutive yet essential player, the graceful ballet dancer of the ensemble. It is the master of motor control, fine-tuning movements while maintaining posture and balance, and orchestrating procedural learning with adept precision [3]. Above the medulla oblongata, rests the pons, the liaison officer of the brain. It serves as a bridge, transmitting information between the cerebellum and the cerebrum. The pons is the cellist in our neural orchestra, providing depth and resonance to the symphony of our cognition.

Midbrain

The midbrain, or mesencephalon, is a central structure that serves as a sentinel, vigilantly coordinating arousal, sleep, and sensory responses. It is a critical component in the orchestration of our responses to sensory stimuli, encompassing visual and auditory processing centers. Specifically, the superior colliculi are involved in the processing of visual stimuli, while the inferior colliculi are involved in the processing of auditory stimuli. Furthermore, the midbrain contains the reticular formation, a network of nuclei that play a key role in arousal and consciousness. The role of the midbrain can be analogised to the brass section of an orchestra; it is bold and commanding, signaling the presence of stimuli with a surge of neural activity. Ultimately, the midbrain is indispensable for the integration and coordination of sensory and motor pathways.

Diencephalon

The diencephalon comprises two primary structures: the thalamus and the hypothalamus. Both play pivotal roles in maintaining homeostasis, the state of internal physiological equilibrium. The thalamus functions as a central relay station, processing and directing a vast array of sensory information to the appropriate regions of the cerebrum, the part of the brain responsible for higher-order functions such as perception, motor functions, and cognition. This encompasses the transmission of information related to visual, auditory, somatosensory, and gustatory stimuli. Concurrently, the hypothalamus orchestrates a multitude of physiological functions. It controls thermoregulation by initiating sweating or shivering, controls hunger and satiety through interactions with hormones like ghrelin and leptin, and manages thirst and fluid balance by regulating the secretion of vasopressin. Additionally, the hypothalamus plays a vital role in regulating the circadian rhythm. Collectively, the thalamus and hypothalamus are essential for the coordination and regulation of numerous fundamental physiological processes.

Cerebrum

Finally, the cerebrum, occupying most of the cranial cavity, is the largest and most complex region of the human brain, responsible for a multitude of essential cognitive functions. It receives and interprets sensory information, converting a diverse array of inputs—visual, auditory, somatosensory—into coherent and integrated perceptions. It coordinates the intricate dance of motor control, shapes the sonnets of language, and weaves the tapestry of memory. It is within the cerebrum that we experience the world, make sense of our surroundings, and construct the narrative of our existence. It spans across two cerebral hemispheres—left and right—connected by the corpus callosum, a dense fiber network resembling a neural bridge, facilitating seamless communication between the two sides [5]. This region's outer layer, the cerebral cortex, is the canvas upon which the four primary territories—frontal, parietal, temporal, and occipital lobes (Figure 2.1, B)—paint their respective functions [6]. The cerebral cortex unveils itself as an undulating ocean, where gyri (ridges) crest like mighty waves, and sulci (grooves) plunge into the depths, creating an intricate origami-like structure that amplifies the surface area for neuronal connections (an evolutionary strategy to maximise the number of nerve cells in limited space) [3], [7].

Each lobe has a specialised set of functions. The frontal lobe—the cerebral architect—excels in executive functions and is largely concerned with short-term memory, laying the groundwork for reasoning, planning future actions, problem-solving, and decision-making, while simultaneously governing voluntary movement through the primary motor cortex [8]; the parietal lobe serve as the brain's cartographer, dedicated to processing somatosensory information, integrating sensory input from a myriad of sources, and constructing a cognitive map for spatial awareness and coordination [9]; the occipital lobe embodies the role of visual artist, painting vivid images by interpreting and making sense of the visual information streaming in from the eyes [10]; and the temporal lobe emerges as the brain's polymath, resonating with the melodies of the world's sounds, while orchestrating language comprehension, emotion, and memory formation [11], [12];

2.2 Neuronal Structures and Functions

The remarkable diversity of human behaviour is underpinned by an elaborate array of sensory receptors linked to an adaptive neural organ—the brain—that selectively processes sensory signals, identifying environmental events of interest. In essence, the brain dynamically filters and organises perception, either storing significant portions in memory for future reference, converting it into immediate behavioural responses, or ignoring it. Perhaps even more remarkable are the cells that underlie this complex operation—brain cells. A dualistic cellular composition lies at the heart of every species' brain: neurons and glia. Glia, aptly named after the Greek word for "glue", with their many forms, are tasked with vital responsibilities such as structural reinforcement, metabolic bolstering, insulation, and steering development. Nonetheless, neurons generally take center stage as the brain's most vital cells [3], and our focus will likewise be exclusively neuronal.

2.2.1 Neurons: The Building Bricks of the Nervous System

Amidst the intricate matrix of the brain, neurons—the trees of the nervous system's forest—orchestrate a vibrant cascade of electrical and chemical interactions, harmonising the cacophony of signals that mold our thoughts, emotions, and actions [3]. A typical neuron has four morphologically defined regions: (1) the cell body, (2) dendrites, (3) axon, and (4) presynaptic terminals (Figure 2.2).

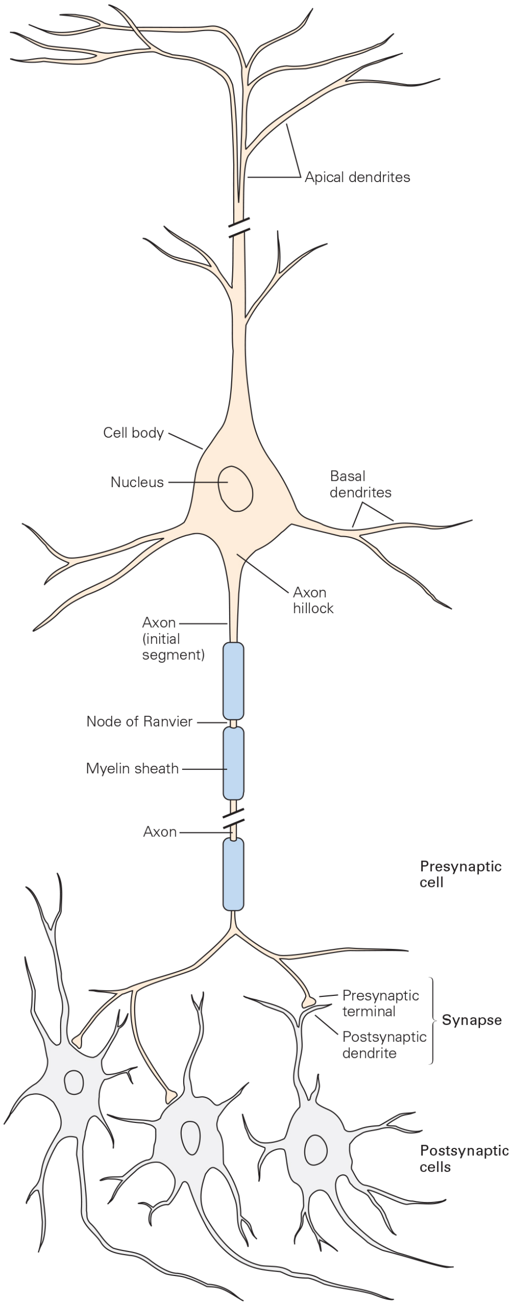


Figure 2.2: Structure of a neuron [3]

Picture the cell body (or soma) as the neuron's control centre, within which the nucleus resides, a secure vault safeguarding the cell's unique genetic blueprint, and alongside the nucleus rests the endoplasmic reticulum, the cellular assembly line that synthesises proteins—the sturdy bricks and mortar of the cell. From this central hub sprout multiple dendrites, short branches that function as receivers for signals from fellow neurons, and one long axon, a messenger carrying signals away from the cell body. Dendrites can be visualised as an intricate network of tiny fishing nets, each delicately cast into the cellular sea to snare incoming signals from other neurons. These dendritic nets capture and channel the information towards the cell body, playing an integral role in the neuron's receptive capabilities. The axon, on the other hand, can be thought of as a road between two fixed locations, stretching from a mere fraction of a millimeter to an impressive two meters. On this thoroughfare, electric impulses, known as action potentials, race along at speeds of up to 100 meters per second. These pulses, originating from a specialized ignition point near the axon's origin, consistently maintain a magnitude of 100 millivolts, thanks to a self-sustaining regeneration process. These action potentials, despite being uniform, convey a myriad of information from our environment—light, touch, smell, sound—yet the nature of the message is not dictated by the action potential itself, but by the path it travels within our neural network. To enhance the speed of these neural signals, larger axons are insulated by a lipid sheath, myelin, which is punctuated by nodes of Ranvier, uninsulated spots on the axon where the action potential is regenerated. Finally, at the axon's end, it branches into presynaptic terminals. These points of division resemble a river delta, where the primary flow branches out into smaller tributaries. Each terminal connects with other neurons at junctions known as synapses, creating a complex, interconnected network. At each synapse, the presynaptic cell dispatches signals from its terminal, delivering them across a minuscule chasm, the synaptic cleft, to the receiving postsynaptic cell. The recipient of these messages could be another dendrite, a cell body, or even another axon. Each incoming message stimulates a cascading dance of signals and pathways, a ballet that gives rise to our sensory experiences and interpretations of the world around us.

These brain-rooted "trees" are classified into various types based on their morphology, function, and neurotransmitter activity, each playing a unique role in the grand orchestra of cognition [13]. Sensory neurons serve as the harbingers of perception, transducing the whispers of the external world into a rich tapestry of neural signals. These neurons detect stimuli from our environment, such as light, sound, and touch, and relay this information to the central nervous system for further processing and interpretation [7]. Interneurons, the mediators of the neural ensemble, weave intricate connections within the central nervous system, integrating and modulating the flow of information between sensory and motor neurons. These neuronal maestros sculpt the emergent patterns of activity that underlie complex cognitive processes, such as learning, memory, and decision-making [3]. Motor neurons, the orchestrators of movement, transduce the directives of the central nervous system into a symphony of muscular contractions, coordinating the graceful dance of our every action. These neurons innervate our muscles, enabling us to navigate the world with precision and dexterity [14].

In the realm of memory, neurons engage in an elaborate *pas de deux* of synaptic plasticity, strengthening or weakening their connections in response to experience [15]. This delicate interplay of activity-dependent modulation forms the foundation of our memories, shaping the neural pathways that define our cognitive landscape and giving rise to the rich tapestry of our inner lives.

2.2.2 Electrical Potentials: The Language of Neurons

Imagine our brain as an expansive, densely intertwined forest, teeming with billions of trees, each representing a neuron. Within this intricate neuronal wilderness, messages—the lifeblood of cognition—traverse an unseen path, rushing from one tree to another. This silent dialogue, a complex language of electrical potentials, from the evanescent whispers of graded postsynaptic potentials, to the thunderous crescendo action potentials [3]. This dynamic interplay underpins our very existence—every thought, sensation, and action—and the processes of memory encoding, consolidation, and retrieval, weaving together the tapestry of our past experiences and shaping our cognitive landscape [16].

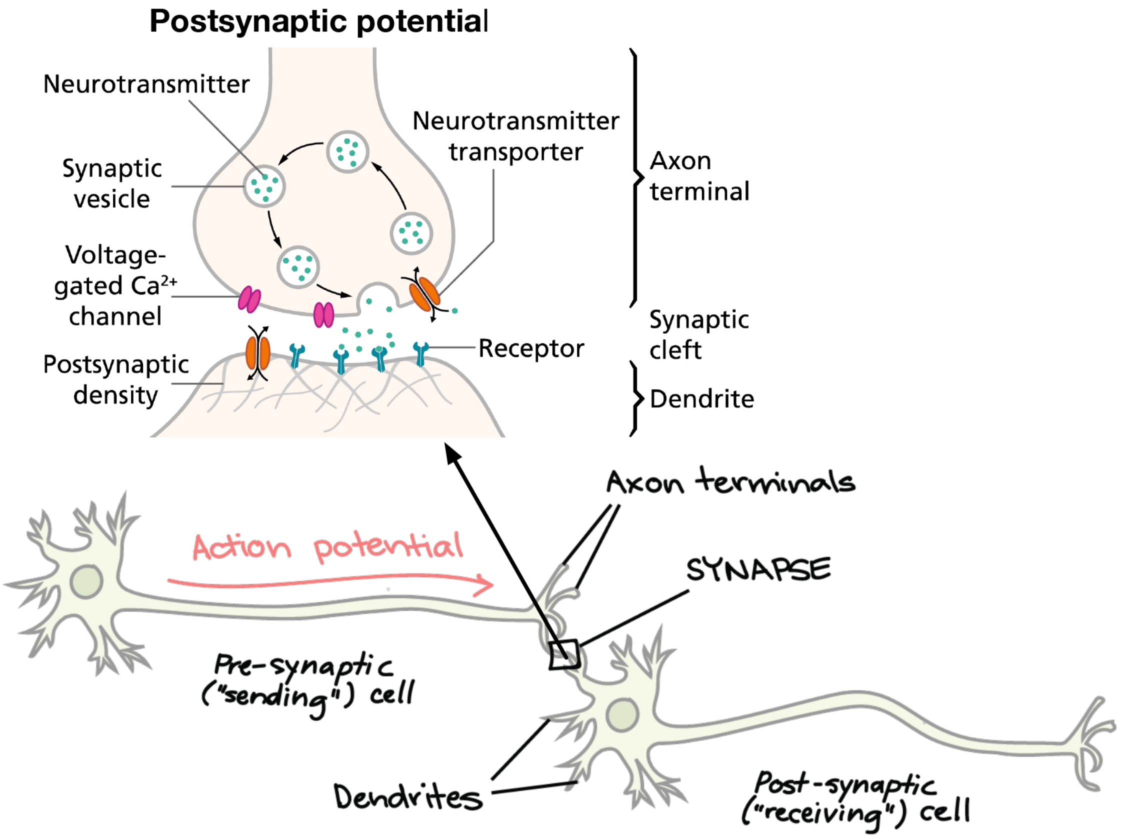


Figure 2.3: Schematic diagram of electrical potentials. Constructed via [17], [18]

Action Potentials

Action potentials act as the tireless messengers in our neuronal forest. They are discrete voltage spikes that travel from the cell body, along the axon to the presynaptic terminal, where neurotransmitters are released. They originate from a delicate interplay of ions, primarily sodium and potassium, in a precise choreography regulated by the neuron's ion channels [3]. The opening and closing of these channels precipitate an influx of sodium ions followed by an efflux of potassium ions, sparking a rapid electrical spike—the action potential. Unlike the whispering wind in our forest, this electrical spike is not gradual; rather, it is all-or-nothing, a definitive decision reached when the electrical charge inside the neuron surpasses a critical threshold. Once initiated, this "spark" races down the neuron, a potent, unchanging message, echoing through the depths of our forest. Even as the electrical storm subsides, its message remains, prompting a cascade of events at the synaptic terminal. Here, we transition from action potentials to their counterparts, the postsynaptic potentials.

Postsynaptic Potentials

Postsynaptic potentials are the voltages that arise when neurotransmitters bind to receptors on the membrane of the postsynaptic cell. The action potential, upon reaching the neuron's terminal, triggers the release of neurotransmitters into the synaptic cleft, which once across the synaptic cleft bind with receptors on the next neuron. This union elicits an electrical response that can be either excitatory or inhibitory—a call to action or a gentle shushing—influencing the likelihood of the receiving neuron firing its own action potential [19]. The varying intensity of these postsynaptic potentials creates a nuanced language, rich in depth and complexity. Unlike the resolute action potential, these potentials are graded and summate [20]. Multiple whispers can combine to create a roar, or conversely, several hushes can quiet the loudest shout. This finely tuned balance enables neurons to integrate diverse inputs and respond appropriately, the crux of neuronal computation.

The intricate interplay of action potentials and postsynaptic potentials forms the essence of the brain's electrical narrative, weaving an intricate tale within the silent forest of neurons—a tale that narrates the enigma of thought and the marvel of consciousness.

2.3 Electroencephalography

Electroencephalography (EEG) is one of many tools that help us bridge the gap between the enigmatic language of neurons and our understanding. It amplifies and translates the vibrant electric dialect into a language we can decipher. Central to the operation of EEG is the concept of volume conduction. The neural dialogue originates in the choreographed flux of ions (Na^{+}, K^{+}, Ca^{++}, and Cl^{-}) across neuronal membranes, which generate an electrical field. However, between the neurons and the EEG electrodes lie various layers of tissues—brain matter, cerebrospinal fluid, skull, and scalp—each imposing their resistance and modulating the signal. Volume conduction refers to the spread of current through a conductive medium, such as the brain, and how it changes as it passes through that medium [20]. Like light refracting through many prisms, each layer adding its signature to the electrical narrative.

Crucially, EEG primarily measures the slower, graded potentials resulting from synaptic inputs—the postsynaptic potentials—which can combine, like tributaries forming a mighty river, generating a sizeable electrical field which can be detected through the skull. The neurons whose activities are primarily responsible for these potentials are called pyramidal cells. Oriented perpendicular to the cortical surface, their alignment and the synchrony of their activity make them perfect broadcasters of the electric currents that EEG picks up. These cells, stacked like rows of miniature radio towers, send their signals skyward, ready to be captured by the awaiting EEG electrodes [20]. Each pyramidal neuron, during transmission of an action potential or upon receiving a barrage of synaptic inputs, generates a current that flows through its body, creating a dipole—an entity with a positive and a negative end. The sum of these dipoles, from countless neurons, generates a fluctuating electric field. However, only those dipoles that align parallel to the scalp contribute significantly to the EEG signal. This architectural detail of pyramidal cells makes them key contributors to the signals we interpret [20].

The strength of EEG is, however, not in its spatial resolution; other techniques like functional Magnetic Resonance Imaging (fMRI) or Positron Emission Tomography (PET) offer finer spatial detail. Instead, EEG's forte is its remarkable temporal resolution. It captures the ebb and flow of neural dynamics in real time, offering a vibrant tableau of the brain's electrical landscape [20]. This oscillatory activity, classified into distinct frequency bands—alpha, beta, gamma, delta, and theta—offers us a picture of the brain's multifaceted processes and states of consciousness. Thus, EEG serves as a vital tool facilitating our exploration of the neural landscape. Combined with machine-learning in Chapter 5, EEG is used to study memory mechanisms and memorability. The chapter outlines two studies that explore the use of EEG data to predict individual recognition of videos upon re-watching, hypothesizing that neural signals at specific moments can indicate whether an event will be remembered or forgotten. It also compares various methodological approaches, focusing on theta band activity over the right temporal lobe, thereby fostering a cross-pollination of ideas between neurophysiology and computational research.

Memory

> *"We are what we remember. If we lose our memory, we lose our identity and our identity is the accumulation of our experiences. When we walk down the memory lane, it can be unconsciously, willingly, selectively, impetuously or sometimes grudgingly. By following our stream of consciousness we look for lost time and things past. Some reminiscences become anchor points that can take another scope with the wisdom of hindsight."*
>
> — Erik Pevernagie

2.4 The Weight of Memory

Memory lies at the very heart of the human experience. Its full weight is masterfully captured in the diametrically opposed stories of two South American authors—Gabriel Garcia Marquez with his tale of pestilence, "One Hundred Years of Solitude" [21], and Jorge Luis Borges with his story of blessing, "Funes the Memorious"[22]. Marquez's story describes a plague that robs the residence of the very essence of human identity—memory. Starting with intimate recollections, the plague progresses to eradicate the names and functions of common objects. One man, in a desperate attempt to hold onto his thinning strings of reality, labels each object in his home, a strategy which reveals itself to be tragically futile as the plague eventually consumes even his knowledge of words and letters. This poignant story serves as a cutting reminder of the pivotal role that learning and memory play in the daily ebb and flow of our existence. Learning, the intricate dance of mental evolution that results in a behavioural shift, hinges upon absorbing knowledge from the world around us. Memory, the elegant partner in this dance, encodes, stores, and later retrieves this acquired knowledge. Marquez, invites us to step into a world devoid of the ability to learn and remember. A world where familiar faces and cherished friends become alien, language loses its meaning and significance, and motor skills once taken for granted vanish like smoke in the wind. Without memory, we would be marooned in the present moment, devoid of past experiences or foresight, the stark spectrum of emotions, once vividly imprinted on the heart, would fade away, eroding our sense of self. Learning and memory emerge, not merely as intellectual faculties, but as integral pillars upholding the structure of human autonomy and survival.

Borges' story, on the other hand, draws us into the world of Ireneo Funes, a young man gifted and cursed with an infallible memory. The story is not merely about Funes' prodigious recall, but about the overwhelming and even paralysing abundance of detail and relentless presentness that comes with it. Funes finds himself perpetually adrift in a sea of intricate detail, each instant blooming into a myriad of perceptions, each as vibrant and as demanding as the other. This may at first sound like a desirable gift, perhaps even a superpower of sorts, but as we delve deeper into Funes' world, we witness the torture inflicted by this unending stream of memory. Borges, through the character of Funes, vividly demonstrates the critical role that forgetting plays in our lives. The ability to forget, to filter out irrelevant details, to compress experiences into manageable chunks, to discard the minutiae, is a vital cognitive process that allows us to navigate our lives effectively. We generalise, we abstract, we categorise. Without these capacities, the richness of every moment becomes an unbearable burden.

Historically, the hypothesis of localising cognitive functions was widely accepted [23]. However, memory, with its intricate interconnectedness to perception, language, and movement, posed a unique challenge to this theory [24]. A significant number of researchers, well into the mid-20th century, doubted the existence of memory as an autonomous function. Their scepticism stemmed from the observation that memory storage is not confined to a single area but rather sprawls across diverse parts of the brain [25]. Nonetheless, contemporary neuroscience has illuminated that while various brain regions partake in the act of memory storage, their importance is not equal [26]. In recent decades, researchers have made considerable progress charting the mechanisms of learning and memory [3]. In this chapter we focus on three key insights. Firstly, the multi-dimensional concept of learning and memory is not a monolith but a constellation of diverse forms, each bearing unique cognitive characteristics, mediated by an array of neural systems [27]. Secondly, memory is not a static entity, but rather a dynamic process that can be broken down into stages of encoding, storage, consolidation, and retrieval [26]. Finally, the occasional slip-ups and distortions of memory, far from being mere defects, serve as invaluable guides illuminating the nature and function of the learning and memory system itself [28].

2.5 The Makings of Memory

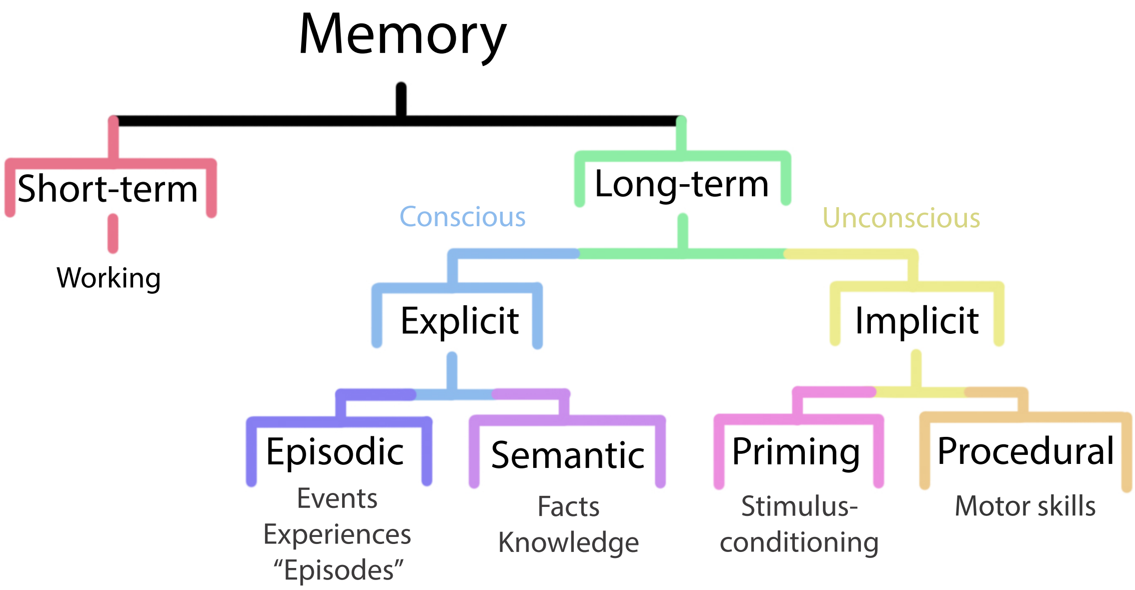


Figure 2.4: Types of memory.

Memories come in many forms [29], from the apathetic and fleeting, to the emotive and everlasting. Theories of memory have long emphasised the range of time scales over which memory operates. A common distinction is between short-term—facilitating the persistence of mental representations across brief temporal gaps—and long-term—enabling acquired knowledge, even from the distant past, to influence current cognition. Long-term memories are those that you can recall days, months, even years after they were originally stored. The information that makes it into long-term memory naturally only represents a fraction of the quotidian experience. Most information held in the brain is only temporary, lasting on the order of hours. The transience of these short-term memories makes them uniquely vulnerable to disruption. For example, short-term memories, but not long-term memories, can be erased by head trauma or electroconvulsive therapy use to treat psychiatric illness [30]. These observations lead to the notion that facts and events are stored in short-term memory, and a subset a subsequently converted into long-term memory via a process called memory consolidation. A second, entirely distinct form of temporary storage, lasting mere seconds, is working memory. Unlike short-term memory, it requires continuous conscious effort (explicit rehearsal, either in the form of mental or verbal repetition), is said to be where we hold information pertinent to a cognitive process we are engaging in (e.g., remembering a phone number before you write it down), and the set of memory procedures that directs associated attention and processing [31]. Working memory's relationship to processing means that it relies most heavily on the prefrontal cortex.

The term "memory", as used in everyday language, typically refers to a form of long-term memory, explicit (or, declarative) memory, which can be consciously recalled. Explicit memories can be episodic—relating to experiences or "episodes" in your life—or semantic—relating to facts or general knowledge [32]. The counterpart of explicit memory is implicit (or, non-declarative) memory, which is a collection of non-conscious abilities. Implicit memories can be procedural—involving learned motor skills—or a product of priming—occurring when exposure to one stimulus influences the brain's response to another [32]. Both explicit and implicit memory are types of long-term memory, but the distinction between them is fundamental as they are supported by different brain systems. Explicit memories rely on the medial temporal lobe: the hippocampus; neocortex; and amygdala, whereas implicit memories rely on the basal ganglia and cerebellum.

2.6 Short-Term Memory

The demarcation between short-term memory, our mind's ephemeral sketchpad, and long-term memory, the grand library of our past, has been a long-standing, yet contentious concept in cognitive neuroscience. This contention stems from attempts to integrate findings from studies of short-term memory painted with broad scientific strokes, originating from different depths of analysis, varying experimental designs, and investigation across an array of species. A review of this rich body of literature reveals that there is no single neural mechanism, system, or process that supports performance on short-term memory tasks. Further complicating our understanding is the notion short-term and long-term memory may be intricately linked [33].

2.6.1 Neural Mechanisms

Imagine short-term memory as a choreographed dance of intricate neural activity. At the center stage, the prefrontal cortex (PFC) and dorsolateral prefrontal cortex (DLPFC) twirl and pirouette, vital to the maintenance and manipulation of this ephemeral memory [34]. It is as though the PFC juggles various bits of information, while the DLPFC, a dexterous performer, can tweak and transform this data mid-act [35]. Accompanying this dance is the subtle hum of theta (4-8 Hz) and gamma (30-100 Hz) band oscillations. These neural rhythms, primarily found in the PFC, serve as the background music, linking the ephemeral dance of short-term memory to the grand ballet of long-term memory [36]. This harmonic connection implies that short-term and long-term memory might not be separate dances, but rather different sequences in the same performance. The synaptic plasticity of our brain forms the very dance floor on which this performance unfolds. Short-term synaptic plasticity flexes and molds to accommodate the quick shifts and twists of short-term memory, while long-term synaptic plasticity provides a lasting foundation for the footprints of long-term memory [3], [37]. This convergence of short-term and long-term adaptation again echoes the intricate interplay between these two memory forms. Lending more depth to this dance is the role of neurotransmitter systems. Dopamine, for instance, conducts the performance, influencing the maintenance of short-term memory in the PFC [38], and simultaneously modulating the formation of reward-related long-term memory [39]. The evidence pointing to the dual influence of neurotransmitter systems underscores the potential interconnectedness of short-term and long-term memory processes [40].

2.6.2 Brain systems

Research using amnesic patients and neuroimaging techniques has found significant involvement of the medial temporal lobe (MTL), particularly the hippocampus, in short-term memory (STM) processes. Several studies have shown that patients with hippocampal damage show immediate memory impairments, especially for spatial relations and arbitrary associations such as the locations of objects within a complex scene or the associations between faces and scenes [41]–[43]. Notably, these effects were not confined to spatial memory. Patients also showed deficits in retaining the associations between objects and their locations after an 8-second delay, further implicating the MTL in short-term memory retention [43], [44]. The same regions implicated in long-term memory retention also appear to play a role in short-term retention, specifically for relational information.

Further research provides additional evidence for MTL involvement in STM tasks that require retention of multiple items, or “memory load” [45], [46]. Both gamma and theta activity in the MTL were found to increase with higher STM loads, which was confirmed by fMRI studies demonstrating similar hippocampal activity increases [46]–[48]. Moreover, studies showed that there is a correlation in activity between the inferior frontal gyrus (IFG) and the hippocampus when the STM load increases during a delay period, pointing to a functional connectivity between these regions in STM tasks [49]. Hence, the encoding and retention of higher loads across a delay seem to involve the MTL, entailing roles for the hippocampus, entorhinal, and perirhinal cortices. Interestingly, while verbalisable items like digit sequences can be maintained independently of the hippocampus, recent evidence suggests that the hippocampus may be especially important for maintaining temporal sequence information. Data shows that neurons in the hippocampus may code different temporal intervals within behavioral tasks, which could mediate the hippocampus' role in memory for temporal order [50]–[52]. Computational modeling supports the idea that these responses could arise from persistent spiking activity of neurons in the entorhinal cortex [53]. This selective firing response to temporal intervals may facilitate the learning of items or events that occur at specific time points.

2.7 Explicit Memory and the Medial Temporal Lobe

The type of memory that is of particular interest in this thesis is explicit memory, which includes the two most pertinent types of memory to everyday individuals—episodic (events) and semantic memory (facts). The three areas of the brain most involved in explicit memory are the hippocampus; neocortex; and amygdala. The neocortex acts as a store of consolidated memories, while the amygdala attaches emotional significance to them. More emotive memories are less easily forgotten, which means that the amygdala can modulate the "permanence" or "stability" of a memory—how effectively it is retained over time [54]. The hippocampus plays a pivotal role in the memory formation and consolidation process [55], and is thus arguably the most important of the three brain regions.

2.7.1 Hippocampal Contributions: Seahorse in the Shell

The hippocampus was first discovered in 1564 by Julius Caesar Arantus, but it was not until the beginning of the 20$^{\text{th}}$ century that researchers began to discover the extent of its functions [56]. After the hippocampal formation (hippocampus proper, dentate gyrus, the subicular complex, and entorhinal cortex) was discovered to be a part of the limbic system (formerly limbic lobe) [57], the hippocampus was found to be a regulator of emotional behavior [58], [59]. In 1953, a patient named Henry Molaison had both of his hippocampi removed during an experimental operation to treat his epilepsy [60]. While his epilepsy was cured, Henry was no longer able to retain anything beyond what was in his short-term memory; he lost the ability to permanently store new information, and his retained memories were limited to rudimentary semantic and episodic memories from long before his surgery. He did, however, retain his procedural memories, and the ability to improve on motor memory tasks despite not being able to remember practicing [61].

The tragic case of Henry Molaison and its ensuing experiments generated five main findings: (1) that memory is a distinct cerebral function, dissociable from other perceptual cognitive abilities; (2) that amnesia spares short-term and working memory; (3) that amnesia is an impairment of explicit memory; (4) that the hippocampus is a core brain structure supporting explicit—but not implicit—memory; and (5) that the hippocampus plays a crucial role in the formation of memories, but that is not the site of permanent memory storage [62], [63]. Current research supports the notion that the hippocampus holds onto memories while they mature, before they are properly stored elsewhere—in the cerebral cortex, the outer layer enveloping the rest of the cerebrum [63]. Additionally, there is evidence to suggest a separation of roles within the hippocampal complex, with the hippocampus proper handling episodic memory and spatial learning, and the parahippocampal and entorhinal cortices handling semantic memory [55].

2.8 Episodic Memory

To be alive is to be subject to a never-ending stream of sensory information. Our memories, however, do not adhere to the time and tide of our experience. Instead, we remember the past as a series of episodes that are discrete and meaningful in nature. Episodic memory refers to this ability to cut up the continuous flow of experience into recollectable packets, and it has three core concerns: what, the narrative context; where, the spatial context; and when, the temporal context [64]. While explicit memory (episodic + semantic) is the common conception of the word memory—as used in everyday language—episodic memory in isolation is arguably the most anthropocentrically germane type of memory as our awareness of the tapestry woven from all our episodes is what gives rise to our sense of having a self. The question is, how are episodic memories formed?

2.8.1 Memory as Reinstatement

The memory as reinstatement (MAR) model [3]—when we remember, the brain returns to a prior brain state. Imagine our brains as intrepid explorers navigating the immense ocean of memory. The voyage begins with a confluence of sensory experiences, emotions, and internal thoughts, represented as a dynamic, bustling seaport of neural activations spanning cortical and subcortical shores. The hippocampus stands as a mighty lighthouse, receiving this bustling activity, constructing a unique sea chart for the current experience—the Hippocampal Neural Pattern (HNP), echoed in the coastline of the cortex, the Cortical Neural Pattern (CNP). A core navigation tool in our oceanic journey is the hippocampal pattern completion, where our hippocampal lighthouse replays the HNP when a memory is called upon. This is sparked by the lighthouse's detection of familiar landmarks or cues, which prompts the recall of the seafaring path, bringing the CNP back into view [65]. Navigational tools such as neuroimaging allow us to chart these exploratory voyages in the brain. A notable tool, the Difference due to memory (Dm) paradigm, compares the navigational charts of remembered versus forgotten voyages. Studies using this approach have found that the Medial Temporal Lobe (MTL) often paints a more vibrant, detailed chart when the memory journey is successfully remembered [66], [67]. Within the MTL archipelago, different islands contribute uniquely to the grand voyage. The hippocampus, like an expert cartographer, captures the intricate topography of our experiences, while the perirhinal cortex (PRC) serves as a more straightforward compass, aligning more closely with the recognition of distinct landmarks or items [68], [69].

2.8.2 Formation: the What, the Where, and the When

At a very high level, sensory information flows through one of many sensory pathways—for vision, touch, hearing, etc.—from its respective sensing apparatus (i.e., eyes, skin, ears) where information about the identity of perceptual objects and events are initially processed, and then projected onto multimodal cortical association areas, where the information is integrated into a cohesive representation, consisting of perceptual and conceptual information about "what" occurred is formed. As for the "where", an analogous yet distinct set of pathways in the cerebral cortex enable the formation of a spatial representation.

Information processed through these distinct streams converge in the medial temporal lobe (MTL) where clusters of neurons represent and contextualise both the "what" and "where". Within the MTL, the perirhinal cortex (PRC) and the lateral entorhinal cortex (LEC) organise populations of neurons in relation to specific object stimuli, whereas the parahippocampal cortex (PHC) and the medial entorhinal cortex (MEC) organise neurons in relation to the spatial context in which an event occurs [70]–[72].

While there is a considerable record of research in relation to how the "what" and "where" are formed and encoded, it is only very recently that the same can be said for the "when". An indispensable feature of episodic memory is our ability to temporally piece together different elements of an experience into a coherent memory. But how is the continuous stream of sensory information broken down into discrete events, and how is their temporal order preserved? The answer is “time cells”, which are neurons in the hippocampus and entorhinal cortex that fire at consecutive moments during an empty interval between two events [51], [52], [73]–[76]. Much as a population of “place cells” provides a map of a spatial context, a population of time cells map the progression of time through a situational context. Time cells keep track of the contextual stability over time in order to group sequential information into events. Similar to how place cells remap when an subject is moved to a new spatial context, time cells “retime” when the temporal structure of the current behavioural context is changed [52], [77], [78]. Remaining in the same spatial context for an extended period of time, such as cooking dinner in the kitchen, may help to organise a sequence of actions, such as preparing the food and then cooking it, into a unified event representation of eating dinner at home [79], [80]. However, when the context changes, such as entering a new room or being interrupted by the doorbell, people tend to perceive an event boundary that defines the end of the current event and the beginning of a new one [80], [81]. Importantly, crossing a single event boundary—including fluctuations in an individual’s surroundings or mental state—can impact a person’s prospective perceptions of the temporal nature of their experience, and suggests that the episodic memory updating that occurs during an event boundary both captures attentional resources, and plays a role in the temporal binding of information [82]. While time cells keep track of the temporal progression within an event, hippocampal ensembles (neurons firing in a synchronous manner) also keep track of time at longer timescales by a slow drift of their population activity over time [83]–[86]. Within the MTL, this representation of the “when” converges with the “what”, and the “where” at the level of the hippocampus. In the hippocampus, these three representations are bound together into a cohesive event and stored as memory engram (an orchestrated ensemble of neurons) which is freely available to be recalled at a later time [87], [88].

2.8.3 Encoding Oscillations

Oscillatory brain activity patterns have started to receive increasing recognition for their contributions to memory encoding, supplementing our understanding derived from event-related potential (ERP) studies—which involve measuring the electrical activity in the brain in response to a specific stimulus or event, providing insights into the cognitive processes involved in tasks such as perception, attention, and memory. This includes activity in theta, alpha, and gamma frequency bands [89]. A key player among these is the theta rhythm, which has been suggested to be instrumental in memory encoding processes, potentially via its link with long-term potentiation (LTP), a cellular-molecular mechanism crucial to memory formation [90].

Gamma frequency bands offer another intriguing facet of the encoding-related activity landscape. Specifically, studies have observed an early surge in gamma power to positively correlate with successful memory encoding, bolstering its role in the encoding process [91].

Contrary to the actively facilitating roles of theta and gamma rhythms, alpha rhythms have been proposed to act as functional inhibitors within the brain. The enhancement of alpha power has been hypothesised to facilitate encoding by actively suppressing task-irrelevant areas, thereby mitigating potential interference [92]. However, the dynamics of this relationship and its impact on memory performance could vary based on factors like the specific task at hand and the brain region involved.

Beyond examining individual oscillatory patterns, a holistic interpretation of EEG data necessitates an understanding of the interactions between different brain regions during memory encoding. Two regions of particular interest in this regard are the medial temporal lobe (MTL) and the prefrontal cortex (PFC), which are both implicated in memory encoding but appear to fulfil distinct roles [93]. MTL regions, with the hippocampus at the forefront, are believed to weave together different aspects of an episodic memory into a unified representation. In contrast, PFC regions are thought to contribute to the optimization of memory formation by curating and structuring the semantic and contextual elements of the episode [94].

2.8.4 The Role of Context

The theory of context-dependent memory, a key component of episodic memory, postulates that the more the encoding context of information matches its retrieval context, the better the memory performance [95]. This model paints a picture where our memories are like fish swimming in the sea of context. When we recall a memory, we're fishing in the vast ocean of our minds, and the more familiar the fishing spot (the context), the more likely we are to catch that fish (the memory). This "contextual sea" could contain everything from the physical surroundings to the mental state at the time of encoding, much like diverse marine life inhabiting different water layers.

The Temporal Context Model (TCM) [96], proposes that an ever-drifting state of context can assist the organization and search of episodic memories. This state, they argue, moves more slowly compared to the rapidly changing items (e.g., a list of words) in memory. Picture a tortoise and a hare in a perpetual race in the brain; the tortoise, the slowly drifting context, helps provide a time-stamped backdrop against which the fast-paced hare, the transient representations, are juxtaposed. Furthermore, TCM posits that memory search is a two-step process: "retrieval" and "reinstatement." During retrieval, the context serves as a cue to remember, while during reinstatement, the context associated with a particular memory is reinstated to aid further recall. It's much like finding an old photograph in a box—the photo (retrieval cue) leads you to recall the event, and the recall, in turn, brings back the associated feelings, thoughts, and more, essentially reinstating the context [97]. But how does our brain keep the "tortoise" of context steadily moving in the race of memory retrieval? Part of the answer lies in the hippocampus' role. Studies suggest that the hippocampus, known for its involvement in episodic memory, binds co-active representations together, including those that drift at different rates [98].

This binding allows words to cue the retrieval of co-active contextual threads, akin to the concept of pattern completion in neuroscience, where partial cues lead to the retrieval of a complete memory trace [99]. Another fascinating aspect of the hippocampus' role in context-dependent memory is the replay of neuronal firing patterns during rest or sleep, which is believed to play a significant role in memory consolidation [100]. It's like the hippocampus is a conductor of an orchestra, replaying the day's music (memories) to its sections (neurons), helping them to remember their parts better. Slow drift can result from the brain representing slow changes in environmental features, such as location changes, or through intrinsic maintenance, where the brain continues to fire patterns of neural activation corresponding to world features and thoughts, even when they no longer exist [101]. This process is like leaving footprints in the sand; even when the person is long gone, the imprints persist, marking their presence.

2.8.5 The Subjectivity of Time

The field of psychology of time has long distinguished between prospective time—time estimation based on perception—and retrospective time—time estimation based on memory—to highlight the difference between our sense of duration during an experience, and our sense of duration in hindsight [102].

Temporal information is integral to our memories, allowing us to remember when an event took place (temporal context), how recently an event occurred (temporal recency), the order in which events unfolded (temporal order), and how much time elapsed between (temporal distance) and during (temporal duration) two events. Temporal information in the LEC does not arise in an explicit clock-like manner, but from the underlying dynamics of the representation of ongoing experience in the LEC [76], [103], suggesting that our temporal representations, like our spatial representations, are allocentric. Additionally, despite clock duration remaining constant, more distinct and numerous sub-events leads to longer recognition memory responses, retrospective duration estimations, and mental event replaying [104].

Strikingly, prospective duration estimation has also been found to be modulated by episodic event structure. When event content and duration are both attended to, more distinct and numerous sub-events also lead to longer duration estimates, but when duration is exclusively attended to, only the number of sub-events leads to longer estimated durations [105]. These findings indicate that incidentally or intentionally encoded episodic event structure modulates prospective duration.

While existing functional magnetic resonance imaging (fMRI) work [106]–[108] certainly supports the notion that human cortical representations reflect subjective—rather than objective—retrospective temporal memory, methodological limitations associated with fMRI make it challenging to provide definitive insight into the nature of neural activity associated with temporal memory. Such limitations make it unclear whether observed changes in neural activity reflect an egocentric passage of time, or simply reflect changes in the quality and/or quantity of externally experienced events. However, given that changes in contextual information, transitioning from one episode to another, and fluctuations in the number and structure of experienced events all have an impact on an individual's memory of the amount of time that has passed between two time points [82], [104]–[107], it is far more likely that temporal memory is indeed subjective.

Remembrance: The Act of Remembering

> *"We do not remember days, we remember moments. The richness of life lies in memories we have forgotten."*
>
> — Cesare Pavese

2.9 Recognition vs Recall

The quantification of memorability is dependent on how we measure remembrance—which is in turn dependent on the modality and measurement paradigm. Broadly speaking, there are two ways to measure remembrance: as *recognition*, where amidst content presentation, participants indicate which items they feel they have previously perceived; or as *recall*, where participants recount as much information as they can concerning previously presented content. These two measures respectively align with the two memory processes posited by the psychological dual process model of memory called *process dissociation* [109]. The first memory process is rapid, unconscious, and typically driven by a feeling of familiarity while the other is slower, conscious, and driven by a detail retrieving intention.

Recognition and recall represent the twin pillars that uphold the vast edifice of memory. While intertwined and symbiotic, they serve distinctly different roles in our quest to retrieve our past. Recognition, the acknowledgment of the familiar, often serves as our first instinctive response to stimuli. In contrast, recall, the recreation of past experiences, is an active and sometimes challenging quest into the depths of our stored knowledge. Understanding the nuanced differences between these two modes of memory retrieval is paramount to the exploration of the web of memory. Each plays its part, an essential gear in the intricate mechanism of remembering, contributing uniquely to the song of cognition. A comprehensive understanding of the subtle interplay between recognition and recall is foundational; it is essential for deciphering the complex mechanisms by which we remember, interpret, and construct our past, ultimately shaping our responses to the present and future.

2.9.1 Recognition: Echoes of Familiarity

Recognition fulfills a crucial role in our cognitive landscape—a search of brain bound impressionist paintings of previously encountered experiences. Imagine walking down a bustling street. You are bombarded by a sea of faces, then, amidst this maelstrom, you spot a familiar face. Instantly, a flash of recognition illuminates your mind—that face, you know it. This is the effortless act of recognition, you did not need to rummage through your memory, trawling for associated details; instead, you experienced an immediate sense of familiarity. The friend's face acted as an intrinsic trigger, which either culminates as is, or unspools into a thread of remembered experiences [110].

The role of recognition, however, extends far beyond mere familiarity with faces. It is an integral layer of our cognitive landscape, incessantly sifting through the sensory deluge, distinguishing the familiar from the novel. Be it identifying a favourite song from its initial notes, recognising your house from afar, understanding the words you read, or successfully employing a well-known strategy in a chess game, recognition is a critical cog in our cognitive machinery [47]. The profound utility of recognition in survival and adaptation cannot be understated. Recognising potentially dangerous entities or situations—whether they be poisonous berries, predacious animals, or the urgency signaled by a car's brake lights—guides our decisions and behaviors, nudging us towards familiar and often safer options [111]. Furthermore, recognition underpins our social existence. Recognising a friend's face, a loved one's voice, or a colleague's unique handwriting fosters our social interactions. These acts of recognition evoke associated memories and emotions, fortifying our social bonds [112]. Perhaps most intriguingly, recognition shapes our identities. Each time we gaze into the mirror, we recognise ourselves, affirming our identity. This self-recognition fosters a continuous sense of self, bridging temporal gaps and providing us with a consistent narrative of our existence [113]. Recognition is a silent guardian constantly checking our present against our past, offering familiarity as a compass in a world of ever-shifting tides. It underpins our understanding, aids learning, fosters social bonding, and shapes our identity [114].

2.9.2 Recall: Threads of the Past

Recall, unlike its sibling recognition, is not a mere moment of familiarity; it is a voyage. It is the act of reaching into the vast sea of memory, casting the net of conscious cognition, and retrieving specific pieces of information [64]. Imagine you're asked to recount the plot of a novel you read in school. It's not just about acknowledging that you've read the book—that's recognition's domain. Instead, it's about diving into the corridors of memory to retrieve the book's storyline. With a bit of initial effort, pulling at the narrative thread, you suddenly recall the eccentric protagonist, the peculiar turn of events in chapter three, or the surprising revelation that came in the book's final pages. You might even re-experience the emotions that initially gripped you as the plot unfolded for the first time. This vivid mental resurrection of the novel's narrative, constructed from fragmented memories, epitomises recall in action.

Recall is more than merely pulling past events into our consciousness; it shapes the way we interact with the world. For instance, recall is key when acquiring a new skill, like cooking an elaborate recipe. You do not just recognise the ingredients or utensils. Instead, you recall the sequence of actions, the subtle nuances of technique—how much force to use when kneading the dough, how long you should knead it for, how many layers of lamination are necessary for the perfect croissant. You draw upon memories—both consciously and nonconsciously—of previous attempts, perhaps even mishaps and lessons, to guide your current actions. This is not just retrieving information; it is the application of experience to navigate a complex task [115].

A key attribute of recall is its constructive nature. We do not possess a perfect recording of our experiences. Instead, we store bits and pieces, which we actively reconstruct during recall. This cognitive reconstruction can be influenced by our current mood, biases, or subsequent experiences, often leading to distortions or false memories [116]. The function of recall extends beyond the mere reconstruction of our past. It allows us to project our experiences into the future, to imagine or plan events that have not occurred, a cognitive ability known as episodic foresight. This capability is central to strategic planning, problem-solving, and decision making, lending us flexibility and adaptability [117]. Recall plays a significant role in shaping our understanding of the world and ourselves. Our personal narrative, our sense of self, is a tapestry woven from recollected experiences. As we recall and narrate our experiences, we construct and reinforce our identities [118].

2.10 The Quantification of Remembering

Measuring memory is much akin to attempting to gauge the hues of a rainbow; we grapple with a complex, multifaceted phenomenon that resists simplistic compartmentalisation. Nonetheless, finding ways to quantify this intricate and core facet of our existence, is key to enhancing our comprehension of its nature. This process of quantification must navigate the dichotomy of rendering an abstract, deeply individualised experience into a tangible, analysable datum. We stand at the threshold of a challenging endeavor as we seek to decode the numerical representation of human memory, a task that is as much a valuable scientific pursuit as it is an embodiment of our Freudian urge to understand ourselves.

2.10.1 Measuring Recognition

A seminal method employed in the study of recognition is the “old/new” recognition task. This paradigm focuses on a fundamental aspect of recognition——identifying something as previously encountered. Participants are initially presented with a list of items—be it words, images, sounds, or video—in the study phase. The test phase introduces a mix of old (previously seen) and new (previously unseen) items, and the participants’ task is to distinguish which items they have encountered during the study phase [119]. There are two main variants of this recognition task, free choice or yes/no, and forced choice. In the free choice variant, participants go through a long list of test items, stating ‘yes’ if they recognise an item from the study phase, and ‘no’ if they do not. The two essential measures of recognition performance obtained from this test are the ‘hit rate’—the probability of correctly identifying old items—and the ‘false alarm rate’—the likelihood of incorrectly recognising new items as old. The difference between the hit rate and the false alarm rate provides a corrected measure of performance that accounts for guessing. To add nuance, researchers may ask for confidence ratings, thereby turning the dichotomous yes/no into a multi-point scale [120]. The forced choice variant differs only in the test phase, where participants are presented with an array of item options—one previously studied item among several distractors—and they are required to select the studied item. The guessing level in such tasks can be calculated based on the number of alternatives—50% for two alternatives, 33% for three, and so on [121].

While the old/new recognition task is a potent frequently used tool, a desire for further granularity lead to the "remember/know" paradigm, which aims at separating recollection from mere familiarity. In essence, participants still categorise items as old or new, but also state whether they "remember"—can recall specifics about the item's presentation—or just "know"—find the item familiar but cannot recall specifics [122]. To study associative recognition memory, "associative recognition" tasks are employed, wherein participants must remember the relationship between two items. For example, recollecting pairings of words or associations between a face and a name [123]. This paradigm explores a different facet of recognition—our ability to link and remember related pieces of information. Lastly, "source memory" tasks delve into the context or source of a particular memory. Participants must not only recognise an item but also remember specific details about its presentation context, such as the location on the screen where it appeared, the voice that pronounced it, or even the list in which it was included [124]. This task uncovers how recognition memory intertwines with contextual recall.

2.10.2 Measuring Recall

A triad of methodological pillars: 'free recall', 'cued recall', and 'serial recall', underpin the measurement of recall in the experimental realm, they provide the foundation for a plethora of nuanced methods that capture the versatility and complexity of our recall abilities. 'Free recall' is the most elemental, a basic yet richly revealing paradigm. Participants are presented with an array of items—words, pictures, sounds, or videos—during a study phase. After a delay, their task is to retrieve as many of these items as possible, with no mandated respect for their original order. Like a miner let loose in a gem-filled cavern, the goal is to emerge with as many precious stones, regardless of where they were originally unearthed. Key measures in free recall include the proportion of items correctly remembered, or hits, to the total items presented [125], as well as the serial position effect—a phenomenon where items at the beginning (primacy effect) and end (recency effect) of the list are often remembered better [126].

'Cued recall' tasks venture a step beyond, imbuing the free recall paradigm with a guiding thread. Participants receive cues or prompts during the retrieval process, hints that illuminate a path in the wilderness of their memory. An especially insightful variant is the paired-association task, where participants learn pairs of items, such as 'pineapple-avalanche', during the study phase. In the test phase, they are prompted with one word (e.g., 'pineapple') and must retrieve its pair ('avalanche'). The cued recall task offers insights into the structure of associative memory—our ability to form and retrieve connections between pieces of information [127].

In the context of 'serial recall' tasks, the order of a sequence holds the highest importance. Unlike free recall, where the recall order is inconsequential, serial recall requires participants to reproduce items in their original sequence. This exacting nature illuminates the mechanics of temporal context memory and exposes the order-dependent character of recall [128].

Diverse variations of these central paradigms continue to enrich our understanding of recall memory. For instance, within cued recall tasks, we encounter 'contingent recall' paradigms, where the cue for recalling an item is often the previous item in a list, probing how recalling one item influences subsequent recall. Similarly, 'source recall' tasks, another variant of cued recall, require not only the recall of an item but also its contextual origin—where, when, or how it was encoded [124]. The 'prospective memory' paradigm introduces a time-based cue into the recall task. Participants must remember to perform a future task upon the appearance of a specific cue or after a certain time has passed. This turns the lens on our ability to 'remember to remember', a vital skill in everyday activities [129].

The chapter was divided into three primary sections, each focusing on different aspects essential for understanding the broader context of the thesis. The "Fundamentals of Neuroscience" section provided a basic understanding of the neural mechanisms of the brain, including an overview of neurons, their structures and functions, and the role of EEG in studying brain activity. The "Memory" section presented a comprehensive overview of biological memory, covering the neural mechanisms of short-term memory, the associated brain systems, and the roles of the MTL and hippocampus in explicit memory. Additionally, this section explored the complexities of episodic memory, including the processes involved in reinstating memories, oscillations during encoding, the significance of context, and the subjectivity of time in memory formation. Lastly, the "Remembrance: The Act of Remembering" section differentiated between the two primary forms of memory retrieval—recognition and recall—highlighting the distinct neural and cognitive processes involved, and outlined common methodologies used for measuring recognition and recall. Ultimately, this chapter aimed to provide a high-level understanding of memory, the act of remembering, and the methods to measure remembering, serving as a fundamental knowledge base for the thesis.

Chapter 3

Memorability and the Measures of Memory

"Memory represents to us not what we choose but what it pleases."

— Michel de Montaigne, *Les Essais*

In the theatre of our existence, the natural world takes center stage, unfolding in a ceaseless stream of sensory threads—from frenzied photons painting our visual landscape to the odious odorants that punctuate our olfactory experiences. Each of us wades through this storm of complex multi-sensory data, our brain serving as both court master and king. It deftly weaves threads into an intelligible tapestry of internal representation, carefully selecting which threads to incorporate and which to cast aside, relegating them to the abyss of the forgotten. Amid this whirlwind of sensory input, a question of profound consequence reverberates: what should be remembered, and what should not? The answer, elusive and enigmatic, is written in the whims of the king—the brain's assessment mechanism for stimuli relevance/importance. This question underpins my exploration into the realm of memorability, acting as the keystone that bridges our sensory experiences with the cognitive retention in the castle of memory.

Memorability—the likelihood that a given piece of content will be subsequently remembered—provides us with the *Rosetta Stone* necessary to decipher the remembering whims of the brain. It illuminates the fundamental principles that guide the cognitive processes distinguishing the memorable from the mundane. Fundamentally, memorability—in its most exact form—represents an index of experience, gauging the divergence between sensory perceptions and memory manifestations. Its proximity to the bedrock of human experience is what ultimately motivates and brings meaning to its exploration.

The examination of memorability holds substantial theoretical significance. It serves as a powerful lens through which we can comprehend the mechanisms that guide our cognitive system in distinguishing the memorable from the mundane. For instance, consider the remembrance of faces. What is it that makes us remember a face we've only seen once at a crowded party? Why do some faces, even without distinctive features, lodge themselves in our memories, while others are quickly forgotten? Memorability, as a concept, stands as a compass in the misty landscapes of such questions, pointing the way towards understanding the relationship between the nature of sensory stimuli and their potential for recall or recognition. From a practical standpoint, memorability transcends the realm of theory and seeps into a multitude of disciplines, carving a niche for itself in areas as diverse as machine learning, cognitive neuroscience, and media design.

Consider the realm of advertising, where creating memorable content is the *holy grail.* Understanding the dimensions of memorability would allow the creation of advertisements that linger in the minds of viewers, thereby increasing brand recall and influence. In the field of artificial intelligence, incorporating a model of memorability into machine learning systems could lead to more intelligent information retrieval and recommendation systems, tailoring content to individual users based on its likelihood to be remembered. Nonetheless, this capability raises ethical considerations. Tailoring content to maximise memorability implies a greater capacity to influence human behaviour, which prompts questions about manipulation and the preservation of decision-making autonomy. Conversely, as memorability serves as proxy indicator of human significance, optimising for memorability could be interpreted as optimising for human importance, a goal that arguably possesses intrinsic merit. This represents a nuanced and ambiguous area, necessitating a careful and responsible examination of the implications. In the field of cognitive neuroscience, a more nuanced understanding of memorability could pave the way for innovative interventions and therapeutic strategies for memory-associated disorders. Comprehending the attributes that render a stimulus memorable enables us to potentially augment memory recall in patients afflicted with Alzheimer's disease or other forms of dementia, thereby enhancing their overall quality of life. For instance, exposing patients to highly memorable stimuli may prompt the activation of specific neural pathways associated with memory, akin to a form of "exercise" for the brain. The underlying principle is grounded in Hebbian theory, which posits that neurons that fire together wire together. Consequently, regular activation of these pathways through exposure to memorable stimuli may help in strengthening synaptic connections, thereby promoting the maintenance of memory circuits and potentially mitigating the progression of memory-related impairments. In the following sections, we will embark on a journey to dissect memorability, tracing its manifestation across various sensory modalities and examining its interactions with other cognitive phenomena. From identifying the distinctive attributes of memorable images to understanding the dynamics of memory formation, we aim to construct a comprehensive understanding of the factors that make a stimulus memorable.

3.1 Visual Memorability: The Eye's Imprint on the Mind

As early as the 1960s, researchers sought to operationalise and quantify memorability as a metric in cognitive psychology. Specifically, in 1966, the Dow Chemical Company used the concept of memorability in an applied study seeking an effective symbol to communicate biohazard risks [130]. This study aimed to create a symbol that lacked intrinsic meaning but was highly memorable—capable of instantly and enduringly etching itself in the minds of viewers. The resulting symbol, still globally recognised today, stands as an early example of memorability's importance in visual cognition. In academic psychology, the formal study of memorability came into prominence from the late 1970s through to the early 1990s. During this time, a particular focus was the investigation of factors that influence facial memory. A crucial development during this period was the conceptualisation of face memory representations in multidimensional face space [131]. In this model, an individual's face is thought to exist as a point in a multidimensional space, where the dimensions represent various attributes of a face. These attributes can range from physical traits such as age, race, and face shape, to more socially-driven features like dominance and valence [132]. Researchers found that distinctiveness—how far a face is from the center or prototype of this representational space—strongly influences memorability. A slew of studies demonstrated that faces with higher distinctiveness tend to be more memorable [133]–[138]. This concept of distinctiveness became so interwoven with memorability that the terms were sometimes used interchangeably [137]. Despite the clear influence of distinctiveness, it could not fully account for the variance in memory performance, hinting at the existence of other factors that affect memorability [137]. This observation sparked questions about what makes an image more memorable and whether memorability, independent of distinctiveness, was a noteworthy attribute to explore further.

3.1.1 Memorability as an Intrinsic Image Attribute

The digital revolution of the 2010s fueled a data-driven approach to deciphering the enigmatic workings of human memory. This era brought forth a novel proposition: the recognition memorability of an image might be an intrinsic attribute of the image independent of an observer's mental constitution at the moment the image is observed, i.e., irrespective of the observer's emotional or personal connection to the image. This hypothesis stemmed from observations made in the pioneering research of Standing (1973) [139] and Brady et al. (2008) [140], where it was found that individuals could correctly recognise an astonishingly large number of images. More importantly, these studies also highlighted variability in the success rates of recognising different items, leading researchers to explore what characteristics might govern an image's memorability. Pioneering this exploration, Isola et al. [141] deployed a massive online experiment, engaging participants on Amazon Mechanical Turk (AMT) in a simple ye t revealing subsequent memory task. Using the Scene Understanding Database [142],a collection of 2,222 images depicting various scenes such as indoor (e.g., bank, pharmacy, bathroom), outdoor natural (e.g., lake, glacier, mountain), and outdoor man-made (e.g., chemical plant, oil rig, campsite), they collected recognition memory annotations from approximately 80 participants for each image, with their combined recognition rates providing a measure of each image's memorability. However, the crux of their research was the consistency analysis, an innovative approach to validating memorability as an intrinsic attribute of an image. In this analysis, they partitioned participants into two randomly selected halves and calculated the memorability for each image separately based on these halves. These separate memorability performance scores were then correlated with each other to assess the consistency in performance between the halves. The process was repeated over 25 iterations and averaged, ensuring that findings were not accidentally attributed to the specific halves chosen. The outcome was compelling—a high correlation (Spearman's rank correlation p=0.73, referred to later as "Human Consistency") signified that images remembered by one half of the participants were likely remembered by the other half as well. In essence, recognition memorability scores proved highly consistent across people, even in a highly heterogeneous online sample. This study convincingly demonstrated that recognition memorability could be conceptualised as an intrinsic, measurable, and a potentially manipulable property of a stimulus. Further bolstering this concept, subsequent investigations conducted by Bainbridge et al. [143] on face images yielded similar patterns of memorability consistency. This was followed by research extending the theory to diverse stimulus types, such as abstract visualisations [144], simple words [145], objects within scene images [146], and even dynamic stimuli like videos [147]. Collectively, these findings paint a compelling portrait of memorability as an intrinsic attribute of images, laying the foundation for a deeper understanding of how images imprint themselves onto our memory. The remaining mystery to unravel is the underlying mechanism of this phenomenon: What intrinsic properties of an image determine its memorability? Unraveling these intricacies will shed new light on our understanding of human memory and cognition.

3.1.2 Image Attributes and Their Influence on Memorability

Despite initial assumptions, the relationship between memorability and other image attributes is far from simple. It seems intuitive to associate memorability with familiar image characteristics such as visual distinctiveness, aesthetics, or visual saliency. However, research to date has failed to pinpoint any combination of attributes that wholly determine an image's memorability. Research has consistently found that basic image features like hue, saturation, or spatial frequency have weak to no correlation with memorability [1], [146], [148]. Interestingly, even scrabbled images devoid of most semantic content (but with preserved low-level visual features such as colour and edges), still held some degree of memorability [149]. However, the effects were transient (existing only within a short period of seconds), suggesting that higher-order perceptual attributes or deeper semantic and conceptual elements in the image might significantly contribute to memorability. Some factors commonly thought to capture attention or trigger emotion were found to have negligible correlations with memorability [1]. For example, the number of objects in a scene or the extent of image coverage by these objects did not influence the overall memorability of the scene. Moreover, factors like aesthetics, visual interestingness (a measure for captured attention), or perceived memorability (subjective estimation of memorability) were not directly linked to actual memorability. However, scene images containing faces or text tend to be highly memorable, and a combination of semantically-based object and scene attributes, such as object/scene category, emotion, actions, and dynamics, are predictive of memorability. This suggests that memorability may have stronger ties with the semantic properties of an image rather than its mere visual attributes.

When examining novel faces, an area with reduced variability in perceptual or semantic features, memorability becomes even more complex. Bainbridge et al. [143] found that attributes such as atypicality (or distinctiveness), attractiveness, emotion, and subjective memorability ratings correlated with actual memorability. Interestingly, memorable faces were often rated as more emotional, irresponsible, unattractive, and unintelligent, but also kinder and more trustworthy. Yet, even this comprehensive set of attributes could only account for 46.6% of the memorability variance, implying that face memorability is not merely a compound of other well-known semantic face attributes. More surprisingly, perceived distinctiveness of faces, even when captured by several terms (atypical, unfamiliar, uncommon), does not fully explain face memorability. Metrics such as Euclidean distances between facial points on an individual face and those on an average face showed no correlation with memorability. This information stands in stark contrast to prior work, which often equated visual distinctiveness with memorability. If visual distinctiveness does not define memorability, what does? One hypothesis is that distinctiveness may manifest in a variety of ways: perceptual distinctiveness, semantic distinctiveness, emotional distinctiveness, and so forth. However, the exact determinants of memorability remain an open question, necessitating future experimentation.

3.1.3 Beyond Recognition: Memorability in Other Cognitive Contexts

It is essential to understand that the concept of memorability transcends its role in continuous recognition tasks, the power and effects of memorability extend into various other memory paradigms. This diversification serves a critical purpose, establishing that the observed memorability effects are indeed attributable to the images themselves, as opposed to the experimental tasks they are presented within [150], [151]. For example, when image memorability is evaluated using memory tasks with distinct study and test phases, intriguing results emerge. Even when the test phase is delayed by a day or a week, the essence of image memorability prevails, demonstrating its long-term effects [151]. This robustness is also evident when the influence of context on image memorability is assessed. While memory performance can be subtly swayed by the similarity of images presented in the same context, an image retains its intrinsic level of memorability, irrespective of the variation in image contexts [150]. While these findings emphasise the durability of image memorability across different paradigms, they also highlight an interesting conundrum that has yet to be fully resolved. Specifically, recent research posits that an image's recognition memorability may not correlate with recall memorability [152]. This suggested absence of interaction is perplexing given that recognition is often considered a precursor to recall, suggesting that a more comprehensive assessment of recall memorability is needed.

Bainbridge [153] conducted a series of online psychophysical experiments to understand how certain cognitive processes such as bottom-up attention, top-down attention, depth of processing, and priming relate to image memorability. Different paradigms were used to test bottom-up attention, including a spatial cueing task and a visual search task, both of which explored the influence of memorable images on performance. The findings showed that while memorable images were easier to hold in memory, they did not significantly alter spatial attention or automatically capture attention in the visual search tasks. As for top-down attention, it was found that even though participants could intentionally modify their memories to some degree, the power of memorability was more potent; participants could not consciously forget a memorable image or remember a forgettable one. Encoding tasks with varying levels of semantic depth also influenced memory performance. However, the impact of image memorability was stronger, implying that regardless of the depth at which images were encoded, memorable ones always had the upper hand in later recognition. Finally, memorability effects were distinct from priming effects, suggesting that memorability has unique implicit effects on memory behavior. Collectively, these findings underscore that memorability is a robust phenomenon, largely unaffected by attention, cognitive control, and priming. This robustness has substantial implications for potential memorability applications. For instance, a memorable image does not need to be visually striking or salient to automatically capture attention. Moreover, the power of memorability is such that it can override other cognitive processes, even when presented alongside unrelated, mundane tasks. These findings leave us with the question of what is happening in the brain when viewing memorable images. While evidence implies that the processing of memorability is automatic and implicit, it does not involve an automatic capture of attention or correlate with the implicit memory effect of priming. Further exploration into our understanding of memorability processing in the brain could shed light on its relationship with the underlying neural substrates for vision and memory. Finally, memorability effects were found to be separable from priming effects. In other words, memorable images did not necessarily lead to quicker processing during repeated presentations, compared to forgettable images. This suggests that there are distinct implicit effects on memory behavior, further confirming the unique and robust nature of image memorability.

3.2 The Neurological Underpinnings of Visual Memorability

3.2.1 Mapping Memorability: Spatial Representations in the Brain

Memorability treads the line between vision and memory. It is defined as a quantifiable attribute of an image, akin to aspects such as color, aesthetics, or emotion.

Yet, at its core, memorability is defined by the behavioral outcome of memory. Given these unique characteristics, how does the brain process memorable images compared to established patterns of visual or memory processing?

To shed light on this, Bainbridge et al. embarked on a comprehensive investigation using rapid event-related functional magnetic resonance imaging (fMRI) [154]. Their experiment involved participants categorising the sex of 360 face images and determining whether 360 scene images were indoor or outdoor. Unbeknownst to the participants, images were composed of an equal mix of highly memorable and highly forgettable images, controlled for various low-level visual features (e.g., color, edges, etc.) and mid-level attributes (e.g., emotion, aesthetics, number and size of objects). A subsequent surprise recognition test allowed for a fascinating comparison: how do classical markers of successful memory encoding compare to the processing of memorable images? This unique approach resulted in a significant activation pattern that spread from higher-order visual areas to memory-related regions when processing memorable images. These visual areas included regions like the fusiform face area (FFA), known for being selective for faces, the lateral occipital complex (LOC) recognised for object and shape selection, and the parahippocampal place area (PPA) which is selective for scenes [155]–[157].

Interestingly, the early visual cortex (EVC), which typically responds to low-level visual properties such as edge information, demonstrated no difference in its response to memorable and forgettable images. This intriguing observation points to the fact that the effects of memorability are not dictated by merely low-level visual differences. Memory-related regions, including the perirhinal cortex (PRC) and parahippocampal cortex (PHC)—regions within the medial temporal lobe (MTL) that have been implicated in memory processing—exhibited marked sensitivity to memorability [158]. Connections to the hippocampus, specifically the anterior hippocampus, were found to be highly sensitive to memorability as well. A striking difference emerged when comparing the neural effects observed for memorability to those found for subsequent memory. The subsequent memory paradigm, used in several prior studies, contrasted remembered and forgotten images based on each participant's individual memory. The resulting signal, in theory, approximates successful memory encoding. However, Bainbridge et al.'s research demonstrated that this signal differs significantly from the neural response to memorability [154]. This divergence was most pronounced when instances occurred where stimulus memorability and individual memory differed, suggesting that these two phenomena might be processed and represented separately in the brain. Whereas memorability-based sensitivity seemed more apparent in ventral visual and memory-related regions, individual memory-based sensitivity manifested more in parietal and frontal regions, in alignment with previously reported literature [154]. Aiming to answer a question of deeper interest, namely how the brain organises memorable or forgettable images, the researchers employed a method called representational similarity analysis (RSA). RSA compares the neural similarity across all pairs of stimuli in the brain and contrasts this with a hypothesised model. The expectation was that forgettable images would cluster closely, showing high similarity, while memorable images would be widely spread, showing high dissimilarity. Surprisingly, the results contradicted this hypothesis. Memorable images were highly similar to each other, while forgettable images were highly dissimilar. This pattern was found in the ventral visual and memory-related brain regions linked to memorability, while individual memory showed a similar geometry in frontal and parietal regions.

3.2.2 Mechanisms of Memorability

At the intersection of perception and memory, a compelling narrative emerges, illustrating the close interaction between the brain and the images it encounters. This narrative, rich in conceptual significance, poses and intriguing notion: memorability is not simply an intrinsic characteristic of an image; instead, it is a dynamic quality, arising from the our brains interpretation of visual stimuli. This intriguing proposition implies that memorability embodies the cerebral sieve, delineating what information will be transposed into long-term memory from the overabundance of visual stimuli bombarding our senses. Investigations within the anterior temporal lobes (ATL) of epileptic patients, courtesy of Xie et al. [159], reveal fascinating insights into this phenomenon. The memorability of a word seems to hinge on its semantic interconnectedness within the vast neural network of word representations. Imagine this network as a sprawling tree; words deeply embedded with numerous connections– the roots of this linguistic landscape—are more likely to be remembered. Conversely, words with sparse connections—the peripheral leaves—tend to be overlooked, implying that the mnemonic weight assigned to a word is independent of its language frequency or concreteness. In a memory retrieval scenario, the brain's quest for the sought word commences from these deeply rooted, highly memorable words, leading to a quicker reinstatement of these entities during a cued recall task [159]. This phenomenon extends beyond the linguistic domain and is mirrored in our visual perception, notably for face and scene images. By examining neuronal pattern similarity in the inferotemporal cortex (IT) and medial temporal lobe (MTL), memorable elements occupy the core of this representational space, with forgettable ones relegated to the periphery [154], [160]. A counter-intuitive yet compelling divergence, however, arises from other research where the neural encoding of memorable images appears more distributed across the entire brain during Magnetoencephalography (MEG) recordings and in rhesus macaque IT [161], [162]. Consequently, this leaves the field with the challenge of reconciling these conflicting observations—whether IT neurons encode memorability via magnitude variation [163], or if a spatiotemporal transformation occurs where visual areas initially capture a stimulus's distinctiveness, which is subsequently consolidated and homogenised in the MTL and ATL.

Further complicating this cognitive landscape, memorability does not always correspond to a stimulus's distinctiveness or atypicality. Indeed, a study involving nearly 14,000 participants and the extensive THINGS database [164] found memorability to vary by object category [165]. Some categories, such as weapons, exhibited higher memorability for typical items, while others, such as kitchen appliances, showed an opposite trend. This dichotomy suggests memorability might be predicated on more than sheer stimulus distinctiveness. Subsequent analysis did uncover stronger associations between memorability and conceptual dimensions over perceptual ones [165]. However, the contributions of perceptual versus conceptual dimensions to memorability remain a nebulous territory. Memorability is observed for semantically-devoid stimuli [149], and in monkeys lacking conceptual understanding for novel object photographs [161]. Ultimately, the neuro-computational underpinnings of this incredibly consistent (Human Consistency Spearman's rank correlation of p=0.73) yet enigmatic attribute known as memorability still remain a tantalising mystery, ripe for further exploration.

3.3 More than Meets the Eye: Memorability Beyond Visual Stimuli

3.3.1 The Persistence of Prose: Textual Memorability

In exploring the space of textual memory, several parallels can be drawn with the realm of visual memorability [166]. However, a distinctive dichotomy emerges with the observation that repeated recall, which incrementally enhances subsequent recognition performance for images, does not yield the same effect for words [167].

Numerous attributes, as varied as concreteness, imagery, emotionality, and lexical associations, contribute significantly to word recall. Words that evoke robust emotions and can be easily visualised, as well as concrete words—those referring to tangible experiences—are often more memorable [168]. Adding to this complexity is the observation that words with smaller sets of associated words have an advantage over those with larger sets [169]. Minimally counter-intuitive concepts have been found to lead to better recall, suggesting that recall memorability is not an inherent property of a concept, but a property of the concept in the context it is presented [170]. At a fundamental level, our cognitive architecture displays a predilection for certain types of words. Concrete words like “pineapple” or “avalanche”, associated with tangible objects or discernible actions, tend to be more memorable than abstract words like “justice” or “love” [171]. This preference could be attributed to the rich sensory and semantic associations invoked by concrete words. Similarly, words that induce vivid mental imagery or evoke strong emotional responses are more likely to be remembered [172]. However, the memorability of words is not just a function of their individual attributes; the semantic relationships they share with other words in a network also play a crucial role. Words that serve as associative nodes in a semantic network tend to be more memorable than those with fewer connections [159]. This association-rich feature enables the brain to prioritise these words during recall, effectively enhancing their memorability. On the neural front, differential activation in the anterior temporal lobes (ATL) associates with the memorability of words. Words with high memorability elicit faster reinstatement in the ATL during recall tasks, indicating a neural prioritisation of these words during encoding and retrieval [159].

Similar to images, the recognition memorability of simple words is highly consistent across individuals, suggesting that it is an intrinsic property of words [145]. Less familiar, lower-frequency words [173]; imageable and concrete words [174], emotionally salient words [175] and the semantic context [176] in which they are presented, all enhance recognition memorability. Consistent with the idea that episodic memory encodes the word and its set of associated attributes, meanings of words are retained in favour of their lexical properties [177]. Diving deeper into the relationship between words and meanings, the seminal study in [145] reveals a fascinating perspective. Underpinning their study is the hypothesis that words are encoded more so by their meanings rather than their surface forms. With this lens, word recognition memorability becomes a factor of how much information the word communicates about its intended meaning and how few alternatives exist—that is, words with less ambiguity and fewer synonyms tend to be more memorable. Their findings present an intriguing twist: the most memorable words appear to be those that maintain a one-to-one relationship with their meanings. In other words, words that encapsulate unique cognitive fidelity—those with fewer meanings and synonyms—seem to leave a lasting impression on memory. This perspective adds a nuanced layer to our understanding of textual memorability, highlighting the interplay between individual lexical attributes (i.e., semantic uniqueness and number of synonyms) in determining the likelihood of a word being recognised.

3.3.2 Sounds that Resound: Auditory Memorability

In the grand concert of our everyday life, each sound is a distinct note contributing to an intricate symphony. A symphony not just heard, but remembered, creating echoes that shape our understanding of the world. Our exploration, thus, turns toward an often understated player in the orchestra of memory - sound. In particular, we delve into the concept of auditory memorability, the dynamic interplay of elements that enables certain sounds to linger in the chambers of our memory. Our quest begins with the fundamental question: what makes a sound memorable? A series of fascinating discoveries offers intriguing insights. Unsurprisingly, our capacity to name or verbalise a sound is strongly tied to its memorability. This phonological-articulation link bolsters the recall of sounds, with verbal sounds, in particular, being recalled more easily than their non-verbal counterparts [171], [178]. This verbal edge underscores the symbiotic relationship between our linguistic and auditory systems, painting a picture of interconnected cognitive landscapes. Yet, the portrait of auditory memorability is not solely etched in linguistic shades. The emotional brushstrokes that colour a sound significantly influence its potential to endure in memory. The emotional impact of a sound and the clarity of its perceived source converge to amplify its memorability [179]. Sounds that echo human activity (e.g., laughter, conversation, applause, music), typically associated with positive valence, are particularly memorable [180]. This positive association not only deepens the imprint of the sound on memory but also enhances its recall [181].

When pitched against its visual counterpart, auditory memorability seems to have a quieter resonance. The prevailing consensus points to the superiority of visual recall over auditory, with the latter decaying at a faster rate [182]. However, adopting such a singular perspective risks striking a discordant note in our understanding of memory. The reality is that our experiences are not confined to single modalities; they are multi-sensory amalgamations that blend sight, sound, touch, taste, and smell into a rich, resonating harmony. In this light, the role of sound in shaping multi-modal media memorability gains profound significance. Research shows that multi-sensory experiences, compared to uni-sensory ones, have a superior recall accuracy [183]. This underscores the vital role that sounds play in contextually priming our memory, providing key information that aids recall [184].

While little research has hitherto been conducted on auditory recognition memorability, interest has started to grow. A recent study suggests that similar to visual and linguistic counterparts, recognition memorability is an intrinsic property of sounds [185]. In a comprehensive crowd-sourced experiment involving 20,000 aural memory games, this research analysed the memorability consistency of various sounds across a wide cross-section of subjects. Key findings elucidated that beyond mere acoustics and cognitive salience, the familiarity of the sound source, its emotional valence, arousal, causal certainty, and the ability to verbalise it, considerably impact the memorability. Strikingly, causal uncertainty (the degree of ambiguity or confusion experienced regarding the cause or origin of a specific sound), visualisability (how easy a mental image can be formed of the source of the sound), and emotional valence (positive or negative) surfaced as the most effective predictors of memorability.

In this chapter, the concept of memorability was defined and contextualised within the framework of this thesis, followed by an exploration of its origins. The chapter investigated the proposition that memorability is an inherent attribute of images, and examined the specific properties that influence it, and how these might be applicable in other cognitive contexts. Additionally, the neurological foundations of visual memorability were discussed, focusing on spatial representations and the mechanisms that facilitate this process in the brain. Moreover, the chapter extended the discussion of memorability beyond visual stimuli, considering the role of textual and auditory stimuli in the formation of multi-sensory memories.

Chapter 4

The Influence of Modality on Memorability Prediction

> *"Our senses are our windows to the world, and sometimes the wind blows through them and disturbs everything within us."*
>
> — Thich Nhat Hanh, *Peace Is Every Step*

Memories are the tethering threads that tie us to the world, testaments to our experiences and interactions, shaping our identities and guiding our responses to the environment. The tensile strength of these threads—memorability—is not a single filament but a complex weave of fibres, each corresponding to a different sensory modality. In this intricate weave, the fibres are inseparable, their interplay obscuring the contribution of each to the overall strength and character of the thread. To gain a comprehensive understanding of memorability, it is necessary to carefully unfurl these fibres and examine their individual and combined contributions.

Sensory modalities, as we understand them, reflect the complex architecture of the brain's sensory processing mechanisms. Each sensory modality relies on a unique network of neurons, each intricately wired to function in harmony with others. While these networks are distinct, they are not isolated; they interlink and share information. This collaboration across modalities generates a unified perception of the world around us [186]. It is through this unified sensory experience that we interact with the world, and it is through this interaction that we create our memories. Hence, it stands to reason that memory, a product of these integrated systems, would also embody this complex multimodal nature.

This perspective underpins the first hypothesis in this thesis (H1): Given the multimodal nature of our sensory experience, it is postulated that memory, and accordingly memorability, should be equally multimodal. Consequently, there should be a measurable interaction of influence between the modalities in multi-modal data. These interactions, a possible reflection of our brain's interconnected sensory systems, could potentially contribute significantly to the tensile strength of our memory threads.

Further, we turn our attention to the well-documented phenomenon of visual dominance among human sensory modalities [187]. This is a fundamental aspect of the sensory system—across numerous species—as vision tends to be the primary conduit for navigating and understanding the world. For humans, this is mirrored in the significant cortical real estate dedicated to visual processing in our brains. In the context of this study, this dominance raises an intriguing question: could the visual data exert the greatest influence on memorability?

This query forms the basis for the second hypothesis (H2): In accordance with the visual cortex's established prominence, I theorise that visual sensory data may play a more potent role in shaping memorability. It is possible that the visual fibres contribute more significantly to a memory thread's tensile strength, essentially influencing its memorability to a greater extent. This chapter investigates these conjectures, aiming to disentangle the modalities' intricate weave by elucidating their individual and combined influences on memorability prediction. It involves an exploration of the interplay between visual, auditory, and textual modalities in the context of multimodal video memorability prediction and concludes with insightful reflections on how these insights can enhance our understanding and guide the creation of more memorable media content.

4.1 The Landscape of Modalities

In the realm of media, the amalgamation of distinct modalities gives birth to immersive experiences that transform raw data into consumable content. Of particular interest is video content, a rich fusion of visual, auditory, and semantic (narrative) elements that engage consumers in multidimensional ways and with a temporal aspect that extends its delivery beyond the instant at which it is first viewed. It is within this framework that we embark on an exploration of these primary modalities, looking to understand how each contributes uniquely to the memorability of the content.

The Visual Modality

The visual modality, a core aspect of video content, plays a significant role in our perception and comprehension of the vast majority of media. It embodies the dynamic and static aspects of the visual elements in a video—movement, colour, contrast, composition, etc. Numerous studies suggest that visual characteristics directly influence viewers' attention, emotion, and consequently, memorability [150], [188]. The hypothesis under investigation in this chapter suggests that the visual modality, in alignment with our natural proclivity towards vision as a primary sensory input, exerts a commanding influence on memorability, and accordingly its prediction.

Auditory Signals

The auditory aspect of videos, although not as direct or explicit as the visual component, has a nuanced and pervasive impact on viewer experience and memory. The distinct contributions of audio within a video—speech, music, or environmental sounds—each interact differently with the viewer's cognition. Speech provides an important semantic context [189] while music and sound effects or background noise can dramatically influence the mood, and thus, the recall of the video content [190].

Textual Information

In the context of videos, the textual modality, often represented as captions or subtitles, plays a subtle yet essential role. Beyond their utility for accessibility, textual elements provide an additional layer of semantic understanding, potentially enhancing the memorability of the content. Studies reveal that reading subtitles can impact the recall of video content [191], suggesting that textual data is a noteworthy player in the memory game.

The Confluence of Modalities

When these modalities intertwine within the structure of a video, they create an environment of complex intermodal dependencies. Not only do they individually contribute to memorability, but their interplay forms a multifaceted, dynamic system that shapes our memory. Importantly, this intermodal interaction is not static, but unfolds over the temporal dimension of the video content. There could be moments in the video where one modality, such as auditory information, provides context to another, like visual data, thereby affecting the memorability of that instance. It is also plausible that specific temporal characteristics or distinctive points in the video where conceptual understanding peaks, might significantly influence its memorability. While these temporal interactions are not a given, their potential existence and role align with our sequential and evolving interaction with the world. Therefore, the investigation of modalities' contributions to memorability should not only consider their individual impacts and synergistic interactions but also the possibility of crucial temporal dynamics within their interplay over the course of the video content.

4.2 Video Memorability Datasets

In the endeavour to decipher the effects of modalities on video memorability, two comprehensive and distinctive datasets are relied upon. The first is the Memento10k dataset, a broad collection of real-world short video clips [192]. The second is a subset of the TRECVid 2019 Video-to-Text dataset [193], with added annotations as part of the 2021 MediaEval Predicting Media Memorability task.

4.2.1 TRECVid 2019 Video-to-Text Dataset

The TRECVid 2019 Video-to-Text dataset [193] contains 6,000, ~ 6 second, videos and was used as part of the TRECvid evaluations in 2019. As part of the 2021 MediaEval Predicting Media Memorability task [194], three subsets of this larger dataset were annotated with memorability scores and distributed to participants in the MediaEval task. The training set encompassed 588 videos, providing a foundation for initial modeling. The development set, composed of 1,116 videos, served to iteratively refine models. Finally, the test set, including 500 videos, was employed to evaluate the participant models' performance and accuracy in predicting memorability. A unique feature of this dataset is that each video is associated with two memorability scores. These scores reflect the likelihood of a video being remembered after two distinct periods of memory retention: short-term, a few minutes after viewing, and long-term, 24 to 72 hours after the initial viewing. These scores were obtained through a unique variant of the subsequent memory paradigm (discussed further in section 5.1), dubbed the "video memorability game", proposed by Cohendet et al. [147].

Two versions of the memorability game were carried out, one was made available on Amazon Mechanical Turk (AMT), and the other hosted on a private server and made available by direct recruitment—an audience essentially made up of students. The game consisted of two phases, a short-term memorisation and recognition phase, and a long-term recognition phase. In the short-term phase, participants were expected to watch 180 videos. The game starts with 20 vigilance fillers which are repeated after a few seconds to ensure that participants are paying attention to the task, then 40 target videos are mixed in with 60 non-target fillers and repeated over the course of a few minutes. The goal was for them to press the space bar whenever they recognised a previously seen video generating binary "recognised/not recognised" short-term memorability annotations. After a period of 24 to 72 hours, the same individuals returned for the long-term recognition phase. This time, they were shown a selection of 120 videos. Among these were 40 target videos chosen at random from the non-target fillers used in the first phase, as well as 80 fillers chosen from new videos. The goal was the same, and binary long-term memorability "recognised/not recognised" long-term memorability annotations were collected. The final short-term and long-term memorability scores were calculated as the percentage of correct recognition for each video.

4.2.2 Memento10k

Memento10k memorability scores were collected through "Memento: The Video Memory Game", a memorability experiment predicated on the old-new subsequent memory recognition paradigm, where crowdworkers from Amazon's Mechanical Turk (AMT) watch a continuous stream of three-second video clips, and are asked to press the space bar when they see a repeated video. To maximise the pace and keep the experiment engaging, videos are shown as a continuous stream. When participants press their spacebar, they receive either a red (incorrect) or green (correct) flash as feedback. If a repeat is correctly identified, known as a "hit", the stream skips ahead to the next video; there is no feedback for missed repeats. Each level of the memory game contains on average 204 videos (with repeats) and lasts $\sim$ 9 minutes. The number of intervening videos between the first and second occurrence of a repeated video is known as the "lag". The game consists of "vigilance" repeats that occur at short lags of 2-3 videos and are used to filter out inattentive workers and "target" repeats at lags of 9-200 videos that provide memorability annotations. The Memento10k dataset [192] consists of 10,000 three-second videos depicting in-the-wild (filmed by non-professionals for social media/home videos) scenes, each with associated short-term memorability scores, memorability decay values (indicating the rate at which memorability scores decrease over time), action labels (e.g., running, jumping, singing, etc.), and five human generated captions (descriptions of what is depicted in the video). The memorability scores were computed with an average of 90 annotations per video, and the videos were silenced before being shown to participants. 7,000 videos were released as part of the training set, and 1,500 were provided for validation. The remaining 1,500 videos were kept for the official test set.

4.3 Training Modality Specific Models

4.3.1 Visual Training

For the visual approach, two methods and several training procedures were implemented depending on the dataset. For the first method a Bayesian Ridge Regressor (BRR) was fit with default sklearn [195] parameters using DenseNet121 [196] features, which were extracted from the first frame of either the TRECVid or Memento10k dataset. For the second method we used an ImageNet-pretrained xResNet50 [197] that was either fine-tuned (for 50 epochs, with a maximum learning rate of 1e-3, and weight decay of 1e-2) on the Memento10k training data and then further fine-tuned (for 10 epochs, with a maximum learning rate of 3e-2, and weight decay of 1e-1) on the TRECVid development set videos, fine-tuned on the Memento10k development data, or fine-tuned on the LaMem [198] dataset (the very first large scale image recognition memorability dataset consisting of 60,000 images from a diverse array of sources) depending on chosen test dataset. The videos in the Memento10k dataset are each three seconds in duration and predominantly depict a single scene, leading to minimal narrative changes between the initial and final frames. However, the quality and resolution of the videos are relatively low, often resulting in intermediate frames that are shaky, blurry, or otherwise distorted. To mitigate the risk of using a distorted frame as a representative example of high memorability during the training phase. This approach ensures that the model is trained with the most clear and representative frame available. Nevertheless, at the testing stage, it is crucial to consider the potential impact of low-quality frames on the overall assessment of video memorability. As a result, the recognition memorability score of a video is determined by averaging the predictions made for the first, middle, and last frames.

4.3.2 Textual Training

For the textual approach, a caption model was implemented, the AWD-LSTM (ASGD Weight-Dropped LSTM) architecture [199], which is a highly regularised and competitive language model. Transfer learning was used in order to fully avail of the high-level representations that a language model offers. The specific transfer learning method employed was UMLFiT [200], which uses discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing. The language model was pre-trained on the Wiki-103 dataset, and fine-tuned (for 10 epochs, with a maximum learning rate of 2e-3, a weight decay of 1e-2, and a dropout multiplier of 0.5) on the first 300,000 captions from Google's Conceptual Captions dataset [201]. The encoder from that fine-tuned language model was then used in each of the caption models, which were either trained (for 15 epochs, with a maximum learning rate of 1e-3, a weight decay of 1e-2, and a dropout multiplier of 0.8) on a paragraph of all five Memento10k training captions, a paragraph of all five Memento10k training captions that were augmented with audio tags extracted using the PANNs [202] network, or additionally fine-tuned on the first TRECVid development set captions to predict memorability scores rather than the next word in a sentence.

4.3.3 Auditory Training

For auditory features, one of two methods, and several training procedures depending on the dataset, were implemented. For the first method Mel-frequency cepstral coefficients (n_fft:2048, hop_length:256, n_mels:128) were extracted from training videos, and stacked with their delta coefficients in order to create three channel spectrogram images. These spectrogram images were then used to train an

ImageNet-pretrained xResNet34 model for 15 epochs with a max learning rate of 1e-2 and weight decay of 1e-3 to predict audio recognition memorability. For the second method, a Bayesian Ridge Regressor with VGGish [203] audio features was fitted. 128-dimensional embeddings for each second of video audio were extracted, resulting in a 384-dimensional feature set per video.

4.4 Modality Specific Model Performance

Since its inception in 2018, the MediaEval Predicting Media Memorability benchmarking task [204]–[208] has driven much of the state-of-the-art work video memorability prediction. Each year organisers of the task share a collection of videos among participants who are asked to compute and submit runs which predict the memorability score of each video in the collection. Once runs are submitted, the organisers compare the participants' submitted runs against human annotated, ground-truth memorability scores, and announce performance evaluation metrics for each participant's runs. The task has run for five years and has led to significant incremental improvements in the performance of automatic memorability prediction for short form videos. The integration of deep visual features with semantically rich attributes, such as captions, emotions, and actions, has been identified as a particularly efficacious strategy for predicting video memorability [209]–[211]. This confluence of modalities not only amplifies prediction precision but also furnishes a comprehensive perspective on the myriad factors that collectively shape video memorability. Results discussed in this section are from the 2021 task, where a Spearman's rank correlation score of p=0.658 [212] on the Memento10k dataset was state-of-the art for short-term video memorability prediction. Current state-of-the-art on the Memento10k dataset is a Spearman's rank correlation score of p=0.724, achieved by my ConceptualDream framework, and discussed in further detail in chapter 7.

Table 4.1: Results for models tested on the MediaEval2021 Predicting Video Memorability TRECVid test set (other participant runs included for reference).

	Short-Term		Long-Term	
Approach	**Spearman**	Pearson	**Spearman**	Pearson
Visual				
BayesianRidge Dense121	0.053	0.071	- 0.007	-0.18
xResNet50 Frames	0.105	0.13	-.021	-.036
Textual				
AWD-LSTM Captions	0.043	0.037	0.071	0.059
AWD-LSTM AUG Captions	0.02	0.026	0.065	0.058
Auditory				
xRestNet50 Audio Spectrograms	0.054	0.044	**0.113**	0.121
BayesianRidge VGGish	0.056	0.039	0.108	0.088
MediaEval2021 Participants				
GTHUPM [213]	0.291	0.305	0.125	0.124
Erika [214]	0.132	0.139	0.11	0.116
HCMUS [215]	0.101	0.11	0.059	0.067
AIMMLAB [216]	0.297	0.312	0.097	0.114
MeMAD [212]	0.222	0.214	0.063	0.098

Examining Table 4.1, which shows the results of the models on the MediaEval2021 TRECVid test set measured using Spearman rank correlation score, we see a rather mixed picture. For visual modality models, the xResNet50 Frames (Spearman p=0.105) slightly outperforms the BayesianRidge Dense121 model (Spearman p=0.053). In the textual category, the AWD-LSTM Captions model (Spearman p=0.043) performs marginally better than its augmented version (Spearman p=0.02). Intriguingly, for long-term memorability, the best performing model is within the auditory category, the xRestNet50 Audio Spectrograms model (Spearman p=0.113), albeit with relatively low scores. This table's results, coupled with unusually low Spearman scores, hint at potential data quality issues within the TRECVid dataset. Table 4.2 presents the results for the Memento10k validation set where we observe a marked improvement in results. In the visual modality, BayesianRidge Dense121 (Spearman 0.524) leads, closely followed by the xResNet50 Frames model (Spearman 0.446). Although the textual modality shows an improvement

Table 4.2: Results for models tested on the Memento10k test set.

Approach	**Short-Term (Spearman)**
Visual	
BayesianRidge Dense121	**0.524**
xResNet50 Frames	0.446
Textual	
AWD-LSTM Captions	0.423
AWD-LSTM AUG Captions	0.410
Auditory	
xRestNet50 Audio Spectrograms	0.2030
BayesianRidge VGGish	0.2913
MediaEval2021 Participants	
GTHUPM [213]	0.656
Erika [214]	0.628
HCMUS [215]	0.516
AIMMLAB [216]	0.648
MeMAD [212]	0.658

from TRECVid results with AWD-LSTM Captions (Spearman 0.423), it still lags behind the visual models. Interestingly, the auditory modality, with xRestNet50 Audio Spectrograms (Spearman 0.203) and BayesianRidge VGGish (Spearman 0.291), lands at the bottom of this set, suggesting that audio might not be as influential on this dataset.

Table 4.3 provides insight into model generalisability, containing the results of the models trained on Memento10k and then tested on the TRECVid test set. Across all modalities, we observe a significant drop in performance when models trained on one dataset are tested on the other, indicative of the models' limited generalisability. For instance, the BayesianRidge Dense121 model in the visual modality falls from Spearman 0.524 (Memento10k test set) to 0.256 (TRECVid test set).

Drawing from these results, we can discern certain patterns that emerge. The visual modality consistently outperforms other modalities in both datasets, evidenced by the higher Spearman scores in Table 4.2 (Memento10k test set) and the relative performance in Table 4.3 (TRECVid test set). This trend is observed even when the

Table 4.3: Results for models trained on Memento10k and tested on the MediaEval2021 TRECVid test set.

Approach	Short-Term (Spearman)
Visual	
BayesianRidge Dense121	0.256
xResNet50 Frames	0.132
Textual	
AWD-LSTM Captions	0.114
AWD-LSTM AUG Captions	0.106
Auditory	
xRestNet50 Audio Spectrograms	0.018
BayesianRidge VGGish	0.021
MediaEval2021 Participants	
AIMMLAB [216]	0.091

models are transitioned from the Memento10k to the TRECVid dataset, indicating a consistent pattern of the visual modality's dominance in memorability prediction. Interestingly, despite the significant drop in absolute performance when moving from the Memento10k validation set to the TRECVid test set, the relative performance across different modalities remains consistent. This consistency suggests that the underlying factors affecting the predictability of memorability across modalities hold stable, irrespective of the dataset. Therefore, while data quality issues[1] with the TRECVid dataset may have affected the absolute scores, they do not appear to disrupt the relative pattern of modality performance. Taken together, these findings underscore the power of the visual modality in memorability prediction while also revealing enduring challenges in developing models that perform consistently well across different datasets.

4.5 Multimodal Video Memorability Prediction

We evaluate the utility of including the audio modality in short-term video "recognition memorability" prediction, and assess a proposed gestalt based video memo-

[1] See Appendix 1 for detailed discussion of data quality issues with the TRECVid dataset

rability prediction system[2] by benchmarking it on the Memento10k dataset [192], comparing it to state-of-the-art solutions[3]. The contributions are two-fold: A) we assess the influence of the audio modality on video memorability, and B) we propose a multimodal deep learning-based late fusion system that uses audio gestalt to estimate the influence of the audio modality on overall video memorability, and selectively leverage audio features accordingly.

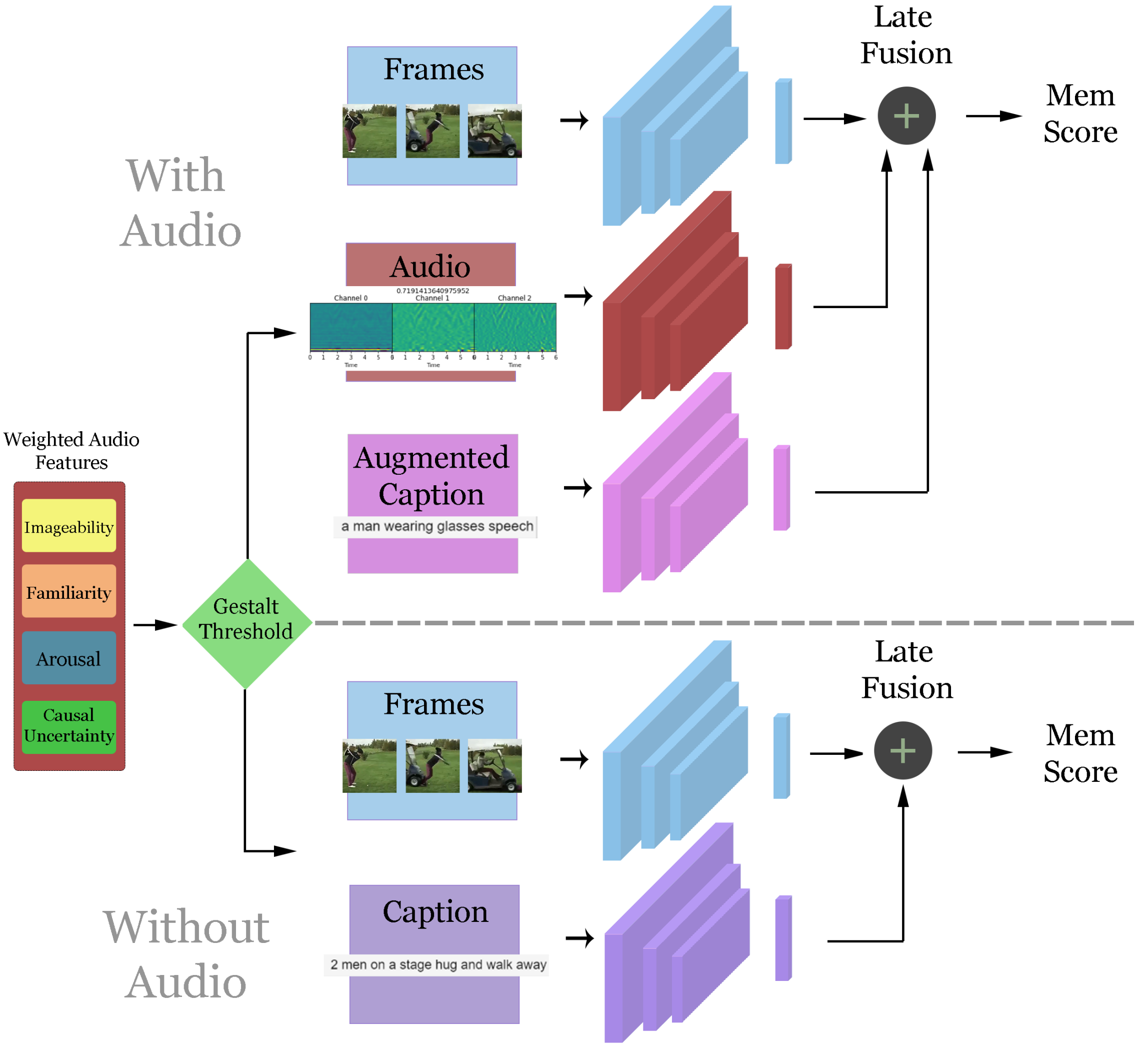


Figure 4.1: Multimodal deep-learning based late fusion framework, using a conditional audio gestalt based threshold.

The system is a multimodal deep-learning based late fusion framework that uses an audio gestalt conditional mechanism to predict short-term video recogni-

[2] https://github.com/lorinsweeney/audio_gestalt_video_memorability

[3] This work was conducted in 2020 before the aforementioned MediaEval2021 task, and the only Spearman's rank score published was p=0.663, by the authors of the Memento10k dataset

tion memorability (Figure 4.1). Depending on an audio gestalt threshold (0.8), one of two pathways—*without audio*, using textual and visual features; and *with audio*, using textual, visual, and auditory features—is used to predict a video's recognition memorability score. The *without audio* stream's predictions are the weighted sum of the *Frame* model (0.38), and *Caption* model (0.62), while the *with audio* stream's predictions are the weighted sum of the *Frame* model (0.4), *Augmented Caption* model (0.47), and *Spectrogram* model (0.13). Both the weightings of the models' predictions and the gestalt threshold are determined using Randomised Search Cross-Validation (RSCV) from 0 to 1, in increments of 0.01.

4.5.1 Audio Gestalt

The Gestalt principles (Figure 4.2) were first introduced by German Psychologists Max Wertheimer, Kurt Koffka and Wolfgang Kohler in 1928 [217], and continue to be relevant in modern psychology. Traditionally thought of as rules that characterise the organisation of visual scenes—helping us understand them better—the Gestalt principles of *similarity*; *connectedness*; *common region*; *spatial proximity* [218], and *goodness* [219] have been shown to benefit visual recognition memorability.

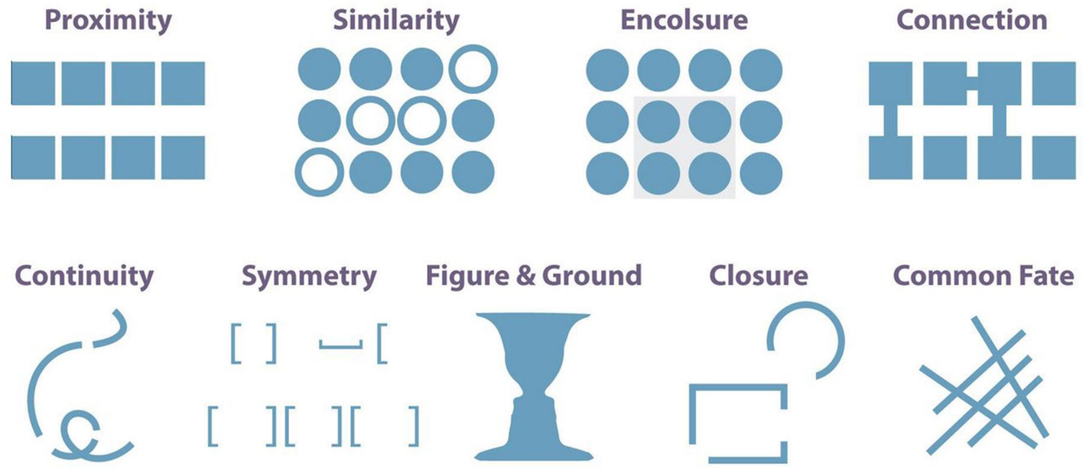


Figure 4.2: Gestalt principles of visual organisation

The very first usage of the term Gestalt was in 1890 in [220], which observed that humans can recognise two identical melodies even when no two corresponding notes have the same frequency. It was suggested that this property indicated the presence of a “Gestalt quality”—a conceptual characteristic that assists our “big picture” understanding of complex sensory data composed of many different parts. Unfortunately, since then, few insights intersecting audio gestalt and other well established audio properties have been discovered. The concept of gestalt in the context of audio was recently reintroduced by [185], using the term gestalt to encapsulate high-level conceptual audio features. They found the following gestalt features: imageability (the ease with which a mental image of the audio source can be evoked); human causal uncertainty (Hcu, the degree of ambiguity or confusion experienced regarding the cause or origin of a specific sound); arousal (the emotional intensity of the audio); and familiarity (how frequently the audio is heard, e.g., your doorbell), to be strongly correlated with audio memorability. I aim to practically apply these findings with the goal of elucidating the role of audio in overall video recognition memorability.

An audio gestalt predictor was built using a weighted sum of the proxy measures for these four features. RSCV between 0 and 1 in increments of 0.05 were used to determine each of the weights. Due to the strong negative correlation between sound imageability and musicality [221], imageability was predicated on whether the audio is classified as music or not. The PANNs [202] network was used to generate audio-tags, labelling the audio as music (giving it a score of 1.0) if a musical tag is present in the top 75% confidence. Hcu and arousal scores were independently predicted with ImageNet-pretrained xResNet34 models fine-tuned on spectrograms from the HCU400 dataset [179]. Due to limited available options, for familiarity, the top audio-tag confidence score of the PANNs [202] network (a large-scale pre-trained audio classification neural network) was used as a proxy (Spearman = 0.305, pval = 4.749e-10 between the two scores in the HCU400 dataset). These four scores were then normalised (scaled into a 0-1 range), and a weighted score (with weights of 0.2, 0.2, 0.2, and 0.4 respectively) was calculated to produce an audio gestalt score.

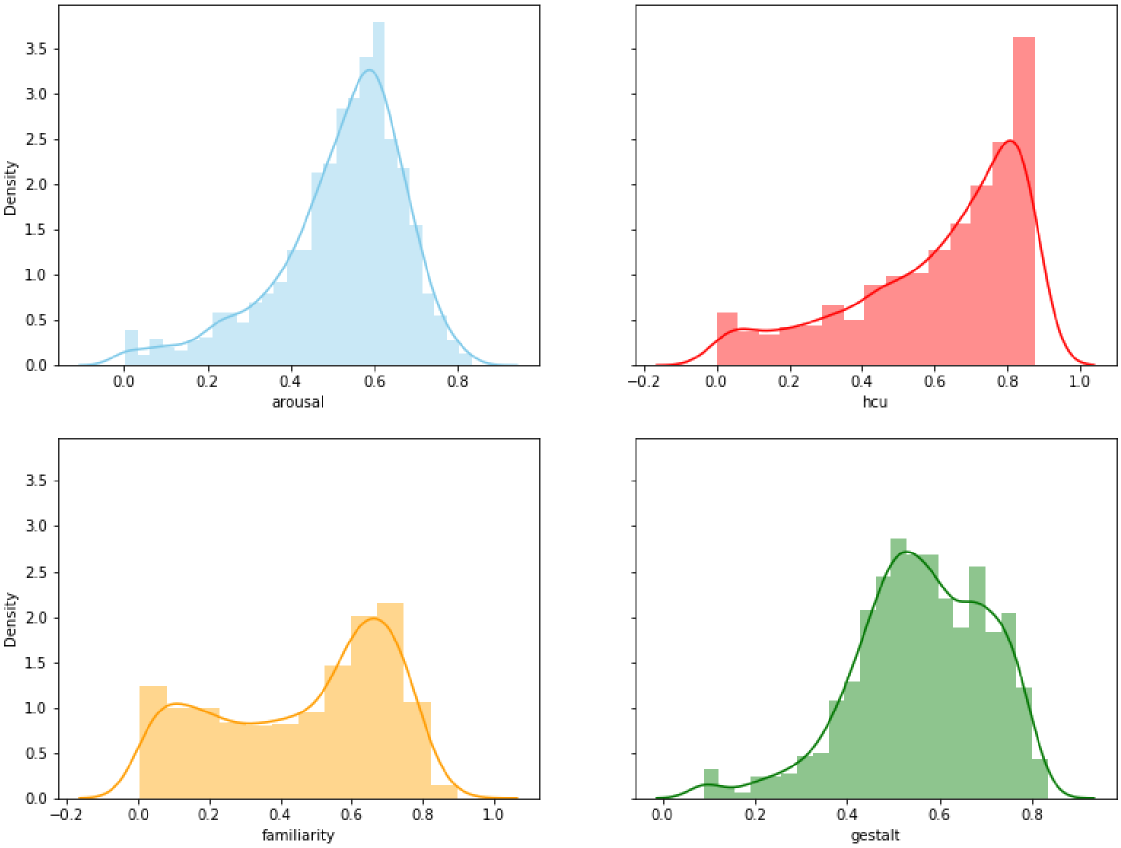


Figure 4.3: Distribution of gestalt related audio features from 1,468 validation videos.

4.5.2 Auditory Features

For auditory features, a network was trained to predict a video's recognition memorability from audio spectrograms—the *Spectrogram model*. Mel-frequency cepstral coefficients (n_fft:2048, hop_length:256, n_mels:128) were extracted from the 6,890 Memento10k [192] training videos with audio, and stacked with their delta coefficients in order to create three channel spectrogram images. These spectrogram images were then used to train an ImageNet-pretrained xResNet34 model for 15 epochs with a max learning rate of 1e-2 and weight decay of 1e-3 to predict audio recognition memorability. Additionally, a Bayesian Ridge Regressor was fitted with VGGish [203] audio features—the *Bayesian Ridge* model. 128-dimensional embeddings for each second of video audio were extracted, resulting in a 384-dimensional feature set per video.

4.5.3 Visual Features

We evaluated the extent to which static visual features contribute to video recognition memorability by training a network to predict a video's recognition memorability from the first frame—the *Frame* model. We trained an ImageNet-pretrained xResNet50 to predict image recognition memorability by first training on the LaMem dataset [198] for 50 epochs with a maximum learning rate of 3e-2 and weight decay of 1e-2, and then fine-tuning on those 6,890 Memento10k [192] training videos which have audio with the same hyperparameters. At test time, a video's recognition memorability score was calculated by averaging predictions of the first, middle, and last frames.

4.5.4 Textual Features

For textual features, we trained a network to predict a video's recognition memorability from a paragraph of text composed of five captions generated by five independent humans—the *Caption model.* Given that overfitting is a primary concern (due to limited variability inherent in short captions), we used the AWD-LSTM (ASGD Weight-Dropped LSTM) architecture [199], as it is highly regularised and is comparable to other state-of-the-art[4] language models. In order to take full advantage of the high level representations that a language model offers, the model was transfer trained using UMLFiT [200], a method that uses discriminative fine-tuning (applying varying learning rates for different layers of a neural network), slanted triangular learning rates (a learning rate scheduling technique that employs a short initial increase followed by a longer gradual decrease in the learning rate), and gradual unfreezing (layers are sequentially "unfrozen" and fine-tuned) to avoid catastrophic forgetting.

[4] At the time of these experiments, circa October 2020, the chosen architecture was SOTA, but that is no longer the case, and there are much more capable models available

A Wiki-103-pretrained language model was fine-tuned on the first 300,000 captions from Google's Conceptual Captions dataset [201] for a total of 10 epochs with a dropout multiplier of 0.5 and max learning rate of 2e-3, resulting in a final language model accuracy of 37% (on the task of predicting the next word in a caption). The encoder from that model was re-used in another model of the same architecture, but trained on captions from the 6,890 Memento10k [192] training videos with audio, for a total of 15 epochs with a dropout multiplier of 0.8 and a max learning rate of 1e-3, to predict recognition memorability scores, rather than the next word in a sentence. An additional network was trained the same way, but fine-tuned on captions that were augmented with audio tags extracted using the PANNs [202] network—the *Augmented Caption model.*

In all cases, models were independently trained on those 6,890 Memento10k training set videos with audio, and independently validated on those 1,484 Memento10k validation videos with audio. All parameter tuning (e.g. RSCV) was performed using the Memento10K training set.

4.5.5 Results

Table 4.4 shows the Spearman rank correlation scores of the individual components of the audio gestalt system, and many of their combinations, and the final implementation of the audio gestalt system on those 1,484 Memento10k validation videos with audio. The best performing individual component was the *Caption model*, achieving a Spearman score of 0.5710. Each of the component combinations are the result of a randomised search weighted summation of their predictions, with the best straightforward combination being *Captions + Frames* (p=0.6175). The audio gestalt based system was the best overall performing approach, achieving a Spearman score of 0.6181.

To evaluate the effectiveness of the approach, it was compared against the Memento10k benchmark scores[3] [192]. From Table 4.5 we can see that the audio gestalt based approach outperforms all other approaches except SemanticMemNet

Table 4.4: Results on 1,484 Memento10k validation videos with audio.

	Memorability
Approach	**Spearman**
Spectrogram	0.2030
Bayesian Ridge	0.2913
Frames	0.4808
Frames + Spectrogram	0.4876
Frames + Bayesian Ridge	0.4992
Captions	0.5710
Captions + Spectrogram	0.5715
Captions + Bayesian Ridge	0.5741
Augmented Captions	0.5555
Augmented Captions + Spectrogram	0.5562
Augmented Captions + Bayesian Ridge	0.5576
Augmented Captions + Frames	0.6068
Captions + Frames	0.6175
Everything Ridge	0.6066
Everything Spectrogram	0.6061
Audio Gestalt Ridge Normal Captions	0.6175
Audio Gestalt Spectrogram Normal Captions	0.6176
Audio Gestalt Ridge	0.6181
Audio Gestalt Spectrogram	**0.6181**

[192]—the framework introduced alongside the Memento10k dataset, which employs a three-stream encoder, processing distinct input modalities: raw frames, the video considered as a cohesive 3D unit, and the 3D optical flow derived from the video. Human Consistency, as mentioned previously in chapter 3, is the Spearman's rank correlation between two random halves of the ground-truth human memorability annotations. MemNet is typically used as a baseline by averaging its predictions across a frame sampling of 1 frame per second [192], and serves as a good point of initial comparison.

With respect to the results in Table 4.4, the general trend for predicting video recognition memorability seems to be that the more modalities used, the better the predictions. Even the addition of a poorly-performing individual audio model (0.2913) with a better-performing individual visual model (0.4808), produces an increase in performance (0.4992). There are however, some very important exceptions to this trend. Indiscriminately tri-modal approaches (those which simply employ the

Table 4.5: Comparison of state-of-the-art on Memento10k. *Trained and validated on fewer videos due to audio constraint, 7,000 vs. 6,890 and 1,500 vs. 1484 respectively.

	Memorability
Approach	**Spearman**
Human Consistency	0.730
MemNet Baseline [198]	0.485
Cohendet et al. (Semantic) [222]	0.552
Cohendet et al. (ResNet3D) [222]	0.574
Feature Extraction + Regression (as in [223])	0.615
SemanticMemNet [192]	0.663
Audio Gestalt	**0.618***

with audio stream of the framework irrespective of audio gestalt scores), *Everything Ridge* (0.6066) and *Everything Spectrogram* (0.6061), achieve lower Spearman scores than the bi-modal combination of visual and textual predictions (0.6175), and their selectively tri-modal counterparts (0.6181).

At first glance, it appears that augmenting captions with audio-tags performs worse than vanilla captions, Augmented Captions (0.5555) vs. Captions (0.5710); Augmented Captions + Spectrogram (0.5562) vs. Captions + Spectrogram (0.5715); Augmented Captions + Bayesian Ridge (0.5576) vs. Captions + Bayesian Ridge (0.5741); Augmented Captions + Frames (0.6068) vs. Captions + Frames (0.6175), however, when selectively used in the audio gestalt system (0.6181), they outperform vanilla captions (0.6175).

4.6 Discussion

Aside from SemanticMemNet, the audio gestalt based system outperforms all of the other tested approaches. Even though the advantage incurred is only marginal, selectively including audio features (0.6181) is ultimately better than both always including them (0.6066), and not including them (0.6175). I believe that this can in part be explained by the fact that sounds have the potential to provide valuable contextual priming information [184], but that some sounds simply add noise, having a deleterious effect on overall understanding of a context. Thinking of audio gestalt as an ontological property that encapsulates high-level auditory features that positively contribute towards the understanding of a context, helps explain the benefit of using it as a measure to discriminate between useful and distracting audio in multimodal content. The effect of different gestalt thresholds is shown in Figure 4.4.

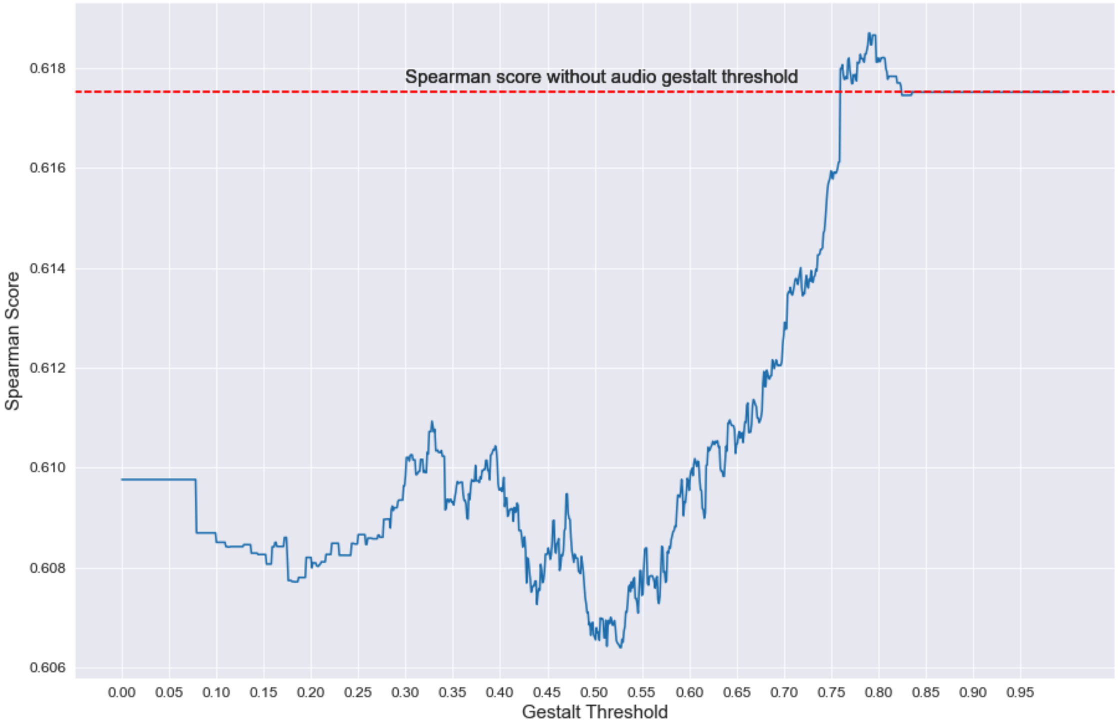


Figure 4.4: Effect of gestalt thresholds on Spearman scores of 1,468 Memento10k validation videos.

It is interesting to note that there is no difference in Spearman score between Audio Gestalt Spectrogram (0.6181) and Audio Gestalt Ridge (0.6181), even though the Bayesian Ridge achieves a noticeably higher Spearman score (0.2913) than the Spectrogram model (0.2030). This indicates that the inclusion of auditory features is not strictly additive, and further suggests that they may act as a contextual signal of some sort.

In [185], the authors found that the strongest predictors of sound recognition memorability were imageability, and causal uncertainty (Hcu). Naturally, we would expect the audio gestalt weightings to reflect this to some degree, but we found that the highest weighted audio gestalt feature is familiarity (top audio-tag confidence score). The gestalt weightings for imageability; Hcu; familiarity; and arousal, are 0.2; 0.2; 0.4; 0.2 respectively. As shown in Figure 4.3, familiarity is the only audio gestalt feature with a bi-modal distribution. Both arousal and Hcu are heavily left skewed, leading us to believe that the models used to predict their scores have been overfit, and accordingly, there is considerable room for refinement and improvement.

4.7 Conclusion

As the exploration of the complex threads of multimodal memorability concludes, they unspool to reveal valuable insights. There is some empirical support for the first hypothesis (H1), suggesting that memory and memorability are inherently multimodal. The interaction of sensory modalities is evident in the performance of the audio gestalt regulated deep learning-based late fusion system, with its selective use of audio features fortifying the prediction of short-term video memorability, even when originating from an independently less successful audio model. The audio modality's influence on video recognition memorability positions it as a contextual element with potential as a recognition aid, depending on the extent of its high-level features. These findings strengthen the proposition that recognition memorability is closely linked with high-level perceptual properties of content, rather than with low-level properties, and that this relationship extends beyond the visual domain.

Nonetheless, understanding of the full extent of the audio modality's role in short-term video recognition memorability remains incomplete. The interplay between a video's auditory and visual content might significantly impact its overall memorability. To fully comprehend this interaction, further experimental investigation is necessary, including refining the measure of audio gestalt. Future research should focus on generating independent memorability scores for each modality—audio, visual, and textual. These metrics could elucidate the roles each modality plays within a multimodal medium like video.

The second hypothesis (H2) examined the role of visual sensory data in multimodal memorability prediction. This hypothesis was inspired by the established prominence of the visual cortex in how humans interact with the world. The investigation tentatively confirmed the substantial influence of the visual domain, while simultaneously highlighting the powerful roll of textual features. However, it is conceivable that the difference between domains can be accounted for by the quality of their encoding method. The textual results suggest that when semantics are isolated, such as in textual descriptions or captions, they can account for a significant portion of a video's memorability without direct sensory perception of the video content itself. Thus, while the findings give weight to the influence of the visual domain, they simultaneously hint at the potentially amodal nature of memorability, suggesting its less of a direct product of single modality, but a fine interplay of sensory inputs, and possibly pointing to an abstraction that sits beyond the grasp of senses.

Chapter 5

Neurally Informed Measures of Memory

> *"The moments of the past do not remain still; they retain in our memory the motion which drew them towards the future, towards a future which has itself become the past, and draw us on in their train."*
> — Marcel Proust, *In Search of Lost Time*

Although the nature and constitution of people's memories remains elusive, and our understanding of what makes one thing more or less memorable than another is still nascent, combining computational (e.g., machine learning) and neurophysiological (e.g., electroencephalography; EEG) tools to investigate the mechanisms of memory has the potential to offer insights that would be otherwise unobtainable. While EEG is not a tool that can directly explain the factors that make an experience more or less memorable, it can help us trim the umbral undergrowth surrounding the subject, and offer a potential leap forward in our understanding of the interplay between the mechanisms of memory and memorability.

The use of EEG has proven to be an effective tool in the investigation of the neural mechanisms that underpin memory formation and recall. Its ability to non-invasively capture real-time neural activity provides insight into the timing and general location of these cognitive processes [224]–[227]. Even though the application of machine learning to EEG is an active area of interest—enabling the classification of various cognitive states and processes, such as emotion [228], and mental tasks [229], and sleep stages [230], and the automation or augmentation of neurological diagnostics [231]–[234]—the use of EEG to predict visual memorability has been limited to static stimuli such as images [235], leaving video entirely unexplored.

5.1 The Subsequent Memory Paradigm

The subsequent memory paradigm, a more direct measure of memory performance, heralds a natural evolution of the traditional behavioural and introspective recall experimental paradigms outlined in chapter 4. It rests on the principle of identifying neural activity associated with successful memory formation at the time of experience, thereby providing unique insights into the encoding processes that determine whether an event will be remembered or forgotten. Here, we delve into the intricacies of this paradigm, reflecting on its benefits, limitations, and key findings, while situating it within the broader narrative of memory research.

The subsequent memory paradigm has its roots in Event-Related Potential (ERP) studies. The method involved recording electrical brain activity via EEG while participants were presented with a series of stimuli. Later, the stimuli were sorted into those that would be remembered versus those that would be forgotten, based on participants' performances on a subsequent memory test—any recognition or recall task [225], [226]. A shift was seen with the advent of functional Magnetic Resonance Imaging (fMRI), where higher spatial resolution allowed for the identification of specific regions associated with successful memory encoding, notably the Medial Temporal Lobe (MTL) and the Prefrontal Cortex (PFC) [66], [67]. The use of the subsequent memory paradigm offers several benefits. Firstly, it provides a more direct measure of memory performance than other techniques. Instead of relying on participants' ability to recall or recognise stimuli, it looks at the neural activity associated with successful memory formation. This allows for an examination of the encoding processes that lead to successful memory formation, offering a more complete picture of how memory works. Moreover, this paradigm allows for the examination of individual differences in memory performance. By comparing neural activity during successful and unsuccessful memory encoding, it is possible to identify the neural correlates of these individual differences, which has far-reaching implications for understanding memory disorders and cognitive decline.

However, the subsequent memory paradigm is not without its limitations. One key challenge is the need for large amounts of data. Due to the inherent variability in neural responses and memory performance, many trials with participants are needed to obtain reliable results. Moreover, identifying the precise temporal dynamics of memory encoding processes can be difficult with techniques like fMRI, which have excellent spatial resolution but relatively poor temporal resolution. Several key findings have emerged from studies using the subsequent memory paradigm. For instance, it has been found that neural activity in the PFC and MTL is higher for stimuli that are later remembered than for those that are later forgotten [66], [67]. This suggests that these regions play a critical role in memory encoding. In addition, recent studies have shown that memorability and subsequent memory show dissociable neural substrates, with memorability effects consistently emerging in the MTL, and individual subsequent memory effects in the PFC [236]. The subsequent memory paradigm represents an evolution from traditional recall experiments, offering a unique window into the neural underpinnings of memory formation. Despite the challenges that it presents, this paradigm holds great promise for the future of memory research, providing unprecedented insights into the complex mechanisms that transform our experiences into lasting memories.

5.1.1 EEG Data Acquisition

In a typical subsequent memory experiment, participants are exposed to a series of stimuli during an initial encoding phase while their brain activity is recorded using EEG. Data acquisition is commonly carried out using two computers, one to run the experiment code and present the stimuli, and the other for recording and monitoring participant EEG data. A typical setup is shown in Figure 5.1. The stimuli are generally simple words or images, selected according to the research question at hand. When displaying stimuli sequences to participants, a timestamp for each stimulus must be recorded and aligned with the multi-channel time-series EEG recording captured on the acquisition computer. These timestamps are commonly referred to as markers. The EEG recording process begins with the placement of electrodes on the scalp, a setup often guided by the International 10-20 system that ensures standardised electrode placement across subjects and studies. During the encoding phase, EEG data is continuously recorded and later divided into epochs corresponding to individual stimuli presentations. An epoch typically includes a pre-stimulus baseline period and extends to several hundred milliseconds or even seconds post-stimulus, covering the full duration of the brain's response to the presented stimulus [20].

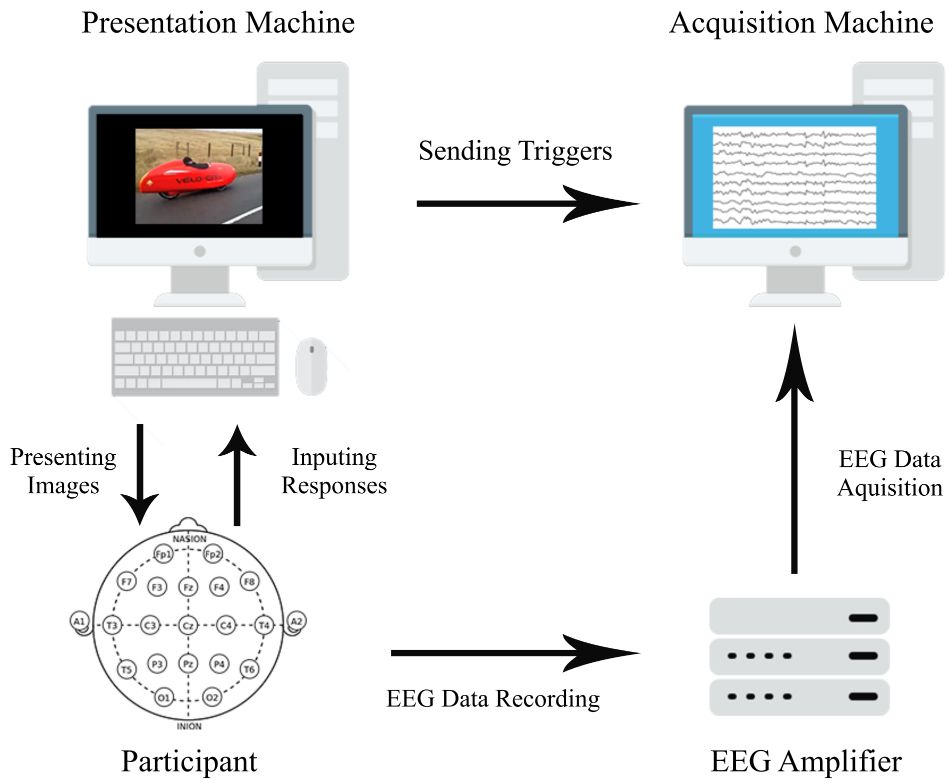


Figure 5.1: Typical EEG data acquisition setup.

5.1.2 ERP Component Analysis in Subsequent Memory Experiments

Following initial EEG data acquisition, researchers turn their attention to identifying and analysing distinct ERP components. ERPs are examined through the averaged time-locked EEG responses to the onset of stimuli and provide an invaluable means of probing the temporal dynamics of cognitive processes. The ERP components of primary interest in subsequent memory experiments include the P300, the Late Positive Component (LPC), and the FN400, all of which have been consistently associated with memory encoding and retrieval processes. The P300 component is a positive deflection in voltage that occurs approximately 300 milliseconds after the presentation of a stimulus, particularly when that stimulus is unexpected or infrequent [237]. The P300 is thought to reflect cognitive processes such as attention and memory, with larger P300 amplitudes generally associated with more cognitive resources being devoted to processing the triggering stimulus [238].

The P300 component is comprised of sub-components, namely, P3a and P3b. The P3a component is often associated with the processing of novel or unexpected stimuli and is thought to reflect the orienting of attention towards these stimuli [237]. It is typically observed as a fronto-central positivity occurring approximately 250-280 milliseconds after stimulus onset. The P3a is thought to be generated by the frontal cortex, especially the prefrontal and anterior cingulate cortices, suggesting a role in the evaluation of stimulus saliency and decision-making [239]. On the other hand, the P3b sub-component is more related to the process of stimulus evaluation and categorisation. It emerges as a parietal positivity between 300-500 milliseconds post-stimulus [240]. The P3b is thought to reflect the updating of working memory representations and the allocation of attentional resources, as larger P3b amplitudes have been correlated with improved performance on memory tasks [237]. The other noteworthy ERP component in the context of memory and the subsequent memory paradigm is the Late Positive Component (LPC). This component, which peaks 500-800 milliseconds post-stimulus, has been found to be associated with the recognition of previously presented items, thus suggesting a role in the depth of memory processing [241]. Studies have also demonstrated that larger LPC amplitudes correspond to better recognition memory performance, reinforcing the link between the LPC and episodic memory [242]. The FN400, a distinct ERP component, is also particularly intriguing due to its clear association with recognition memory processes. This component, named for its characteristic frontal negativity peaking around 400 milliseconds after stimulus onset, has been directly linked with familiarity-based recognition processes. Familiarity-based recognition, a key aspect of our memory system, allows us to identify an item or stimulus as having been encountered before, even in the absence of any specific recall of the details surrounding the original encounter [243]. In ERP terms, the FN400 manifests as a less negative-going waveform for items previously experienced, compared to new items. This effect, typically observed over fronto-central electrode sites, is often taken as an index of the engagement of familiarity processes during recognition [244]. Intriguingly, FN400 effects are often observed irrespective of whether recognition is accompanied by recollection, the ability to retrieve specific contextual information about an earlier episode, suggesting that the component indexes processes relatively independent of recollection [245]. Furthermore, a wealth of studies employing different methodological approaches, including manipulations of memory strength and tests of associative recognition, have provided converging evidence that the FN400 reflects a relatively automatic process sensitive to the perceptual and conceptual overlap between a test cue and memory representations formed during study [246].

5.1.3 Data Pre-processing

Pre-processing of some nature is generally a required precursor to any meaningful interpretation or use of EEG data. It not only prepares the data for subsequent analysis, but also plays a significant role in enhancing the reliability of the findings. The pre-processing pipeline typically involves re-referencing, filtering, epoching, trial/channel rejection, and artifact removal. Initially, we re-reference EEG data, which involves redefining the EEG signal with respect to a new reference electrode or calculated average. The aim is to minimise the influence of the reference on the acquired signal and to provide a more accurate estimation of the neural activity at each electrode [247]. The choice of reference (e.g., common average reference, mastoid electrodes, or nose tip electrode) largely depends on the research question at hand, and the topology of the neural activity being studied. Filtering the signal is the next step. EEG data is generally filtered to retain the frequency band of interest and to exclude other signals and noise. For instance, a band-pass filter could be used to exclude slow drifts (low-frequency noise) and high-frequency noise like muscle activity. Notably, muscle-related artifacts are substantial, owing to the uncontrolled contraction of facial, neck, and scalp muscles by a subject. These are mainly present in higher frequencies (above 20 Hz) and can be effectively minimized with an appropriate high-frequency cut-off [248]. Following filtering, the continuous EEG data is divided into epochs, which are smaller time windows corresponding to specific events or stimuli. Epoching facilitates the examination of the EEG data time-locked to these events, a crucial step for subsequent memory studies, which allows the study of ERPs and oscillatory dynamics around the event [20]. The process of trial/channel rejection is an integral part of EEG data pre-processing. It involves the identification and exclusion of corrupted epochs or channels from further analysis. These could arise due to various reasons such as equipment malfunction, subject movement, or excessive physiological noise. Automated algorithms, as well as manual inspection, are typically employed for this task [249]. Finally, artifact removal is another critical step. One common type of artifact in EEG data arises from eye movements and blinks. Given that these eye-related artifacts have a distinct spatial and temporal pattern, techniques such as Independent Component Analysis (ICA) have been developed to identify and remove these artifacts. By decomposing the EEG signal into spatially fixed, temporally independent components, ICA can effectively separate neural signals from artefactual signals, which can then be excluded from the data [250].

Before delving into feature extraction from EEG data, it is vital to address some intrinsic characteristics of EEG signals that could influence the design of an effective predictive system:

- High-dimensionality: EEG data is high-dimensional because it encompasses multiple channels (spatial dimension) and numerous time points (temporal dimension). This vast space introduces a major challenge for model training as it can lead to overfitting, where the model learns the noise along with the signal in the training data;

- Limited training sets: EEG subsequent memory paradigms typically involve a restricted number of participants due to the demanding nature of the setup. This limitation constrains the volume of available training data, and, in the high-dimensional space of EEG, it is a major hurdle;

- Overlapping epochs: If a brief inter-stimulus interval (ISI) is used in a subsequent memory paradigm, there can be overlap between adjacent epochs. This overlap can obscure the distinctiveness of memory-related ERPs, thereby confounding the feature extraction process;

- Low Signal-to-Noise Ratio (SNR): EEG data is characterised by a low SNR where the subtle memory-related ERP-effects are often overwhelmed by ongoing EEG background brain activity. This characteristic makes it challenging to extract reliable, meaningful signals on a single-trial basis [251];

- Imbalanced datasets: In subsequent memory paradigms, the instances of remembered and forgotten stimuli are often imbalanced. This disparity introduces bias in the training of predictive models, which could lead to overestimating the performance on the majority class while overlooking the minority class [252].

Understanding and tackling these inherent properties are pivotal for building an effective subsequent memory predictive system using EEG data. Possible solutions could involve dimensionality reduction techniques, adequate data collection strategies, suitable noise reduction methods, and strategies to handle imbalanced data. An essential distinction between the subsequent memory EEG paradigm and other ERP paradigms lies in the necessity for single-trial detection. Conventionally, ERP analysis computes grand average ERPs by averaging phase-locked activity across trials. This process leads to the attenuation of non-phase-locked or background EEG activity, thus improving the signal-to-noise ratio (SNR) of the EEG system [20]. In paradigms like the P300 speller, a desired symbol is represented several times, and epochs corresponding to each row or column are averaged across trials, further reducing the influence of random background EEG oscillations. However, in the subsequent memory paradigm such an approach is not feasible as each presented stimulus should be unique if it is not a target, and only repeated once otherwise. Consequently, subsequent memory paradigms rely on the detection of memory-specific ERPs on a single-trial basis. This methodology introduces an added challenge, as the inherently low SNR of EEG data hampers the reliable detection of discriminative ERP activity [243]. Therefore, the subsequent memory paradigm's need for single-trial detection underlines the unique difficulties and requirements associated with utilising EEG for memory prediction.

5.1.4 Performance Evaluation Metrics

In assessing the predictive power of EEG-based machine learning models, especially within the confines of subsequent memory experiments, a bouquet of performance metrics have been cultivated, each offering unique perspectives on model efficacy. Of these, two in particular emerge as the most popular and robust to target ratio imbalances.

Area Under the Receiver Operating Characteristic curve (AUC-ROC) is the most widely used evaluation metric for a wide array of EEG based machine learning research [228], [233], [253]. It provides a visual representation of a model's ability to distinguish between classes—typically "remembered"/"forgotten" for subsequent memory paradigms. This curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds, allowing us to appraise a model's discriminative capacity across these thresholds. A model with perfect discriminatory power would have an AUC of 1. Conversely, an AUC of 0.5 suggests the model performs no better than a random classifier [254]. Its utility extends to the inherent imbalance present in subsequent memory experiments, as AUC-ROC remains immune to changes in class label distributions. Nevertheless, while the AUC-ROC provides an excellent overview of a model's performance, it can obscure nuances in model behaviour at different thresholds, which may be crucial in certain applications [255].

Balanced Accuracy (BA), another noteworthy metric, emerges as an antidote to the limitations of traditional accuracy in the presence of imbalanced classes. The traditional accuracy measure, the ratio of correct predictions to total predictions, tends to exaggerate the performance on the majority class. The BA metric, however, takes the arithmetic mean of sensitivity and specificity, where sensitivity ($\frac{\text{true positives}}{\text{true positives}+\text{false negatives}}$) is the proportion of actual positives correctly identified, and specificity ($\frac{\text{true negatives}}{\text{true negatives}+\text{false positives}}$) is the proportion of actual negatives correctly identified. This offers a more representative measure of performance across classes, an attribute crucial in the context of EEG subsequent memory paradigms [256]. Notably, BA reduces to traditional accuracy for balanced classes. Yet, BA may mask the distinction between different types of errors; a model with high False Negatives may have the same BA as a model with high False Positives, making the choice of the appropriate metric a function of the specific cost associated with different types of misclassifications in the application at hand.

5.1.5 Feature Extraction Methods

A core challenge inherent in EEG feature extraction methods is finding characteristics of the EEG signal that relate to cognitive responses of interest. Feature extraction plays a significant role in EEG research as it can drastically alter the

SNR and classification strategies employed, and ultimately determines the overall performance of a system.

Spatial Features

Spatial features, derived from the relative positioning and activity of EEG electrodes on the subject's head, allow us to probe into the distribution and interaction of neural signals across the brain's cortical regions [257]. While these features can provide glimpses into the spatial organization of brain activity, their utility in predicting cognitive states such as memory encoding often pales in comparison to their temporal counterparts [258]. The reason lies in the limitations of EEG's spatial resolution, which can often obscure finer spatial patterns. Consequently, spatial features in EEG are often employed as preliminary steps in data pre-processing, primarily for dimensionality reduction through techniques such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) [259].

Principal Component Analysis (PCA) is a stalwart tool in multivariate data analysis, employed not only in EEG but across diverse scientific disciplines [260]. PCA operates by calculating an orthogonal transformation of the data to identify the 'principal components', the directions of maximal variance in the data. The result is a series of uncorrelated components, each successively accounting for the highest remaining variance in the dataset. This reduction in dimensionality while retaining maximal variance facilitates pattern detection and noise reduction in EEG data [260]. The downside, however, is that the resulting components may not be interpretable in physiological terms, rendering the extracted features less meaningful in the context of neuroscientific research.

Independent Component Analysis (ICA), another widely employed technique, alleviates some limitations of PCA. This method identifies statistically independent sources within EEG signals [261]. The assumption at the heart of ICA is that the EEG signals are linear mixtures of independent non-Gaussian sources. By separating these sources, ICA provides a powerful tool for isolating artifacts and components of interest within the complex EEG signals. While the interpretability of the components is still not guaranteed, ICA offers a higher probability of mapping components to underlying neuronal sources, rendering it more suitable for neuroscientific applications [261].

Time-frequency Representations

Temporal features, on the other hand, lie at the heart of EEG-based predictions. They capture the oscillatory dynamics and event-related fluctuations in EEG signals over time, providing insights into the temporal evolution of neural processes [262]. Techniques for extracting these features—such as event-related potentials (ERPs), event-related spectral perturbations (ERSPs), Fourier transforms, and wavelet transforms—enable us to examine the intricate temporal structure of EEG data [257], [263]. These techniques find their strength in their ability to delineate the complex temporal patterns of brain activity, offering a richer and more predictive set of features for modeling cognitive states [264].

Event-Related Potentials (ERPs) encapsulate the brain's time-locked response to a specific event, such as a sensory stimulus or a motor action. ERP waveforms, averaged across multiple trials, offer an enhanced signal-to-noise ratio, revealing a reliable picture of the brain's prototypical response to the task at hand [20]. Alterations in the characteristics of these waveforms, whether they occur in amplitude, latency, or shape, can shed light on the neural processes underlying cognitive functions, including memory encoding and retrieval [265]. However, ERPs require assumptions of linearity and stationarity, and the averaging process, while enhancing signal-to-noise ratio, could obscure subtle but significant variations in individual trials.

Event Related Spectral Perturbations (ERSPs) allow us to examine how spectral power within distinct frequency bands (e.g., theta, alpha, beta, gamma) change in response to a task or event [266]. Unlike ERPs, which focus on average time-domain signals, ERSPs investigate how the spectral characteristics—the event-related synchronization (ERS) or desynchronisation (ERD) in the power of specific frequency bands—evolve over time [267]. This detailed view of the modulations in oscillatory activity offers critical insights into how these oscillations and their synchrony contribute to various cognitive operations [268].

Fourier Transforms deconstruct complex signals into simpler sine waves of varying frequencies, unveiling spectral components ranging from delta (1-4 Hz) to gamma (30+ Hz) bands [257]. While powerful, Fourier Transforms overlook the non-stationary nature of EEG signals, which harbor time-dependent spectral characteristics. To address this, the Short-Time Fourier Transform (STFT) incorporates a temporal aspect to the Fourier Transform by employing a sliding window approach. This added dimension offers a local spectral view within the moving window, allowing the tracking of spectral changes over time—a critical element when considering the rapidly evolving spectral characteristics of EEG [269]. The STFT introduces an inherent trade-off between temporal and frequency resolutions, a balance determined by the window length, which calls for careful thought based on the question at hand [257].

Wavelet transformations bring an elegant solution to the resolution trade-off challenge. Unlike the STFT, which uses a fixed-sized window, the Wavelet Transform employs a window that expands and contracts, thereby enabling variable resolution at different frequency levels. The transformative power of this method lies in its adaptability: it offers high-frequency resolution at low frequencies and high temporal resolution at high frequencies, mimicking the logarithmic frequency perception of the human auditory system [270]. A myriad of wavelets—each with a distinct shape, or ‘mother wavelet’—can be chosen to match the characteristics of the EEG signal, thereby enhancing the signal-noise separation [270]. Among these, the Morlet wavelet has found widespread acceptance in EEG analysis due to its balance between time and frequency localisation. However, the choice of wavelet largely depends on the nature of the EEG signal and the particular research question. Other wavelets, such as the Daubechies, Haar, or Mexican hat, might be more suitable in different contexts.

Another fundamental approach in EEG analysis is the selection of an appropriate reference. While often overlooked, the choice of reference can dramatically affect the perceived topographical distribution of scalp potentials, thereby influencing downstream analyses [20]. The standard method of referencing, the *average reference*, uses the mean of all channels as the reference. This method is often used when the goal is to represent activity at each location relative to the overall average of activity. It works well when the number of electrodes is large enough to provide a good estimate of the average, but it can be influenced by a few noisy channels [20].

5.2 Video Memorability EEG Pilot

In the pursuit of expanding the boundaries of memorability research, we set out to explore uncharted territories by conducting the first of its kind pilot study—an exploration into the utility of EEG data within the context of video memorability prediction. This endeavor embarked with an underlying twofold motivation. Firstly, we sought to examine the potential applicability of EEG features to video memorability prediction. The human brain offers rich, dynamic data that can be captured via EEG, potentially holding subtle, yet crucial, markers of memorability. The pilot study was conceived as a stepping stone, an initial foray into investigating whether these neural signatures could be harnessed to predict video memorability or an adjacent phenomenon. The investigation, thus, not only enhances our understanding of the neurophysiological underpinnings of memorability but also uncovers novel avenues for potentially predicting the same. Secondly, the aim was to broaden the research horizons of the memorability community. By exploring EEG features in relation to video memorability, we hoped to provide an impetus for researchers from various disciplines to delve into this rich data without necessitating prior domain expertise in EEG. The endeavour was to furnish a bridge between the neurophysiological and computing research realms, fostering a vibrant cross-pollination of ideas and methods. Further supplementing these objectives, the dataset collected and processed during the study was made publicly available [1]. In doing so, we sought to foster a culture of open science, democratising access to this novel data, thereby facilitating future explorations in this burgeoning area of research.

5.2.1 Dataset

The stimuli used in the study were a subset of the subtask 1 data (i.e., the TRECVid2019 short-term video memorability prediction task) in MediaEval'2021 [271], and consists of 395 unique videos, 100 of which were designated as targets and selected to reflect the bottom and top 50 memorable videos from the dataset, 200 were designated fillers and selected to reflect the next top and bottom 100, and another 95 were designated as fillers, and selected to reflect the middle 100 memorable videos from the original set of 1500 videos (Figure 5.2). This selection framework was chosen to offer a balanced representation of the distribution of the original set.

EEG data was collected from a total of 20 subjects (resulting in a final 11 after data quality filtering) while they completed an adapted subsequent memory experiment—a short-term recognition memory game—which was used to annotate the videos for memorability. EEG data acquisition[2] was carried out in two separate locations using a common experimental procedure, and each location annotated the same set of videos. Despite the use of disparate equipment, the likelihood of experimental compromise was minimised by the consistency of all other experimental factors. The memory game began with a preparation phase, a pseudo phase carried out right before the main experiment which afforded the opportunity to explain and test equipment and settings. In this preparation phase participants were given a verbal description of the experiment procedure, then presented with a set of written instructions, and finally taken through a practice run of 3 test videos to familiarise

[1] Dataset and examples of use, as well as the code to replicate the results in this paper, are available at `https://osf.io/zt6n9/`

[2] Data collection for participants 1–5 was carried out at Dublin City University (DCU) with approval from the University's Research Ethics Committee (DCUREC / 2021 / 171), and for participants 6–11 at the University of Essex (UoE) with approval from the Ethics Committee there (ETH2122-0001). Data at DCU was collected using a 32-channel ANT Neuro eego system with a sampling rate of 1000 Hz. Data at UoE was collected using a 64-channel BioSemi ActiveTwo system at a sampling rate of 2048 Hz.

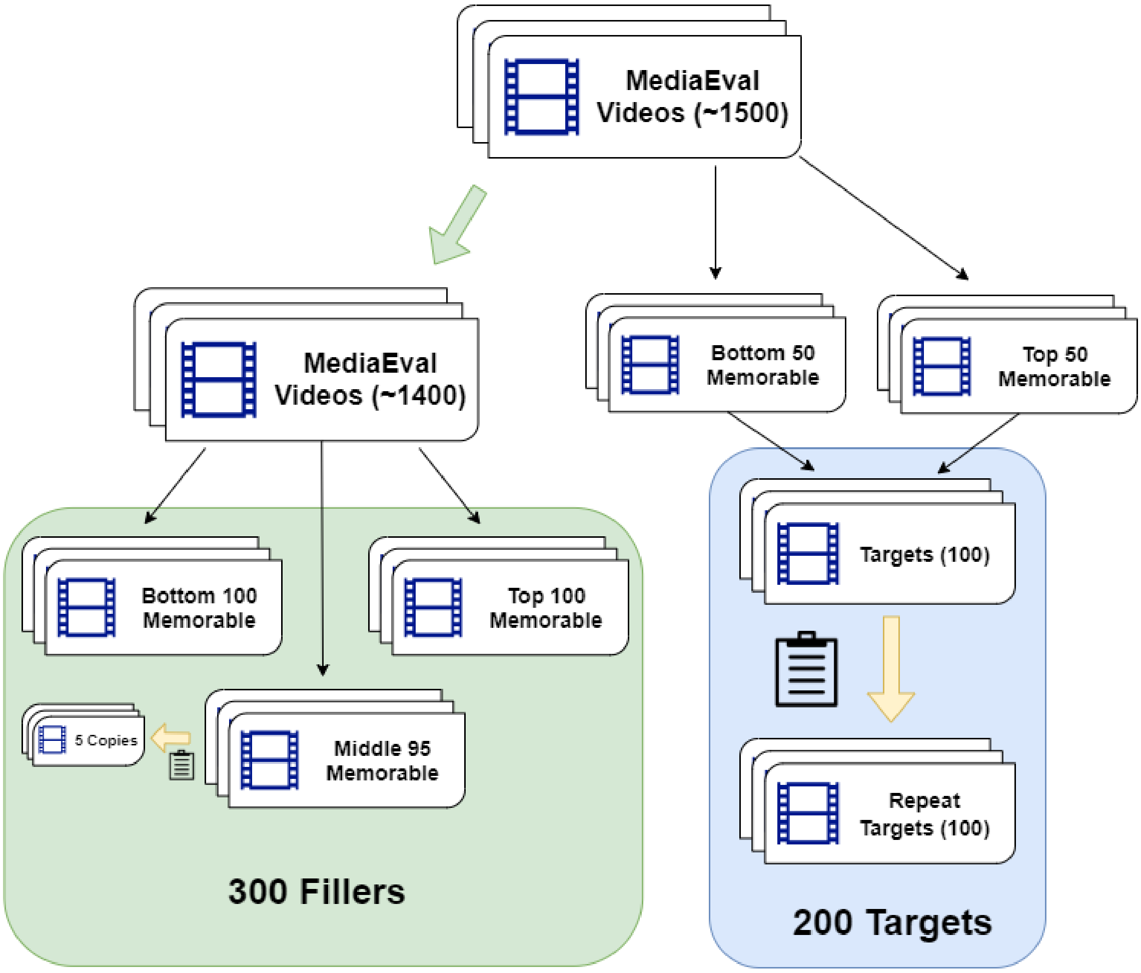


Figure 5.2: EEG video selection procedure.

them with the experiment. Rather than being split into separate encoding and recognition phases like the traditional subsequent memory paradigm, the memory game was continuous in nature—involving the presentation of a video followed by an interstimulus delay, then followed by recognition response input. This was done in an attempt to keep the data collection paradigm as close to the original "Video Memorability Game" proposed by [222]. The experiment used a total of 450 videos, 192 of which were the target videos (96 targets, shown twice), and the remaining 258 videos were the fillers. The experiment was broken into 9 blocks of 50 videos, where a fixation cross was displayed for 3–4.5s, followed by the video presentation for its ∼ 6 second duration, followed by a "get ready to answer" prompt of 1–3 seconds, followed by a 3s period for recognition response (Figure 5.3). The time per block was approximately 700 seconds (∼ 12 minutes) without accounting for 30-second closed/open eye baselines and breaks, which occurred between blocks. In order to account for recency effects, the first 50 videos presented did not include targets, but had 5 filler repeats, and the presentation positions of targets between each of the participants was pseudo-randomised, with the distances between target and repeat videos designed to fit a uniform distribution, and the position of each block aside from block 1 being shifted by 1 for each participant.

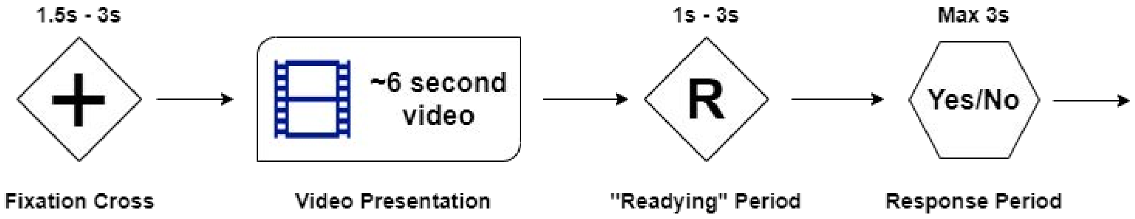


Figure 5.3: EEG video presentation procedure.

5.2.2 Methodology

A standardised processing procedure was implemented for the EEG data collected from the two distinct locations by selectively including the 28 channels common across both setups. The initial phase of pre-processing involved re-referencing the data using a common average reference, a technique that computes the average across all EEG channels and subtracts this average from each individual channel [20]. This strategy nullifies the potential influence of non-cerebral signals and reduces the likelihood of extraneous signal distortions across channels. Subsequently, the EEG data were subjected to band-pass filtering, a technique designed to allow only signals within a specific frequency range (in this case, 0.1-30 Hz) to pass through [272]. The application of a symmetric linear-phase Finite Impulse Response (FIR) filter enabled a uniform phase delay for all frequencies, mitigating signal distortion and preserving the original temporal characteristics of the EEG signals [272]. Addressing the omnipresent issue of artifacts in EEG data, Independent Component Analysis (ICA, see section 5.1.5) was employed—a statistical technique that separates multivariate signals into independent non-Gaussian components [261]. ICA is extensively used in EEG analysis for its effectiveness in isolating and removing artifact components such as eye blinks or muscle activity, thereby enhancing the overall quality of the EEG data. The implementation of trial rejection, employing subject-specific thresholds, further ensured the exclusion of trials that might compromise the data quality [20].

Once the pre-processing was complete, the focus was on the extraction of meaningful features from the EEG data. The choice of features for this pilot study was informed by two principal considerations. First, recognising the novelty of the undertaking—the exploration of EEG features for video memorability, an area hitherto uncharted—we sought to harness features well-established within the broader EEG research community. ERPs and ERSPs (see section 5.1.5), extensively utilised in numerous cognitive studies [89], [243], [266], thus, presented themselves as a sensible starting point. Second, amidst the plethora of potential EEG features at our disposal, we intended to ensure coverage of both the time and frequency domains, fostering a comprehensive insight into the temporal dynamics and oscillatory behavior of the EEG data. This strategic choice, we hoped, would facilitate the initial exploration of the utility of EEG for video memorability prediction while also providing a robust foundation for further, more specialised, investigations. First, we centered our attention on the time-domain, specifically, the extraction of ERP features. To obtain these features, the EEG data were low-pass filtered with a cutoff frequency of 15 Hz and subsequently downsampled to 30 Hz, reducing the data size and ensuring the retention of frequencies most relevant to cognitive processes [20]. Baseline correction was then applied using the 250-ms pre-stimulus interval, ensuring the minimization of non-stimulus-related EEG variance. Following this, data corresponding to the first second of each repeated video clip from each of the 28 channels were extracted and concatenated into a feature vector, yielding the ERP features. For the frequency domain, the ERSP features, we segmented the EEG data into 4-second epochs and computed a trial-by-trial time-frequency representation using

Morlet wavelets for frequencies ranging from 2-30 Hz [257]. This method provided a nuanced view of the EEG, capturing temporal dynamics and oscillatory activity within specific frequency bands. In contrast to the comprehensive channel coverage in ERP feature extraction, the ERSP analysis was restricted to four specific EEG channels (Fz, Cz, Pz, and O1). This decision was primarily informed by a strategic approach to this initial exploration of the rich, high-dimensional EEG data in the context of video memorability. The four selected channels provide comprehensive coverage of frontal, central, parietal, and occipital areas, which are broadly implicated in visual processing and memory tasks [224]. By focusing on these locations, we aimed to reduce the dimensionality of the data and concentrate on regions of interest where specific cognitive processes are anticipated to be manifested. This approach allows for a more targeted exploration, especially crucial in establishing a foundational understanding of how these complex features may correlate with the task. It is also worth noting that ERSPs, due to their encapsulation of more complex and varied spectral changes, are not as straightforward to interpret across all channels compared to ERPs [224]. Hence, limiting the analysis to these key channels also aids in enhancing the interpretability of the ERSP features.

5.2.3 Results

The results section of a study typically serves to highlight key findings, underscore significant trends and discrepancies, and facilitate understanding of the study's potential implications. However, in the case of this pilot EEG video memorability study, an unexpected but critical shift in our research question mandates a slight departure from the norm.

Due to unforeseen circumstance during the data collection process—participants remembered on average 93% of the videos—the final dataset had an very large class imbalance, which impeded our initial investigation into whether EEG data could be used to predict subsequent remembering of the videos. Class imbalance is a significant issue in machine learning, where one class heavily outweighs the

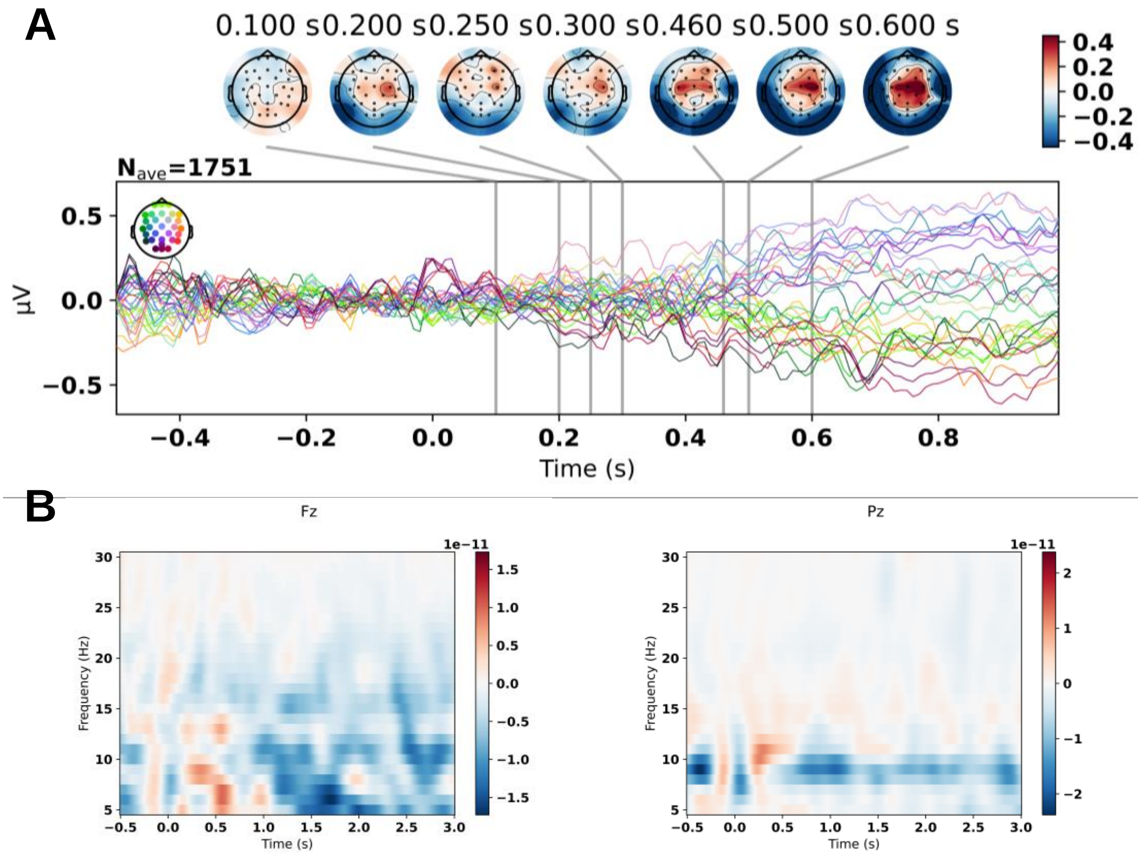


Figure 5.4: Grand-averaged butterfly plot showing differences in EEG activity for the second minus first presentation for videos for the first second (top-A). Averaged time-frequency differences in power for the second presentation minus that for the first presentation of videos for the first 3 seconds for channels Fz and Pz (bottom-B left and right, resp.).

other in terms of samples. It poses a significant challenge as most algorithms are designed to maximise accuracy and reduce error, thus they tend to be biased towards the majority class. In our case, the vast majority of videos were remembered, resulting in a dearth of examples of forgotten clips. To address this issue and make the most meaningful use of the collected data, we chose to pivot our research question. Rather than focusing on prediction of subsequent memory, we re-framed the investigation to ask, “Can we distinguish between the first and second viewing of clips that were successfully remembered, based solely on EEG data?” This shift in focus opened a new avenue of exploration within the framework of the study, focusing on the task of predicting whether an EEG sample was from a participants first viewing, labeled “unseen”, and second viewing, labeled “seen”. By comparing the EEG data from the first and second viewings of remembered clips, we aimed to uncover neurophysiological patterns that might correlate with the repetition of memorable videos. This unique focus allowed us to continue our investigation despite the unexpected class imbalance in our data, providing an innovative perspective within the field of video memorability research.

We began by standardising the data to have a mean of zero and unit standard deviation, an essential pre-processing step in machine learning to ensure that different features are on the same scale. Standardisation eliminates the influence of disparate scales between features, thereby enhancing the stability and convergence of the model [273]. To discern any significant patterns between the first and second viewings of the remembered clips, we employed a Bayesian Ridge Regressor (we turned our "seen"/"unseen"" labels into 1 and -1 respectively). This method was chosen due to its robustness and ability to handle multicollinearity, a common issue with EEG data. The regressor was implemented with the *scikit-learn* Python library, using its default parameters [195]. We applied this model independently to the two feature sets derived from the EEG data (ERPs and ERSPs) to determine their individual classification power. We adopted a 20-fold cross-validation approach, splitting the data into a 80% training and 20% testing set. This helped to ensure a robust estimation of the model's performance, and its generalisability to unseen data. The results of this cross-validation are reported in Table 5.1 for each participant, along with the average performance across all participants. Performance varied across participants, reflecting the inter-individual differences in EEG patterns, possibly the different levels of attention or cognitive strategies employed during the experiment, and the different numbers of training/testing data were used per subject (following trial rejection), with some participants having very little data to support the machine-learning analysis. The mean AUC-ROC values obtained are 0.591 ± 0.06 for ERP-based classification and 0.575 ± 0.06 for ERSP-based classification. Both of these values are above the 0.5 baseline, which represents a random classifier. This demonstrates the potential utility of EEG features for the task of differentiat-

Table 5.1: Mean AUC-ROC values obtained for each participant across all folds, separately for ERP and ERSP features.

Participant	ERP-based classification	ERSP-based classification
1	0.564 ± 0.09	0.522 ± 0.09
2	0.585 ± 0.11	0.558 ± 0.07
3	0.520 ± 0.07	0.532 ± 0.07
4	0.666 ± 0.07	0.626 ± 0.09
5	0.714 ± 0.06	0.649 ± 0.08
6	0.555 ± 0.11	0.522 ± 0.10
7	0.601 ± 0.10	0.525 ± 0.08
8	0.590 ± 0.08	0.674 ± 0.08
9	0.609 ± 0.09	0.489 ± 0.06
10	0.628 ± 0.06	0.618 ± 0.09
11	0.477 ± 0.08	0.611 ± 0.12
Mean	0.591 ± 0.06	0.575 ± 0.06

ing between the first and second viewings of remembered videos. The higher mean AUC-ROC value for ERP-based classification indicates that time-domain EEG features may provide more discriminative power in this context. However, the variance across participants suggests there might be value in exploring individualised models or adaptive algorithms that could take into account the individual characteristics of the EEG signals.

5.2.4 Conclusion

This pilot study represents an initial venture into the unexplored territory of EEG-based video memorability research. Despite the significant challenges posed by the experimental context and the complexity of EEG data, this endeavor has opened up an entirely new perspective for understanding and predicting video memorability.

The investigation was prompted to redefine its primary question due to the skewed class distribution in the memorability scores of the selected video clips. Instead of predicting whether a clip would be remembered or forgotten based on EEG data, the question became whether the first and second viewing of clips that were successfully remembered could be differentiated. This re-framed question led down an interesting, albeit unintentional path, revealing interesting patterns and adding a new dimension to the understanding of video memorability.

The key insights from this study emanate from the comparison between ERP and ERSP features in their ability to differentiate between the first and second viewings of the remembered clips. Although the performance of the models was only slightly above chance, the variance across participants hinted at the potential individual-specific characteristics that may influence the memorability of videos, an intriguing area for future research.

Moreover, the public release of the processed EEG features alongside the dataset [1] extends an invitation to the broader research community to dive into this dataset. Although no additional research has been conducted using the released data from this pilot study, the EEGMem [274] dataset, which was developed and publicly released (detailed below in section 5.3.1) with knowledge acquired from conducting the pilot study, has been used in two independent studies. In the study by Hamelink [275], Event-Related Potentials (ERPs) released with the dataset were analysed across 1,000 trials to explore the neurophysiological differences between videos that were later remembered versus those that were not. The posterior brain region was examined across three channels (Oz,O1, O2), and the right temporal cortex was considered through one channel (P8). Amplitude differences were noted in the 340 – 408 ms post-onset window and around the 476 ms mark in the visual cortex between remembered and non-remembered videos. Additionally, in the right temporal cortex, a distinct amplitude difference was observed within the 306 – 816 ms window. These findings suggest a pronounced P300 component (see [276] for more detail) for remembered videos in the right temporal lobe. Conversely, in the visual cortex, a greater positivity was associated with videos that were not remembered. In contrast, Kleinlein et al. [277] presented two processing pipelines aimed at predicting whether a video would be subsequently remembered. Motivated by the variability in how different subjects respond to the same video, the first approach involved the aggregation of statistical vectors for each trial (i.e., subject and video pair), followed by the application of a random forest model. The second approach centered on ERP channel coherency (a measure of the strength of the coupling between the signal recorded by two sensors at specific frequency bands). For each subject and video, the coherency between each ERP channel was computed pairwise across various power bands, resulting in a 28x28x4 matrix that described channel interactions within these spectral bands. This matrix was then transformed into a vector embedding, which was used as input for a shallow neural network with 256 neurons, a ReLU activation function, and used the Adam optimiser. While their results were consistent with random chance, exhibiting an average AUC-ROC score of 0.506, it is positive to see the dataset promoting interdisciplinary research, adding to the overall field of video memorability. While acknowledging that the findings from this study and my own are cursory and adjacent to the subject of video recognition memorability prediction, they represent a significant stride forward, and lay a solid foundation for future research in this intriguing space. Further work (outline in the next section) could consider larger sample sizes, more nuanced EEG features, and more diverse sets of videos, allowing for the understanding of the relationship between EEG signals and video memorability to be deepened.

5.3 Memories in the Making: Predicting Video Memorability with Encoding Phase EEG

In a world awash with fleeting moments, the river of time relentlessly splashes us with the experience of being. Amidst this mercurial torrent of sensory droplets—each vying for a place in the precious annals of our memory—our brain stands as the vigilant gatekeeper, painstakingly managing the flow of water and deciding which droplets will reach the reservoir of our memory. However, our reservoir—like any storage system—is subject to constraints of capacity and encoding efficiency. We accordingly posit that a critical moment of memorability should exist, an ephemeral yet potent point in time which captures the essence of an experience, and assigns it a “remembering priority”, which will ultimately determine its fate within the annals of our memory when consolidation comes around.

This section outlines a second exploration into the use of EEG in the context of video memorability prediction, building on the aforementioned pilot study with a refined experiment procedure, and completely new data collected from different participants. More specifically, we investigate the utility of encoding phase electroencephalography (EEG) signals, recorded from subjects during video stimulus presentation, to predict subject specific recognition upon subsequent (24–72 hours later) re-viewing. EEG signals were transformed into the visual domain by turning them into scaleograms with a continuous wavelet function, which allowed us to avail of state-of-the-art visual deep learning techniques. By leveraging temporal and spatial information contained within the EEG data, we position ourselves to capture the moment of memorability—a moment of encoding that corresponds to a remembering moment. We hypothesised that the neural signals recorded during this moment of memorability will differ from those recorded during forgettable moments, and that these differences can be used to predict whether a given subject will remember a given video. We employed a two factor study design—comparing subject-independent (SI) and subject-dependent (SD) training approaches, and compared single electrode and composite 28 electrode scaleogram images—in order to evaluate the generalisability of our approach and whether theta band (4–8Hz) activity over the right temporal lobe (channel P8), which has been implicated in memory formation [90], [278], leads to more accurate predictions.

5.3.1 Dataset: EEGMem

EEGMem [274] represents an extension and enhancement of the Memento10k [192] video memorability dataset. As opposed to the prior pilot study described earlier in this thesis, this enriched dataset boasts a significantly wider variety of videos, offering a broader, more diverse range of stimuli to enhance the robustness of our findings. It comprises encoding phase EEG recordings gathered from 12 participants as they engaged with a subset of the Memento10k videos. In total, 45 participants were recruited (16 recorded at DCU and 29 recorded at UoE[3]), but strict inclusion criteria—including EEG trial data quality and a false positive recognition rate of less than 30%—led to the final dataset comprising data from only 12 participants. The recordings were captured as part of the 2022 MediaEval Predicting Video Memorability task [274]. The data collection process of EEGMem was designed to address the class imbalance that surfaced in the preliminary study. Distinct from the immediate recognition memory examined in the pilot study, this investigation shifted focus to long-term recognition memory, assessed 24 to 72 hours following the initial viewing. This amendment was grounded in well-established memory decay research, implying that memories naturally fade with time. Hence, we hypothesised that fewer videos would be recalled after this extended interval, ensuring a better balance between remembered and forgotten classes.

Specifically, the EEGMem data collection process encompassed two distinct phases. The encoding phase recorded participants' EEG data as they watched a continuous stream of 1,000 videos. Subsequently, the online recognition phase took place between 24 to 72 hours post the encoding phase. In this phase, participants re-watched the original 1,000 videos, interspersed with an additional 1,000 unseen videos from the Memento10k dataset. Participants were then prompted to indicate, via a keyboard press, whether they recognised a video from the initial encoding phase. This method enabled the collection of binary annotations signalling recognition, thereby facilitating a more balanced and meaningful exploration of the potential of EEG data in predicting video memorability.

The allocation of the video stimuli during the encoding phase was undertaken with careful consideration to ensure full coverage of the original 10,000 videos from the Memento10k dataset over the course of data collection. The allocation process also ensured that each participant viewed a unique subset of videos while also expe-

[3] Data collection for participants carried out DCU was done with approval from the University's Research Ethics Committee (DCUREC/2022/100), and for participants at UoE with approval from the Ethics Committee there (ETH2122-0001). Data at DCU was collected using a 32-channel ANT Neuro eego system with a sampling rate of 1000 Hz. Data at UoE was collected using a 64-channel BioSemi ActiveTwo system at a sampling rate of 1000 Hz.

riencing the most and least memorable videos from the set. The 1,000 videos viewed by each participant during the encoding phase were composed of the 75 most memorable videos, the 75 least memorable videos, and an additional 850 unique videos. A minimum of 12 participants were required to fully cover the Memento10k dataset. The most and least memorable videos were selected to be viewed by all participants, facilitating direct comparisons of EEG responses across participants to these particularly memorable and non-memorable stimuli. The additional 850 videos were selected pseudo randomly, with each participant viewing a unique set of these videos. In order to guarantee reproducibility, the randomisation process utilised a fixed seed. If at any point the remaining pool of videos (excluding the top and bottom 75) fell below 850, the selection pool was replenished by reshuffling the original list, excluding the videos that had already been viewed. The 1,000 videos were presented to the participants in 8 blocks, with each block consisting of 125 videos. Between blocks, participants were shown a 2-minute break video. The allocation procedure can be represented in pseudo code as follows:

Algorithm 1 Video Allocation for EEG Participants

```
procedure VIDEO ALLOCATION($files, n_{subjects}, n_{videos}, top_{75}, bottom_{75}$)
    $shuffle(files)$ with a fixed seed
    for subject in 1 to $n_{subjects}$ do
        $unique_{850} \leftarrow select_{850}(files)$
        $files \leftarrow files - unique_{850}$
        if $length(files) < 850$ then
            $shuffle(files)$ with a fixed seed
            $files \leftarrow files \cup unique_{850}$
        end if
        $video_{set} \leftarrow top_{75} \cup unique_{850} \cup bottom_{75}$
        Save $video_{set}$ in a block-structured file for $subject$
    end for
end procedure
```

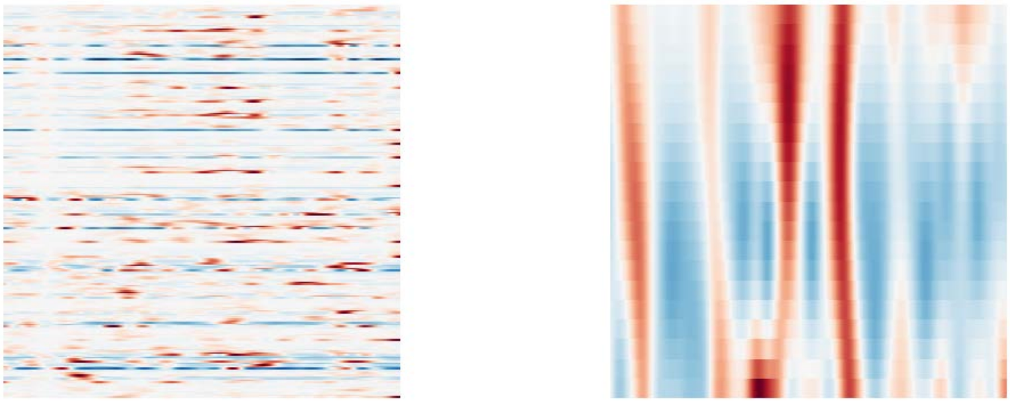
Through this process, we ensured the collection of diverse data across the Memento10k video dataset, while eliminating participant-specific bias and enabling the comparative study of responses to specifically memorable and non-memorable videos.

5.3.2 Pre-processing

A standardised pre-processing procedure, which aims at enhancing signal quality and reliability, was implemented for the raw EEG. The goal of this process was to eliminate non-neural sources of variability, reduce noise, and label the data for subsequent analyses. First, the data was referenced using a common average reference, a technique which subtracts the mean voltage of all electrodes from each individual electrode at every time point. This method minimizes the effects of volume conduction and common noise. Next, the data was band-pass filtered between 0.1–30Hz. This effectively removed slow drifts and high-frequency noise from the EEG signal, leaving the relevant neural oscillations (delta, theta, alpha, and beta waves) that are typically associated with cognitive processing. The filtered data was then subjected to ICA (see section 5.1.5), a statistical method used to separate mixed signals into their independent components. ICA effectively identified and removed artifacts, such as eye movements, muscle activity, and heartbeats from the EEG data. Lastly, binary annotations from the recognition phase were used to label each participant's specific encoding phase EEG trials and their associated Memento10k videos. Videos correctly remembered, our "remembered" class, were labeled as "True" (true positive), and all other outcomes were labeled as "False", our "forgotten" class. Through these pre-processing steps, the EEG data was effectively cleaned, standardised, and labeled, paving the way for accurate and reliable subsequent analyses.

5.3.3 Feature-extraction

With the goal of leveraging the state-of-the-art Vision Transformer (ViT) architecture, which provides significant advantages over traditional CNNs due to its ability to model long-range dependencies across the input space [279], we turned the filtered EEG data into scaleograms with a Morlet wavelet function. Two types of scaleograms were generated: a single-channel variant and a variant incorporating all 28 electrode channels (Figure 5.5). The decision to use Morlet wavelet scale-



(a) 28-channel composite

(b) single channel

Figure 5.5: Examples of each of the two types of scaleogram images

ograms was primarily due to their high time-frequency resolution, which effectively captures the temporal evolution of spectral features in the EEG data. This choice also aligned with the goal of assessing the predictive potential of the theta band frequencies (4–8Hz), particularly over the right temporal lobe (electrode P8 in this case). For the single-channel scaleograms, 20 frequencies were linearly spaced from 4 to 8Hz, focusing on the theta band. In contrast, the 28-channel scaleograms used 8 linearly spaced frequencies ranging from 3 to 17Hz, thereby encompassing a broader spectrum that includes delta, theta, alpha, and low beta frequencies. The number of cycles employed in the wavelet function for both frequency sets was dictated by a logarithmic function that efficiently balanced the time-frequency resolution trade-off inherent in wavelet transformations [257]. We incorporated a baseline period from -0.25 to 0 seconds for both scaleogram types. Z-score normalisation was applied, effectively standardising each frequency's amplitude to a common scale. Scaleograms were then averaged across trials for visualisation within the time window of -0.5 to 3 seconds, corresponding to the full duration of the Memento10k videos. The final 28-channel scaleogram image consists of a 4 by 7 grid, each cell displaying an individual scaleogram corresponding to one of the 28 electrode channels. This comprehensive representation encapsulates a wide array of spectral features across different scalp regions, providing a rich dataset for the subsequent application of the ViT model.

5.4 Methodology

The ViT architecture was selected for all experiments primarily because of its minimal inductive bias towards image-specific features, which allows it to generalize well across diverse visual tasks. Given its success in various image recognition challenges, the ViT architecture has emerged as the current state-of-the-art solution, providing the potential for effective, high-performance processing in the EEG scaleogram image analysis.

5.4.1 Vision Transformer (ViT) Architecture

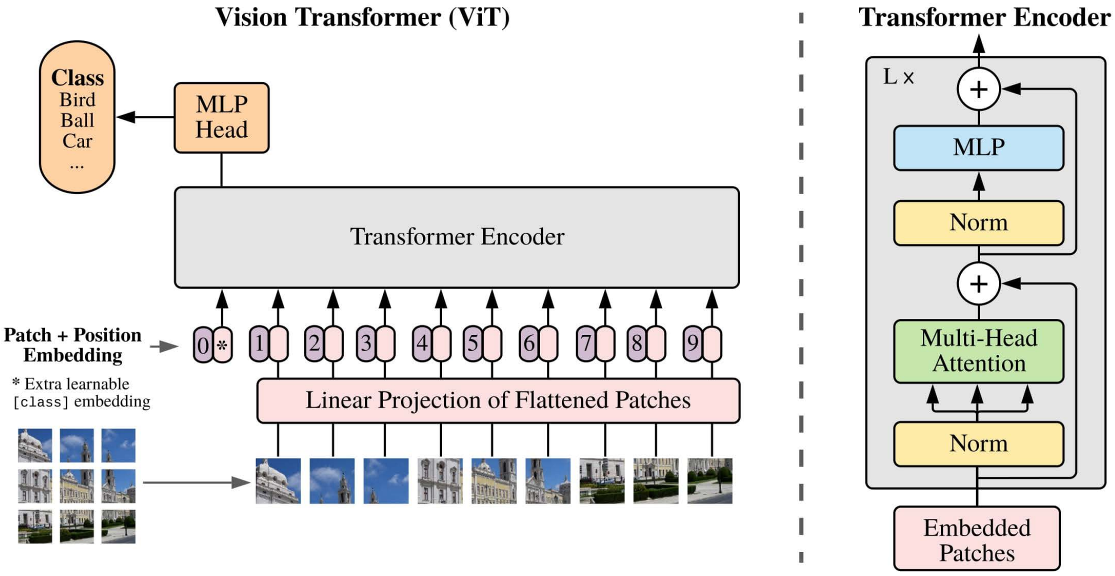


Figure 5.6: Vision Transformer architecture (image from [279])

Introduced to the world in 2020 by Dosovitskiy et al. [279], the Vision Transformer (ViT) has since captivated the field of computer vision-—-a refreshing departure from the ubiquity of convolutional neural networks (CNNs). Jettisoning the convolutional approach, the ViT architecture instead relies on the transformer architecture—originally a tour de force in natural language processing tasks [280]—to sequences of image patches [279]. A ViT model takes an input image and dissects it into fixed-size, non-overlapping patches—these small cut-outs then undergo linear embedding, thereby transforming them into a sequence of vectors. This technique mirrors the way words are embedded into sentences in transformer-based language models. To this sequence, a learnable classification token is appended—an essential ingredient for the ultimate task of image classification [279]. Following the creation of this sequence, the model channels it through multiple layers of transformer encoders. Each encoder is a multifaceted unit, featuring multi-head self-attention and feed-forward neural network layers—all interleaved with layer normalisation and residual connections [281]. Leveraging the self-attention mechanism, the model assimilates the global context of the image, enabling it to consider all patches simultaneously and coherently [279]. Subsequently, the transformer layers churn out a sequence of vectors, with the premier one representing the classification token. This vector navigates through a final linear layer, producing the output classification probabilities. The ViT's architectural elegance enables it to not only rival, but at times outperform, CNNs in large-scale image classification tasks—an impressive feat considering the considerably different (CNNs relying on local convolutions and hierarchical structures, whereas ViTs leverage self-attention mechanisms over flattened image patches) methodologies [279]. Despite being resource-intensive, requiring large-scale training datasets and computational power, ViT presents a paradigm shift in computer vision. By illustrating the successful application of the transformer architecture to image data, it spotlights the architecture's versatile potential, inspiring novel research directions.

5.4.2 Fine-tuning ViT Models

Fine-tuning is a commonly adopted technique in deep learning that leverages a pre-trained model, originally trained on a large-scale dataset, and retrains it on a target task-specific dataset. The idea is to harness the knowledge captured by the pre-trained model and transfer it to the target task, an approach known as transfer learning [282]. When fine-tuning a pre-trained Vision Transformer (ViT) model, no layers are typically removed. The ViT model consists of several Transformer encoder layers that capture a rich understanding of the image representations in their weights [279]. The pre-existing architecture and its weights serve as a robust initial state for training. The primary update to the model during fine-tuning is in the output layer. The output layer of a pre-trained ViT model is task-specific, typically designed for a multi-class classification task (e.g., 1000 classes for ImageNet). When fine-tuning for a specific task, this output layer is often replaced with a new one appropriate for the target problem (e.g., binary classification would need only two output nodes). This output layer is randomly initialised and updated during the training process. The rest of the model is then fine-tuned on the target dataset. All the weights in the layers are updated during the training phase. Although one could freeze some layers during this process to preserve the pre-trained weights, it is generally beneficial to update all weights, particularly for the Transformer model. This is because the self-attention mechanisms in the Transformer layers are data-dependent and updating these weights can help the model better adapt to the target task [281].

In order to optimise the predictive power of the models, we fine-tuned (i.e., no layers were frozen, a new task specific head was created, and the model is re-trained) a total of 42 distinct ViT-large models pre-trained on the ImageNet-21k dataset. Each model was fine-tuned on one of four types of data: 28 Channel, Fp1, P8, and Frames. The fine-tuning procedure was executed under two distinct training categories: Subject-Independent (SI) and Subject-Dependent (SD), with 12 SI models (one per subject), and 4 SD models trained for each data type. In the case of SI models, a leave-one-out cross-validation approach was implemented. This method involved training the models on the data from all subjects excluding one, and subsequently testing the model on a test set of the data from the excluded subject. This approach is widely used to obtain an unbiased estimate of model generalisation on independent data, and has been shown to provide robust performance measures in studies with limited subjects. In contrast, the SD models were fine-tuned on data from all subjects, excluding a combined test set. This combined test set was created by pooling together a stratified test set comprised of 15% of each subject's data. The rationale behind this approach was to provide the model with a greater diversity of data during training, while still providing a robust test set for model evaluation. For both SI and SD training categories, all models were trained using the AdamW optimiser, a popular and efficient gradient-based optimization algorithm that computes individual adaptive learning rates for different parameters [283]. The learning rate was set to 2e-5, and a cosine learning rate scheduler was employed. This scheduler adjusts the learning rate based on the progress of training, effectively reducing the rate towards the end of the training process. A weight decay of 0.1 was applied to avoid overfitting by adding a penalty to the loss function, thereby reducing the complexity of the model. The implementation of both early stopping and a dropout rate of 0.65 further served to minimise overfitting. Early stopping is a form of regularization technique where the training process is halted once the performance on a validation set stops improving, thereby preventing the model from learning the noise in the training data. The dropout method randomly drops units and their connections during training, which helps to prevent co-adaptation of feature detectors and improve model generalization.

5.4.3 Results and Analysis

Model performance was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) as there were only two prediction classes—remembered and forgotten. Individual classification results for each subject (column headed SID) are displayed in Table 5.2, for both the SI and SD models. This shows that the ViT models trained on channel P8 scaleograms are the best performing, with a mean AUC-ROC of 0.567 for SD trained models, and a mean AUC-ROC of 0.563 for SI trained models. A set of ViT models were trained on a random channel's (Fp1) scaleograms in order to account for the possibility that data resolution was the factor driving the 28 channel and P8 single channel performance difference, which the results suggest was not the case. An additional set of ViT models (Frames) were

Table 5.2: AUC-ROC scores for all models. (Subject Mean excludes Frames column data)

	28 Channel		Fp1		P8		Frames		
SID	SI	SD	SI	SD	SI	SD	SI	SD	Sub Mean
S-2	0.445	0.540	0.488	0.509	0.558	**0.572**	0.487	0.477	0.518
S-4	0.482	0.466	0.483	0.470	**0.634**	0.577	0.472	0.496	0.518
S-9	0.437	0.530	0.459	0.478	0.569	**0.613**	0.498	0.501	0.514
S-10	0.462	0.457	0.489	0.502	0.568	**0.578**	0.515	0.483	0.509
S-13	0.430	0.474	0.464	0.508	0.574	**0.596**	0.492	0.500	0.508
S-16	0.383	0.523	0.497	0.528	0.534	**0.564**	0.489	0.482	0.503
S-19	0.393	0.405	0.509	0.516	0.578	**0.636**	0.514	0.488	0.506
S-30	0.388	0.441	0.331	0.349	0.340	0.436	0.502	**0.492**	0.381
S-31	0.626	**0.655**	0.521	0.556	0.510	0.520	0.479	0.506	0.565
S-36	0.696	0.540	0.599	0.532	**0.707**	0.564	0.511	0.481	**0.606**
S-37	0.562	0.323	0.531	0.401	0.614	**0.632**	0.503	0.486	0.510
S-41	0.566	0.386	0.501	0.343	**0.567**	0.538	0.500	0.506	0.484
Mean	0.489	0.478	0.489	0.474	0.563	**0.567**	0.497	0.492	

trained on video frame data—three frames were extracted, 1 per second and the majority prediction was chosen—rather than EEG data, the results of which were random, with no discernible difference between SI and SD training, nor notable difference in subject performance, which is logically coherent with the fact that video frame data is not influenced by the subject viewing it.

The repeated-measures design (each participant is in both subject-dependent and subject-independent conditions, and their data contribute to both scaleogram types), allowed us to perform a paired t-test, finding that the AUC-ROC scores for models trained with P8 channel scaleogram images are significantly (t= 3.243, p = 0.0036) greater than those trained with 28 channel scaleogram images, adding weight to the hypothesis that theta band (4–8Hz) oscillations over the right temporal lobe (channel P8), predict encoding of declarative memory [278]. Figure 5.7 and Figure 5.9 provide visual illustrations of the difference between P8 and 28 channel trained scaleogram images. Although we did not find a significant interaction between SD and SI training, we can see that more points in Figure 5.7 lie below the line of best fit, indicating that more models trained on SD performed better than their SI trained counterparts, which makes sense from a training perspective, as EEG data can be highly subject specific, and not having any subject training examples can hinder

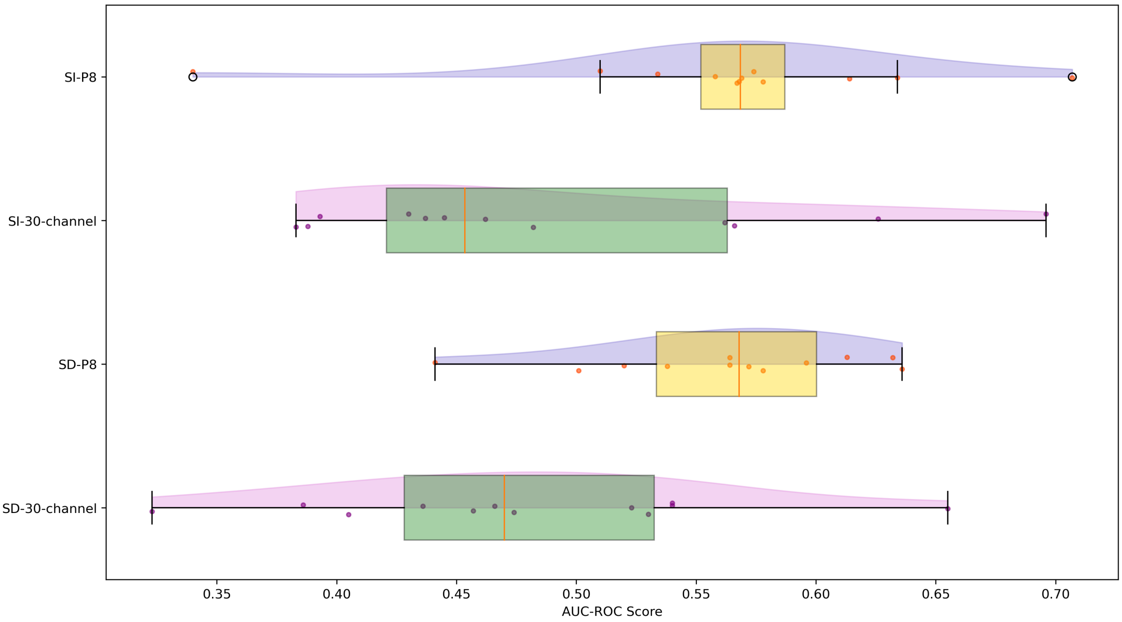


Figure 5.7: Raincloud plot comparing 28 channel and P8 channel scaleogram trained models.

prediction performance. Figure 5.7 additionally highlights the impact of 28 channel vs channel P8 scalograms on a per subject basis, with a green line indicating that either SD or SI AUC-ROC scores improved and red lines indicating that they both worsened.

Subjects S-30 and S-36 sit on two ends of the performance spectrum, with S-30 producing the worst performances across the board (an average AUC-ROC of 0.381), and S-36 producing the highest average AUC-ROC of 0.606 and the highest absolute AUC-ROC of 0.707. While both subjects boast an inordinately high imbalance (<90%) in their ground truth responses—subject S30 responding “forgotten” for 1,828 videos out of 2,000, and subject S36 doing so for 1,871 videos—subject S30’s results are likely a reflection of the quality of their response data as a large portion of their response reaction times consistently rested within the 1 - 1.5 second mark. This suggests a more rhythmic rather than innate and earnest nature to their responses, whereas subject S36’s average response reaction times were more varied and typically given <1 second after stimulus onset.

Reaction Times

Figure 5.8 shows the per-subject distributional differences in video ground-truth memorability (population level) scores for remembered (tp) and forgotten (fn) videos. A series of statistical tests were carried out in order to assess the relationship between response reaction times and memorability scores.

A Pearson correlation analysis between response times and ground-truth memorability scores for each video was conducted across all subjects. The findings revealed no statistically significant correlations between these variables for any subject, suggesting that no clear linear relationship exists between response times and memorability scores. Then, two per subject independent t-tests with Bonferroni correction were performed, testing the differences between mean response reaction times and mean video memorability scores for remembered and forgotten videos. For mean response reaction times, significant differences for subjects 10, 13, 16, 19, 30, 36, 37, 9, and 2 were found, where the mean response reaction times for remembered videos were consistently higher. For subjects 41 and 4, the inverse was found. Subject 31 was the only subject with no statistically significant difference in their mean response reaction times. For ground-truth memorability scores, only subjects 2 and 4's results showed significant differences in mean memorability scores. For subject 2, the mean memorability score was higher for remembered videos (M=0.842) compared to not remembered videos (M=0.771), t=6.801, p=3.33e-11. Likewise, for subject 4, the mean memorability score was higher for remembered videos (M=0.809) compared to not remembered videos (M=0.780), t=2.891, p=0.00395. Given that only subjects 2 and 4 demonstrated a significant difference, it highlights the fact that the influence of memorability on long-term video recognition might be inconsistent across individuals, or potentially affected by individual differences or data quality variations. Further investigation is required to determine if there is a causal relationship or if other factors might be influencing the observed differences.

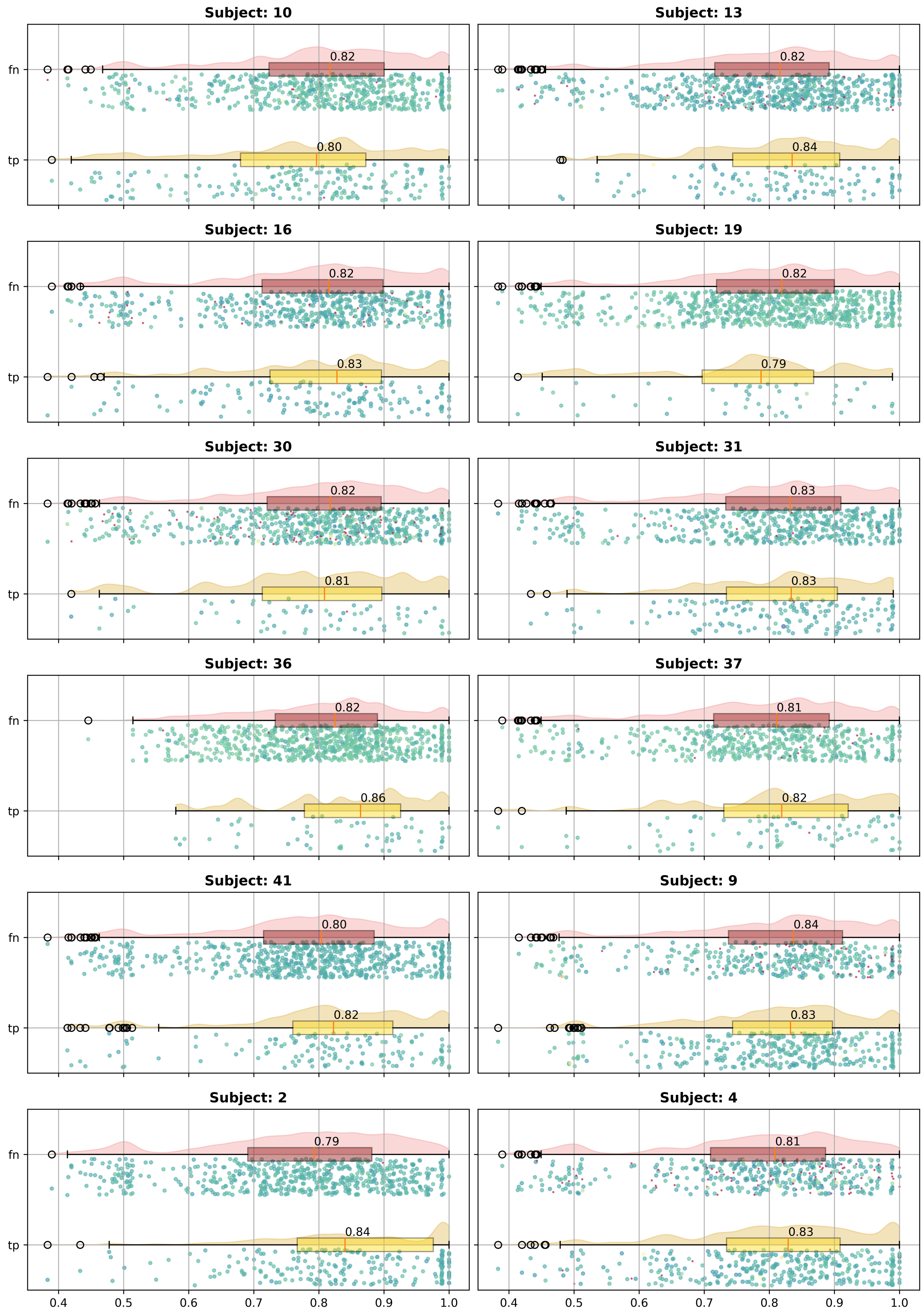


Figure 5.8: Raincloud plots for each of 28 subjects plotting the difference in ground-truth memorability scores (x-axis) for remembered (tp) and forgotten (tn) videos. The colour and size of each dot is proportional to the subjects response reaction time when viewing the video for the second time.

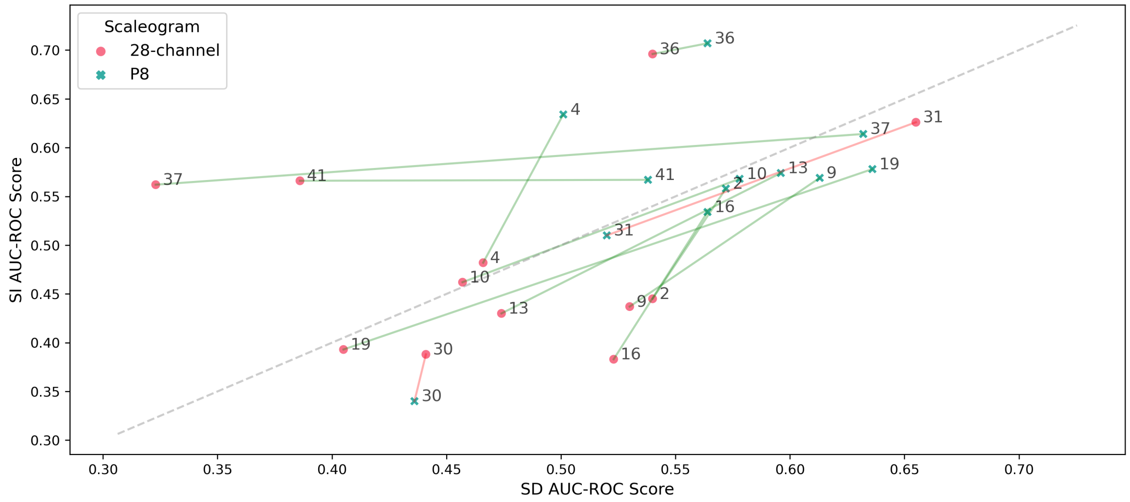


Figure 5.9: AUC-ROC scatter plot for scaleogram comparison. Subject specific P8 vs 28 Channel model performance difference is illustrated with coloured lines—green indicating P8 improvement, and red indicating 28 Channel improvement.

5.4.4 Moment of Memorability

The concept of a 'moment of memorability' refers to a specific point during video viewing at which the viewer comprehends the underlying concept of the video. At this pivotal moment, the brain assigns an encoding priority: high for highly memorable content and low for less memorable things, which is likely to be tied to the content's human information utility [284]. Interestingly, the analysis of reaction times in relation to memorability scores shows that while there is no direct correlation between reaction times and the memorability of a given video, there is a significant relationship between the mean reaction times and whether a video was remembered or forgotten. This observation can be elegantly explained by the concept of the 'moment of memorability'. When a viewer starts to feel a sense of familiarity with a video, they might wait until this moment of memorability to respond. Here, the feeling of familiarity culminates, leading them to confirm that they remember the video. This could account for the longer reaction times for remembered videos. Conversely, in the absence of this feeling of familiarity, viewers might automatically respond that they do not remember the video, leading to quicker response times for forgotten videos.

Having established the implications of mean reaction times and identified sub-

Table 5.3: AUC-ROC scores for models trained on scaleograms generated with video mean reaction time slice

	Fp1		P8	
SID	SI	SD	SI	SD
S-36	0.615	0.636	0.715	**0.742**

ject S-36 as the "cleanest participant" (characterised by their lowest false positive rate, indicating that they responded to the experiment in earnest, and the attentive nature of their EEG, evident through low alpha frequency), we embarked on a supplementary investigation to test the 'moment of memorability' hypothesis. This hypothesis posits the existence of a critical time window within the viewing period that significantly influences memory encoding. To investigate this phenomenon, we chose to exclusively use subject S-36's data. While one might argue that focusing on subject S-36's performance could seem like zeroing in on an outlier, we have good rationale. Many other participants demonstrated markedly higher false positive rates, diminishing the reliability and utility of their samples for our research goals. Given our limited sample size—a comprehensive, in-depth analysis requires data from a minimum of 30 high-quality subjects—rather than incorporating a few more subjects of debatable quality, we reasoned that leveraging data from our "cleanest participant" would provide a more clear-cut preliminary exploration. This strategy is to gauge the viability and value of a future larger-scale experiment. We generated scaleograms for the Fp1 and P8 channels, but with a key distinction: we limited the Morlet wavelet generation to a time window of 0.25 seconds before and after the videos' mean recognition response. This timeframe was chosen based on the previous mean reaction times findings. The results of this exploration are shown in Table 5.3 below, listing the AUC-ROC scores for these models:

The supplementary examination of the 'moment of memorability' hypothesis offers intriguing, though preliminary, insights that deepen the grasp of memory encoding dynamics. The augmented AUC-ROC scores, particularly evident in the P8 channel models, emerged when centering scaleogram training around the mean reaction time. This enhancement indicates that a pivotal viewing juncture—termed here as the ‘moment of memorability’—where the content is conceptually encapsulated by the viewer, exerts a degree of influence on memory encoding and retrieval. This revelation is in harmony with the cognitive perspective that memories are not merely static impressions, but are intrinsically linked with the comprehension and significance derived from experiences.

5.4.5 Conclusion

The results of this research support both the hypothesis that theta band (4–8Hz) oscillations over the right temporal lobe (channel P8) are involved in the encoding of declarative memory, and the potential existence of a distinct “moment of memorability” (H4) that can be leveraged to predict subsequent subject-specific recognition. Although a significant interaction between subject-dependent (SD) and subject-independent (SI) training approaches was not found, the higher performance of several SD trained models suggests the potential importance of subject-specific EEG data for prediction. This highlights the need for further investigation into the adaptability and generalisability of models across different individuals.

The current study, of course, is not without its limitations. The binary nature of ground truth responses may oversimplify the complex processes of memory encoding and recognition. Additionally, the issue of volume conduction in EEG measures requires further attention to ascertain its impact on the predictive power of the models. Nevertheless, the study has unveiled promising avenues for further exploration and has demonstrated the feasibility of using EEG signals in conjunction with deep learning techniques to predict video memorability. Notably, the ‘moment of memorability’ hypothesis (H4) offers a valuable new perspective on how we consider video content and its memorability. Rather than viewing memorability as a static property inherent to the video, this concept proposes that memorability may be tied to the point in the video where comprehension or mental representation of its content culminates. This suggests that the underlying concept being conveyed in the video, and the timing of when that concept is fully comprehended, may play a critical role in the video's memorability. This opens up the exciting prospect of exploring not only what is remembered, but when and why certain content is remembered, paving the way for a deeper understanding of the intricacies of memory encoding and recall. This fresh approach has the potential to greatly enrich future investigations into the predictability of video memorability.

Chapter 6

Reconciling the Rift Between Recognition and Recall

> *"Every act of perception, is to some degree an act of creation, and every act of memory is to some degree an act of imagination."*
>
> — Oliver Sacks, *Man's Search for Himself*

The tapestry of human memory is intricate, woven with multiple threads of cognition, each one vital in the creation of the final picture. Two such threads, recognition and recall, form the basic fabric of remembrance [285]. These mechanisms, though conceptually distinct, are nonetheless entangled on a neural level, performing an intricate dance of interconnectedness in the cerebrum's amphitheater [286]. The nuances of this performance, however, are not yet entirely understood. Existing computational models of memorability stand as impressive monuments to our understanding of recognition memory. These models, however, gaze singularly upon recognition tasks, fixed solely upon the binary response of "yes" or "no" to indicate whether an individual believes they have encountered a particular stimulus before [1], [147], [192], [198].

Yet this approach harbours a silent presumption, one that overlooks the vast complexity within the recognition process itself. Consider a key distinction: an item might be recognised based on a mere sensation of familiarity or on the basis of "recollection"—where distinct, contextual details about the item can be recalled from the caverns of memory. Though seemingly subtle, this distinction is profound, causing ripples of implications through our understanding of recognition. It is a caveat that current models, with their monolithic focus on simple recognition tasks, often neglect. Recognition, as it turns out, is a delicate game of reliance. If the mechanisms of recollection stumble and falter, recognition's footing is expected to find its balance in the underpinning feeling of familiarity [285]. The intertwined nature of these two mechanisms presents a conundrum: to what extent do they influence each other? Can one truly be isolated from the other?

Further complicating the matter is the work of Bainbridge et al. [152], which suggests an absence of correlation between recognition and recall. Their study found no significant connection between the number of participants who recognised an image and the number who were able to recall the same image. Additionally, the correlation between the quantity of objects recalled for an image and its recognition rate was found to be equally unremarkable. Yet, this conclusion challenges the aforementioned intimate connection between recognition and recall, setting the stage for an academic rift that demands resolution. While Bainbridge et al.'s contributions to the field are undeniable, their chosen means of analysis raises concerns. Their approach, while elegant in its simplicity, arguably oversimplifies the complexity of recognition and recall, relying on measures of comparison that, when examined closely, appear to be unrefined and unequal (straightforwardly comparing simple recognition and recall metrics without accounting for innate processing capacity differences, e.g., individuals can correctly recognise upwards of 10,000 images[139], [140], which is orders of magnitude greater than recall limits [287]). A more nuanced approach, one that acknowledges the multifaceted nature of these processes, seems imperative in order to enrich our understanding. Providing a strong counterpoint to Bainbridge et al.'s conclusions, Broers and Busch [288] employed the more refined "remember/know" procedure—participants indicate directly, after an old/new statement, whether they remember specific episodic details about the item (recollection) or whether they only know that the item is old (familiarity) [32], [289]—finding evidence to suggest that an image's memorability scales with a greater likelihood of episodic recollection but not familiarity. They also noted considerable variability in the judgements across individual images: some memorable images were recognised almost exclusively based on recollection, others mostly on familiarity. In essence, images with high recall memorability also tend to have high "yes/no" recognition memorability. This variation provides both clarity to the relationship between recognition and recall, and raises further substantial questions, such as: why are certain images highly memorable, but are consistently more strongly associated with pure familiarity rather than recall?

The existing body of literature, while rich and thought-provoking, paints a complex and multi-dimensional landscape, where contradictions, oversights, and questions blend into a blurry picture. As fascinating as this panorama is, it leaves us with a yearning for clarity and a quest for comprehension. In light of this, the forthcoming sections of this chapter aim to provide clarity on the subject of visual recall and seek to resolve inconsistencies in the literature pertaining to the relationship between recognition and recall by detailing an innovative video recall experiment that leveraged drawings as a measure of recall.

6.1 Setting the Stage

Research into visual recall memorability has been primarily foundational, focusing on basic effects to support memory system theories, and resulting in few concrete insights into the visual attributes influencing recall performance. Probability of recall is generally regarded as a function of serial presentation position, with two basic effects emerging in serial-position curves—a primacy effect, increased recall probability of items near the start of a presentation list; and a recency effect, increased recall probability for items near the end of a presentation list [125], [290], [291]. This primacy effect can be attributed to the increased rehearsal of the first few items of a list, resulting in better long-term storage for these items and can be eliminated by ensuring all items receive equal amounts of rehearsal [292]. The recency effect can be eliminated with a short mental task, following presentation and preceding recall, indicating that the effect can be attributed to items still being held in short-term memory [293]. The degree of vividness with which a person reports being able to visualise imagery is predictive of their recall performance [294], [295]. The strongest determinants of recall are list length and the complexity of items, with short lists of low complexity items exhibiting the greatest recall [125], [290], [291], [293], [295]. Given that more complex stimuli also eliminate the primacy and recency effects [291], many past studies' use of simple stimuli—line drawings [167], [294], [296], or images with simple depictions of objects [297], [298]—and low resolution verbal metrics—a single word [167], [296], [297], or a brief verbal description [290], [291], [294], has resulted in very little insight into the content and contributing factors of memory formation. A recent study explicitly set out to address many of these past limitations, providing deeper insight into recalled memories and assessing the relationship between "recall memorability" and "recognition memorability" [152]. They found that drawings from Delayed Recall (with an 11-minute digit span task following presentation and preceding recall) accurately reflect aspects of their original images, containing visual information beyond a simple construction from the scene category label. Drawings made while viewing an image or immediately after encoding it, display a greater degree of diagnosticity, indicating time modulated memory decay. Memory drawings were found to preserve an accurate spatial map of the original image, and contain very few incorrect objects. It was also suggested that recall could be driven by semantic meaning captured in an image—with visual saliency and meaning maps explaining aspects of memory performance. Ultimately, they purported to have found no relationship between the "recall memorability" and "recognition memorability" of individual images.

6.1.1 Myopia in the Mind's Eye

The ability to conjure up colourful images and examine them in the mind's eye has long been thought of as fundamental to a thinking mind. This assumption was first articulated by Aristotle in *De Anima*—"the psyche never thinks without an image . . . the reasoning mind thinks its ideas in the form of images"—and has since established a long history in philosophical psychology. The belief that the character of one's mind is like any other is likely to be at the heart of this intuition. Given the impossibility of inspecting the qualia of a mind other than one's own, what reason would one have to assume otherwise? In 1880 this widespread intuition was formally assessed for the first time by Sir Francis Galton, who was interested in the natural varieties in mental disposition. Aiming to define the different degrees of vividness with which individuals can recall familiar scenes, Galton pioneered the quantitative study of mental imagery with his "breakfast-table survey", reporting a wide variation in subject reported mental vividness, and some participants describing "no power of visualising" [299]. Even though mental imagery abilities surveys [294] have consistently suggested that 2-5% of people are non-imaging/imaging impaired, contemporary mental imaging literature still largely views non-imaging/imaging impaired individuals as 'repressive'/'neurotic', or outright denies their existence [300]. However, with the phenomenon's recent acquisition of a name—aphantasia: a condition of reduced or absent voluntary mental imagery [301]—the subject of inter-individual variability in internal mental representations has garnered more serious attention. Research into the neural correlates of individual differences in imagery vividness suggest that the early visual cortex plays an important role [302], however, there is also evidence to suggest that inter-individual variation in the vividness of mental imagery depends on an interaction between frontal, parietal, and visual regions, underscoring a more intricate interplay than previously thought [303].

The intricate relationship between vivid mental imagery and memory recall is eloquently expressed in several notable theories and research endeavours. Paivio's dual coding theory [304], for instance, articulates that encoding of information is significantly enhanced when both verbal and visual channels are engaged, resulting in more vivid mental representations and consequently, improved recall. Empirical evidence of this dynamic is observed in studies employing the method of loci, an ancient mnemonic strategy based on the creation of detailed, spatially structured mental images to facilitate information retrieval [305]. Echoing this, [306] reinforces the robust correlation between vivid mental imagery and recall, asserting that memories associated with detailed mental images are more likely to be successfully recalled. Moreover, research into the unique phenomenology of episodic memory further illuminates the central role of vivid mental imagery in recall. Such memories, often experienced as rich mental images [27], are generally characterised by higher detail, thus aiding recall [307]. Taken together, these studies strongly suggest a deep-seated nexus between vividness of mental imagery and the complex landscape of memory recall. However, with the introduction of aphantasia—a condition marked by an individual's reduced or absent ability to generate mental imagery—an intriguing paradox presents itself: despite the lack of vivid (or any) mental imagery, individuals with aphantasia often exhibit recall abilities akin to those with typical mental imaging capacities [308]. Although this may seem counterintuitive and somewhat contradictory to previous research touting the importance of vividness to recall, it simply implies that the mechanisms of recall are resilient to the absence of vivid mental imagery, albeit with a different experiential quality. This complex interplay highlights that accounting for inter-individual variation in mental imagery vividness is crucial when attempting to create a more faithful and individual-agnostic conceptualisation of memorability.

6.2 A Picture Paints a Thousand Words: Drawing Video Recall Experiment

The nature and content of mental visual imagery has historically been difficult to quantify. As we attempt to revisit past experiences lodged in our memory banks, what do we truly remember? Is our recall a high-resolution, lifelike reproduction of the event, a vague and diluted version, or perhaps merely a verbal account of the visual phenomena we once witnessed? Dissecting these facets is a pivotal step in unravelling the complexities of memory. It involves understanding what information is encoded and retained, how these memory traces deteriorate over time, and what aspects are resurrected when we summon these memories.

A recent study attempting to address these questions demonstrated that object and spatial details can be captured with a drawing-based visual memory experiment [152]. Representations of recollection, captured in the form of drawings, provide a window into the contours of one's mental terrain in ways that other methodologies may not permit. In this context, a drawing task becomes a potent tool in our analytical arsenal, equipping us to assess the nuances of mental imagery with a degree of precision hitherto unattainable. It offers a tangible, visual output that embodies the idiosyncrasies of individual cognition, revealing subtleties that may remain veiled within conventional measures of memory. With this in mind, a novel drawing based video recall experiment was devised and carried out in order to address the third hypothesis in this thesis (H3), which seeks to clarify the nature of the relationship between recognition memorability and recall memorability.

6.2.1 Experimental Design

In order to facilitate a more direct comparison between recognition and recall, videos from extreme ends of the recognition memorability spectrum were selected as the stimuli to be used in the video recall memorability experiment where drawing was leveraged as a recall tool. A total of 32 videos were selected from the Memento10k dataset. Half of these videos were chosen from the top 100 most memorable videos (Figure 6.2), and the other half were picked from the least memorable ones (Figure 6.3). Videos were picked with "drawability" in mind, and representing as broad an array of depiction categories as possible. In this context, drawability refers to the inherent qualities of a visual scene or event within a video that make it amenable to being accurately represented or reconstructed through simple sketches or line art. Several factors influenced our choice of videos based on this principle. Uniqueness was a primary factor; scenes chosen had distinct visual elements that set them apart from typical visuals, ensuring that drawings can be specifically attributed to a particular video. Simplicity was another essential criterion; videos with straightforward yet striking visuals were prioritised, avoiding overly intricate scenes that might hinder recall accuracy. Additionally, the cultural and cognitive accessibility of content was assessed, giving preference to scenes that have universal resonance, as opposed to those tied to specific cultural or sub-cultural contexts. The experiment was structured into eight rounds, where each round consisted of an encoding phase, in which participants watched four unique videos; a recall drawing phase, in which participants were tasked with drawing a scene from a "target" video—one of the four videos—from memory; and a perceptual baseline, in which participants were presented with a frame from the target video, and were tasked with drawing it. This perceptual baseline serves as an essential point of reference for each participant's innate drawing ability. By incorporating this baseline, we can account for individual variations in drawing proficiency, ensuring that the assessment of recall is not confounded by the participants' ability to draw.

Encoding Phase

A video selection algorithm was implemented in order to create a dataframe of video ordering and target selection unique to each participant. The algorithm ensured a balanced representation in each round, where two videos were highly memorable, and two were highly unmemorable. To introduce an element of randomness and mitigate the risk of pattern recognition by the participants, the order of these videos was randomised for each round. The algorithm also assigned a target video for each round. This target assignment was pseudo-random, meaning that it was randomly chosen, but in an increasingly constrained manner that ensured every video was assigned as a target at least once across all participants. This strategy ensured that all video used in the experiment would produce data that could be used to analyse its recall memorability, and that we could account for serial position and recency effects. The video selection algorithm can be represented in pseudocode as follows:

Algorithm 2 Select Videos for Experiment

```
procedure SelectVideos($num\_participants$, $list1$, $list2$)
    Randomly shuffle $list1$ and $list2$
    Initialize an empty list $data$
    Initialize $targets$ to be $list1 + list2$
    for $p$ in range of $num\_participants$ do
        Initialize $i, j$ to 0
        while $i <$ length($list1$) or $j <$ length($list2$) do
            if $targets$ is empty then
                $targets = list1 + list2$
                Randomly shuffle $targets$
            end if
            Create blocks of 4 videos
            Select a target video from $targets$ and remove it from $targets$
            While $target$ is not in $block$, repeat the target selection
            Find the index of $target$ in $block$, which is $target\_index$
            Append the row [p, block, target labels] to $data$
            Increment $i$ and $j$ by 2 accordingly
        end while
    end for
    Create a DataFrame $df$ from $data$ with columns 'Participant', 'Video1', 'Video2', 'Video3', 'Video4', 'Target1', 'Target2', 'Target3', 'Target4'
    return $df$
end procedure
```

The algorithm commences by shuffling both lists of selected videos (high and low memorability). It then iterates through each participant, forming blocks of four videos for each round. If both video lists still contain elements, a block is formed with two videos from each list, and the order within the block is randomised. A target video is then selected. If the initially chosen target video is not part of the current block, the algorithm continues to randomly select a target until it finds one that is in the block. Once a valid target is found, its index within the block is recorded, and the participant's id, the four videos, and the target index are stored as a row of data. If one of the lists is exhausted before the other, the remaining videos from the non-exhausted list form the remaining blocks, again with randomized order and target selection following the same methodology. After cycling through all participants and rounds, the resulting data is used to create a dataframe that contains the participant IDs, the videos presented to each participant in each round, and the target for each round. This controlled pseudo-randomised structure of video presentation is designed to ensure a balanced and unbiased experiment, while the random elements keep the experiment challenging and engaging for the participants. A total of 52 participants took part in the experiment, with a final 35 fully completing the experiment. Aside from exclusion following inclusion criteria: age 18-65, no cognitive impairment, no personal or immediate family history of epilepsy, no personal history of neurological illness or brain injury, no demographic information was collected. This decision was grounded in our aim to prioritise the central cognitive variables of the study and minimise potential biases in the interpretation of results. Collecting demographic data can inadvertently introduce an array of confounding variables, such as socio-cultural or educational influences on drawing styles and memory, which could detract from the primary focus of understanding recall mechanisms in the context of our experiment. Moreover, in line with the principles of equitable research and to prevent unintentional biases, it was deemed essential to approach the data without preconceptions linked to demographic factors. This ensures that any findings derived from this experiment pertain directly to the human cognitive processing of memory recall in the context of video-based stimuli, rather than being influenced by socio-demographic characteristics.

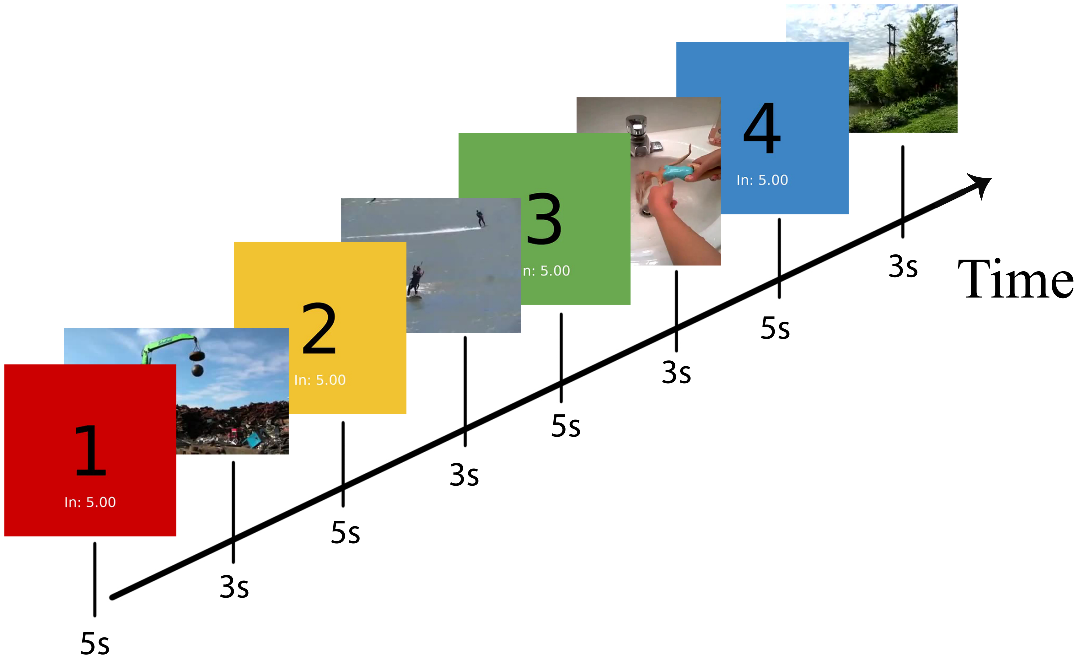


Figure 6.1: Encoding phase in online drawing experiment.

During the encoding phase, videos were displayed at their native resolution, ranging anywhere from 200px by 600px to 600px by 200px. Each video was 3s in duration, and they were presented following an anchoring 5s, consistently coloured, countdown, interstimulus interval as depicted in Figure 6.1.

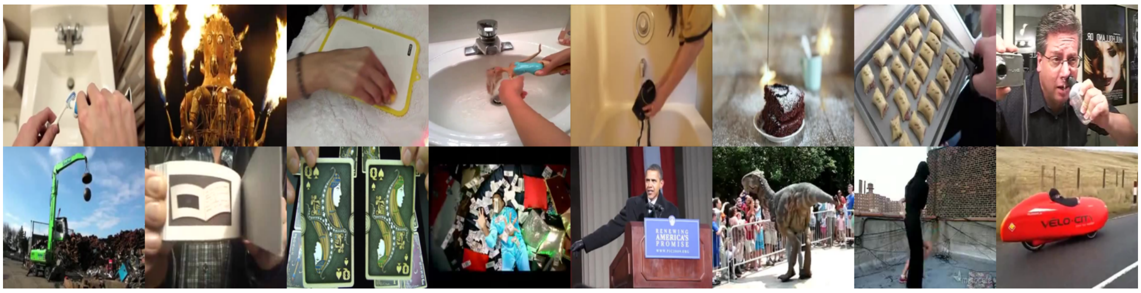


Figure 6.2: High memorability videos used in experiment (resized to fit into the grid with a 1:1 ratio).

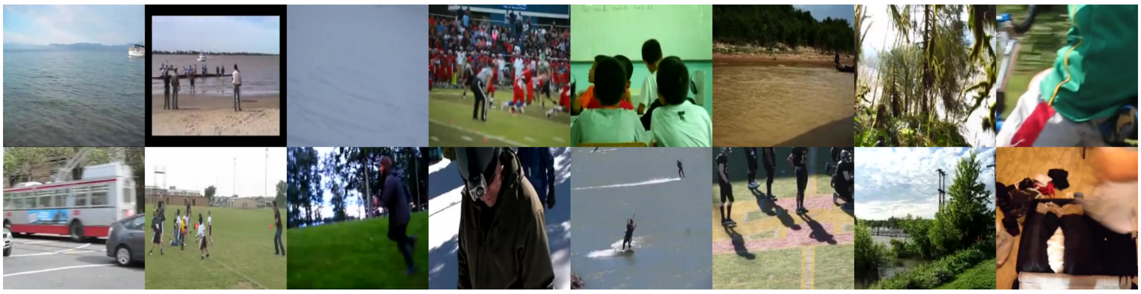
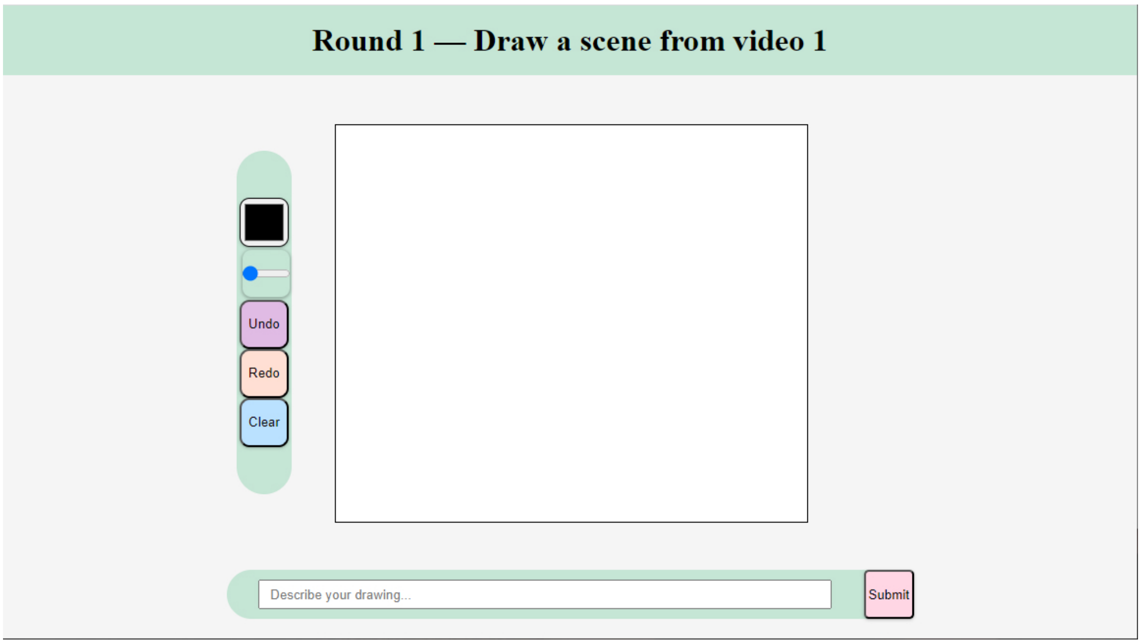


Figure 6.3: Low memorability videos used in experiment (resized to fit into the grid with a 1:1 ratio).

Recall Drawing Phase



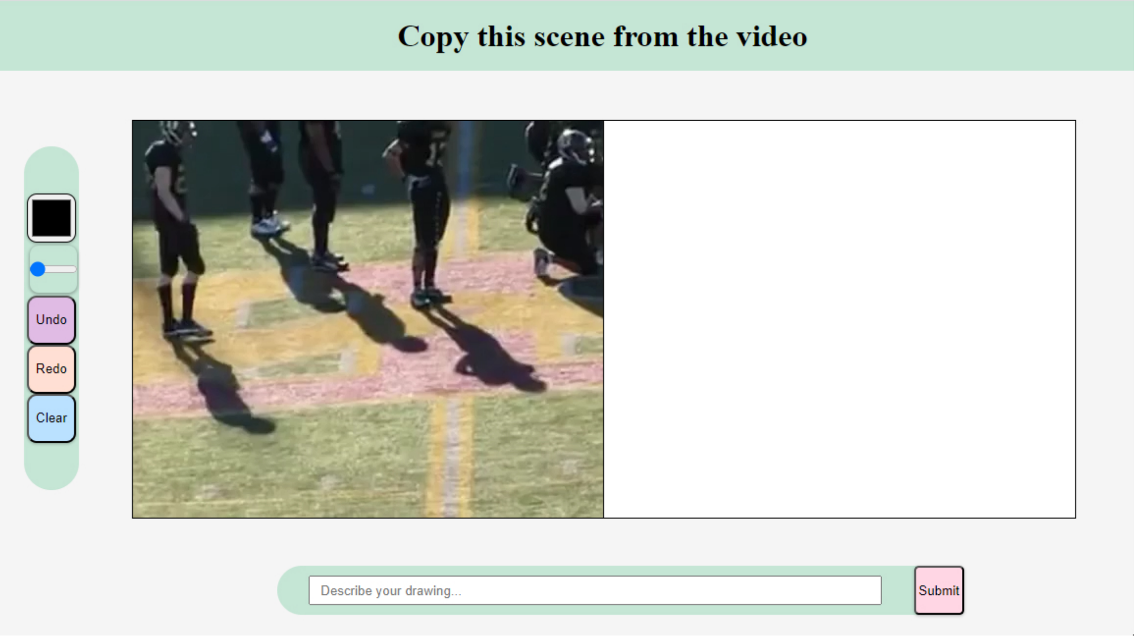
Round 1 — Draw a scene from video 1

Figure 6.4: Drawing page for video recall in online drawing experiment.

After all four videos in a round were displayed, participants were redirected to a drawing recall page, where they were instructed to draw a scene from the target video for that round, and then caption their drawing before submitting. As shown in Figure 6.4, the drawing recall page consisted of a heading which indicated the current round and the target video; a drawing canvas which was the same dimensions as the target video; a drawing toolbar enabled participants to change the colour of their brush, resize it, undo or redo an action, and clear the canvas; a caption bar for the participants to describe their drawing; and a submit button to move onto the next phase.

Perceptual Baseline

After submitting their recall drawing and caption, participants were redirected to a perceptual drawing page where they were instructed to copy a scene depicted from the target video, and then caption their drawing before submitting. As shown in Figure 6.5, the interface was the same as the previous phase, but this time a video scene was depicted adjacent to the drawing canvas.



Copy this scene from the video

Figure 6.5: Drawing page for perceptual drawing in online drawing experiment.

Vividness of Mental Imagery

Upon the successful completion of the eight rounds of the experiment, the participants are navigated to the concluding section of the study. This section features the Vividness of Visual Imagery Questionnaire (VVIQ), a psychological tool that was developed by David Marks in 1973, to assess the vividness of an individual's visual imagery [294]. The VVIQ is a self-report measure used to probe the subjective vividness of mental imagery in different individuals. It is predicated on the idea that the capacity to visualise varies widely among individuals and this variance can be systematically measured. The VVIQ is widely recognised in both psychology and neuroscience as an effective method of gauging this variable trait. The questionnaire consists of 16 items in which participants are asked to visualise four scenarios and rate the clarity and vividness of their mental imagery on a five-point scale. The scenarios involve familiar everyday experiences, such as the appearance of a friend, a rising sun, a shop they frequent, or a countryside landscape. Each scenario is rated on vividness in four different aspects, creating a total of 16 separate ratings. The five-point scale ranges from no image at all, which scores 1, to perfectly clear and as vivid as normal vision, which scores 5. Therefore, the total possible score ranges from 16 (poorest imagery) to 80 (most vivid imagery), with a score below 40 typically being an indication of some degree of visual mental imagery impairment, such as aphantasia. The inclusion of the VVIQ in this experiment offers valuable supplementary data on a participant's ability to mentally visualise. While it is not the main focus of the study, understanding a participant's ability to mentally visualise provides context to their drawing ability and potential drawing recall memorability.

6.3 Quantifying Recall

An essential aspect of assessing the efficacy of the drawing recall experiment lies in effectively quantifying recall. Common measures in recall experiments often include recall accuracy (the proportion of items correctly recalled), recall frequency (the number of times an item is recalled), and recall latency (the time taken to recall an item). While these metrics have proven valuable in providing insight into the efficiency and capacity of memory, they fail to capture the complexity and nuance inherent in the process of remembering, providing a relatively sterile and artificial view of memory processes, particularly when dealing with rich and dynamic stimuli such as videos. Traditional measures of recall focus on granular, often binary aspects, of remembering, reducing what it means to recall to the point where it can be quantified in a straightforward manner. When interacting with the world around us, our cognition is not meticulously cataloging the number of objects encountered or the precise order of events. Such an approach would not only be incredibly taxing but also provide limited utility. Instead, our cognition is inherently geared towards understanding the world in a holistic and meaningful way. What we recall is not merely a dry, objective list of details, but a fluid, interconnected web of information imbued with personal meaning, contextual associations, and emotional resonance. Additionally, the traditional metrics of recall often overlook the temporal narrative that unfolds in our memories. When we remember, we are not just recalling a static snapshot of an event but a dynamic, unfolding narrative that evolves over time. This narrative aspect is integral to our experience and understanding of the world, shaping not only what we remember but also how we remember it. Recall, especially of complex stimuli like videos, should therefore encapsulate the "what, where, and when"—the essence, meaning, or narrative of the perceived stimuli. This more comprehensive perspective, focusing on the holistic and semantic richness of recall, can offer a more ecologically valid measure of memory that is more in tune with our natural cognitive processes. A focus on the "what, where, and when" emphasises the importance of understanding the underlying content and context of our memories, moving beyond the simple and reductionist measures of recall, to a deeper understanding of the richness and complexity of our memory processes.

6.3.1 Semantic Similarity

In assessing recall within the framework of complex stimuli, such as videos, it is crucial to move beyond the narrow and overly simplistic scope of traditional recall measures. Traditionally, the primary dimensions of recall assessment have included the sheer quantity of accurately recalled items, the correctness of the sequence in which these items were originally presented, the spatial precision of the recalled items in relation to their original contexts, and the temporal accuracy of event recall within a designated timeline. These measures, while providing and easy-too-quantify metric, strip memories of the narrative and contextual associations that imbue them with value. To circumvent these limitations, a paradigm shift is necessary towards embracing a more encompassing approach: the measurement of semantic similarity. Semantic similarity transcends the reductionist perspective, taking into account the "what, where, and when" of memory. It provides a more ecologically valid, and holistic comparison between the perception of a stimulus and its reconstruction from memory. While the notion of quantifying a stimulus' semantics might initially appear daunting, recent advancements in machine learning and language processing present robust tools to tackle this challenge.

One such tool is the Contrastive Language–Image Pretraining (CLIP) model [309], which is a versatile and robust model that has been trained on a vast corpus of image and text paired data. Its unique training methodology allows it to understand and translate between visual and textual modalities, making it suitable for extract semantic information from both images (e.g., video frames, and drawings) and text (e.g., captions). The primary technique that facilitates this comparison is the generation and comparison of vector representations, or embeddings, of the input data. CLIP generates these embeddings by mapping input data—be it text or image—into a high-dimensional space where the semantic similarity between data points is represented as spatial proximity. That is, embeddings from similar pieces of data—regardless of whether they are text or image—are closer together, while those from dissimilar data are farther apart. The measure of semantic similarity is then computed using cosine similarity, a metric that quantifies the cosine of the angle between two vectors. With embeddings, the cosine similarity effectively captures the angle between the two vectors in the high-dimensional space. As vectors closer together in this space indicate greater semantic similarity, a smaller angle (and thus higher cosine similarity) means the two pieces of data are semantically more similar.

Within this study there are two main axes of recall from which we can measure semantic similarity: drawings and text.

6.3.2 Drawing Based Measures

As previously emphasised earlier in the chapter, drawings provide a window into the contours of a subject's mental terrain in ways that other methodologies may not permit, offering a tangible, visual output that embodies the idiosyncrasies of individual cognition. In the context of the drawing experiment, there are two distinct types of drawings: the drawings created by participants as a result of their recall—"recall drawings"—and the drawings produced while viewing a frame from the original video stimulus—"perceptual drawings". CLIP image embeddings for both of these types of drawings can be generated, enabling a straightforward semantic similarity between them to be calculated.

However, a key challenge arises when attempting to calculate the semantic similarity between a drawing and the ground-truth video frames. This comparison is anything but straightforward because the video frames are not drawings; they don't have the same properties and structural peculiarities inherent in human drawings. Hence, a direct comparison between a drawing and an image may not yield a useful measure of semantic similarity. To bridge this gap, we turn to a ControlNet [310] conditioned Stable Diffusion [311] model.

Stable Diffusion is a state-of-the-art open-source text-to-image model (covered in more detail in the next chapter) that can generate high-quality synthetic images. In combination with ControlNet, synthetic images that closely align with the underlying semantics and structure of the recall and perceptual drawings. These synthesised images serve as a "semantic bridge", allowing for a more valid calculation of semantic similarity between the drawings and the ground-truth video frames.

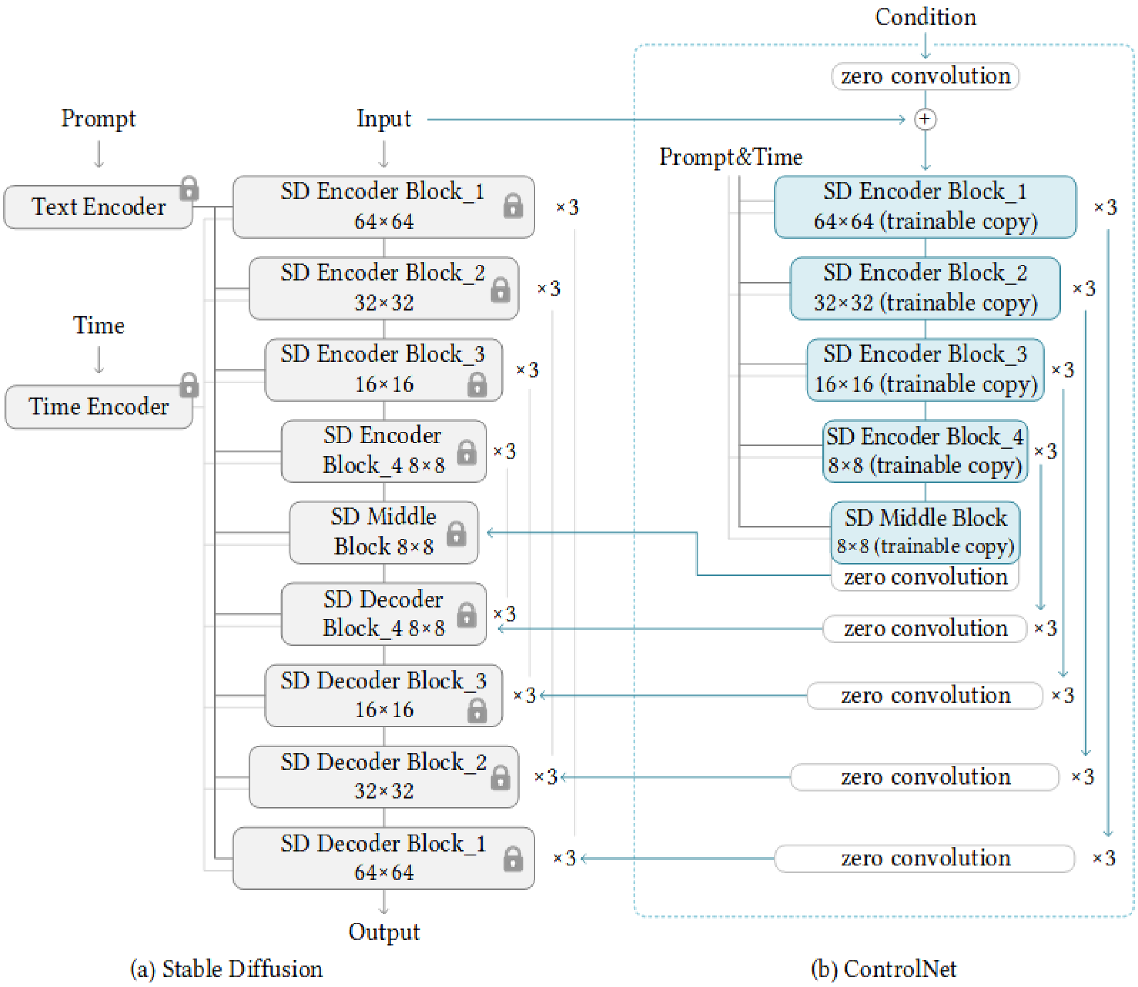


Figure 6.6: ControlNet in Stable Diffusion. Reproduced from [310].

ControlNet

ControlNet [310] is a neural network structure that imparts greater flexibility and control to large diffusion models used in text-to-image generation. The architecture is uniquely designed to support additional input conditions, significantly expanding the application potential of these models. At its core, ControlNet, shown in Figure 6.6, employs an end-to-end learning approach that involves the creation of a "trainable copy" and a "locked copy" of the pre-existing weights from a large diffusion model like Stable Diffusion [311]. The locked copy remains static, preserving the extensive knowledge previously learned from billions of images. In contrast, the trainable copy is dynamically adapted to learn from task-specific datasets, thereby facilitating conditional control. This strategy leverages the power of existing models while still allowing for customisation and adaptation to specific tasks. The trainable and locked copies of the network parameters are interconnected using a unique type of convolution layer termed “zero convolution.” This layer is initialised with zero weights and biases, ensuring that the initial application of ControlNet does not alter the deep features of the model. The layer’s parameters are progressively updated during training, which helps to maintain the speed and efficiency of fine-tuning a diffusion model. A ControlNet trained with human scribbles was used to generate images from the participants’ drawings.

Recall Drawings vs Perceptual Drawings

While this comparison offers potential valuable insights into the overall correspondence between recalled and perceived content, and at face value seems sensible and straightforward, it is not without its limitations. While the CLIP model typically excels at mapping visual data to a high-dimensional space, it can struggle with the inherent ambiguity and idiosyncrasies of hand-drawn images. For instance, it is vulnerable to producing high similarity scores between drawings with minimal semantic content.

Consider two drawing samples produced by a participant (Figure 6.8). Both consist of black scribbles. Despite a lack of discernible semantic features in these drawings, the CLIP similarity score between them is 1.0. However, this actually makes a lot of sense if we consider what the model is doing, and the fact that the sole semantic quality of both drawings—which they equally share—is being a black scribble. In the absence of more complex semantic features, this will be the case for any two images that share colour characteristics. This highlights the importance of participant drawing ability for this specific vector of analysis.

If a participant produces drawings with a high level of detail, CLIP can extract a more meaningful higher-dimensional representation, and accordingly more nuanced and accurate semantic similarity scores can be calculated. The use of colour can also play an impactful (albeit less crucial) role, providing an additional layer of

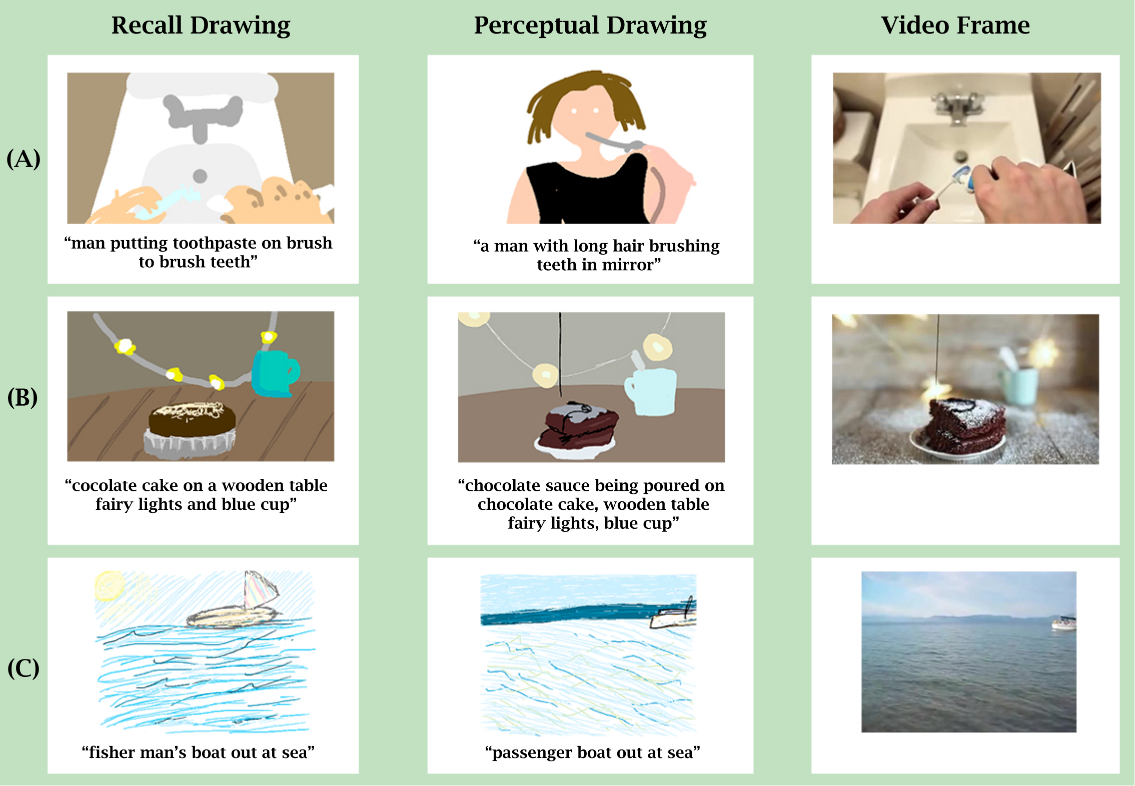

Recall Drawing
Perceptual Drawing
Video Frame
(A)
"man putting toothpaste on brush to brush teeth"
"a man with long hair brushing teeth in mirror"
(B)
"cocolate cake on a wooden table fairy lights and blue cup"
"chocolate sauce being poured on chocolate cake, wooden table fairy lights, blue cup"
(C)
"fisher man's boat out at sea"
"passenger boat out at sea"

Figure 6.7: Example participant recall and perceptual drawings with their associated captions, alongside ground-truth video frames.

information, potentially leading to more accurate similarity scores. Interestingly, for participants who demonstrated a high degree of drawing ability, even without consistent use of colour, this approach proved to be quite effective, as demonstrated in Figure 6.9. An analysis of this subset of high-quality participant drawings revealed a subtle difference in the semantic similarity scores based on the recognition memorability of the videos. High recognition memorability videos yielded a slightly higher average semantic similarity score ($\bar{x} = 0.73, SD = 0.06$) compared to low recognition memorability videos ($\bar{x} = 0.65, SD = 0.07$). This difference, while small, approached statistical significance ($t = 2.09, p = 0.051$). This weak correlation could potentially be explained by a phenomenon reminiscent of the "Matthew effect" in the field of cognitive psychology [312], a term borrowed from sociology that refers to an accrued advantage phenomenon, where, for instance, early advantages in reading ability lead to increased reading experiences, further enhancing the skill and widening the gap between proficient and struggling readers. In the context of memory recall, this suggests that videos with high recognition memorability,

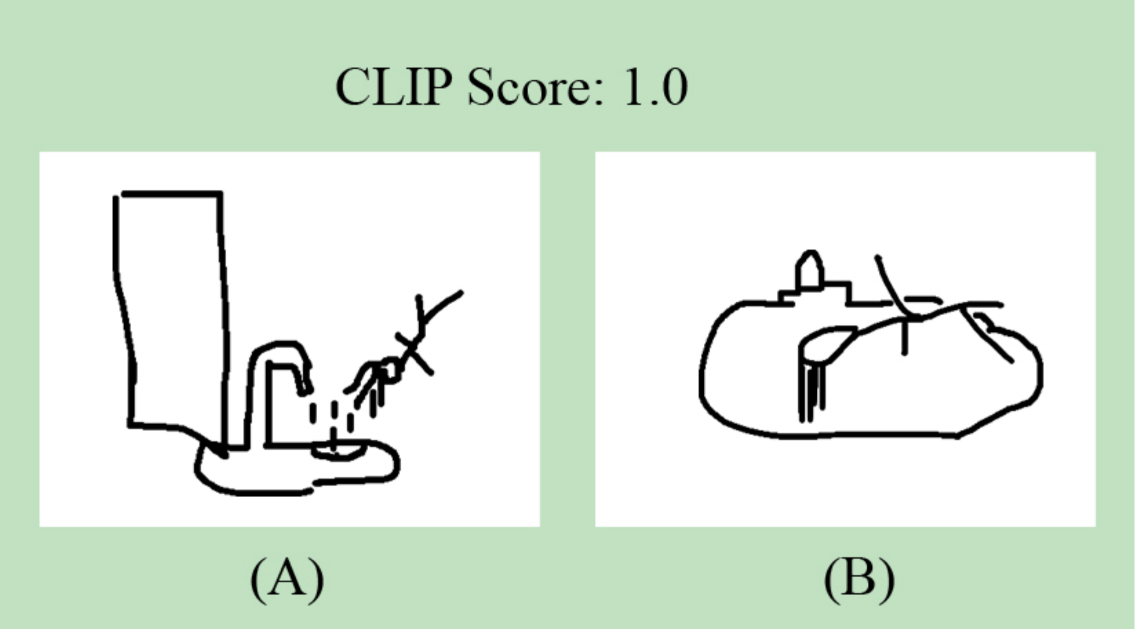


Figure 6.8: Example Similarity score between a participant's recall (A) and perceptual (B) drawings.

due to their distinct and memorable content, may stimulate more comprehensive and precise drawings. However, given the small effect size and marginal statistical significance, this finding should be interpreted cautiously.

Recall Drawings vs Ground-Truth Video Frames

As previously mentioned, a direct comparison between drawings and video frames is unlikely to be of much utility as they have drastically divergent visual and structural properties. Accordingly, a ControlNet conditioned Stable Diffusion model was leveraged to create high-fidelity image representations of the participants' recall drawing. The intent was to transform the relatively low resolution and potentially abstract recall drawings into more detailed images, which can be compared more effectively with the actual video frames. The process of generating the surrogate images involves feeding the caption into the Stable Diffusion model and feeding the recall drawing into the ControlNet. The caption is used to convey the desired conceptual properties of the synthesised image, and the drawing is used to guide its structural composition.

Once the synthetic image has been generated, it is then compared against the

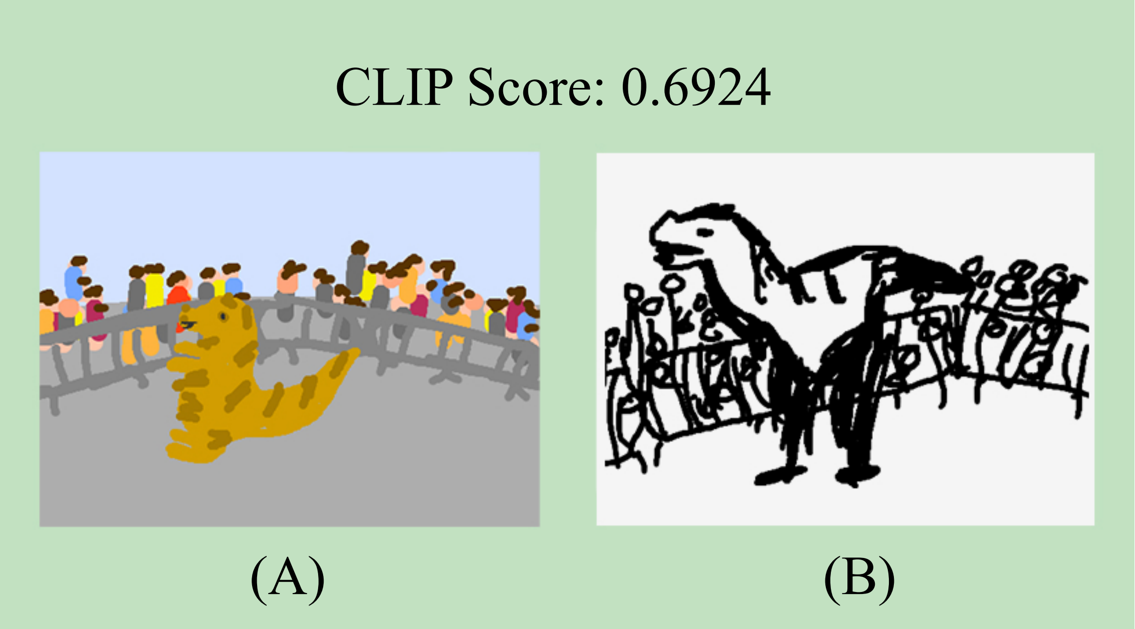


Figure 6.9: Example of similarity score between recall (A) and perceptual (B) drawings.

first, middle, and last frames of the ground-truth video to compute semantic similarity scores. This was done to account for any temporal changes in the video's narrative content, thereby providing a more comprehensive and accurate representation of the video's overall semantics. The final similarity score was chosen as the highest score from these comparisons, representing the closest match between the synthetic image and the video frames. While the synthesis of recall-drawing-based surrogate images facilitated a comparative analysis with actual video frames, the results were somewhat mixed. A weak but statistically significant positive correlation was observed between the semantic similarity scores and the recognition memorability of the videos ($r = 0.256$, $p = 0.018$). In other words, videos that were more memorable (high category) tended to have higher semantic similarity scores compared to less memorable (low category) videos. More specifically, high recognition memorability videos yielded an average similarity score of 0.66 (SD = 0.07), slightly higher than the low memorability videos which averaged at 0.61 (SD = 0.06). A t-test performed on these averages did not yield a statistically significant difference ($t = 1.56$, $p = 0.124$).

A second method of comparison considered both recall-drawing-based synthetic

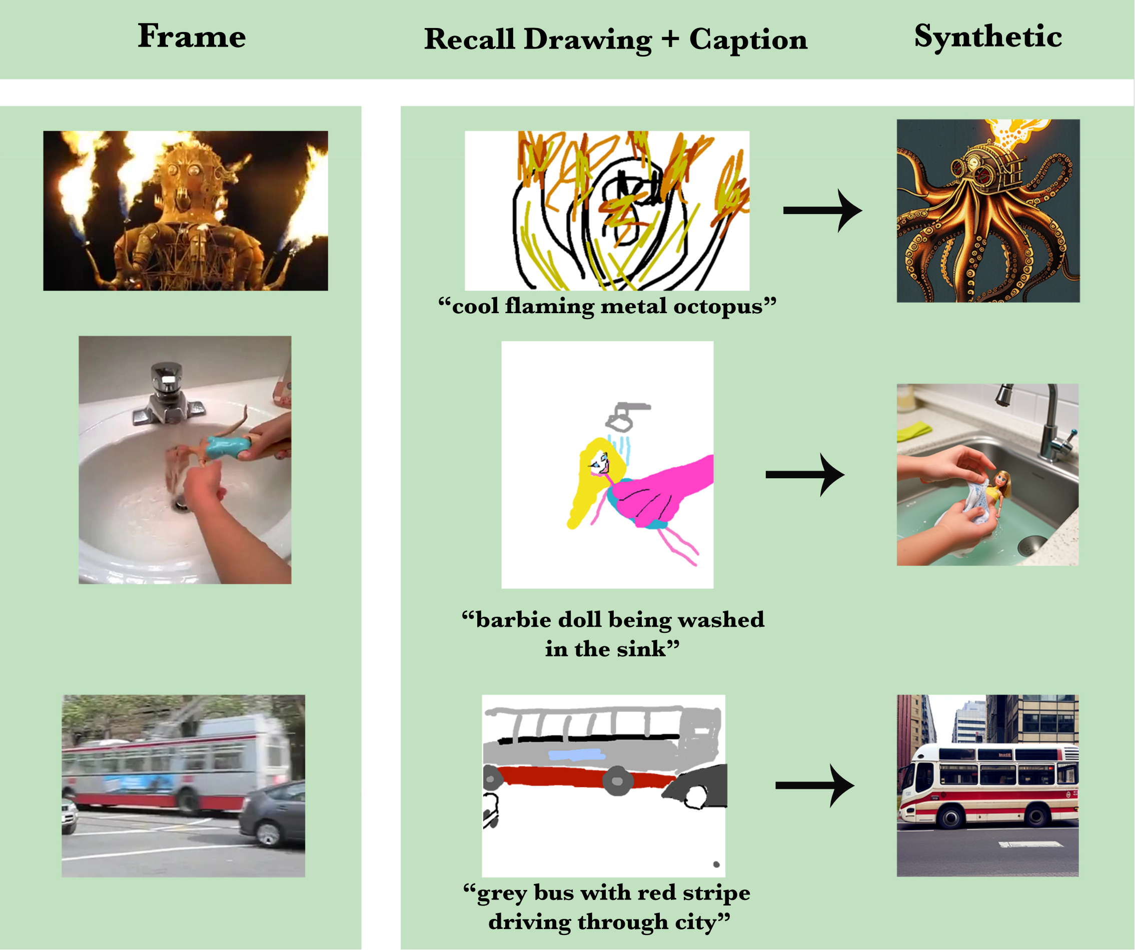

Frame
Recall Drawing + Caption
Synthetic
"cool flaming metal octopus"
"barbie doll being washed in the sink"
"grey bus with red stripe driving through city"

Figure 6.10: Examples of participant recall drawings and captions, and the resultantly synthesised images.

images and synthetic images generated for the ground-truth videos. Three synthetic images—using the first, middle, and last frames—were generated for the ground-truth videos by passing the first ground-truth caption and a frame as inputs to the ControlNet conditioned Stable Diffusion model. This process facilitated a more direct comparison between the semantic interpretations of the recall and ground-truth content, effectively quantifying the fidelity of the recalled information to the original video's semantics. The final similarity score between a recall drawing and video was chosen from the highest comparison score between the synthetic recall image and each of the synthetic video-frame-based images. Upon analysis, a more pronounced, statistically significant positive correlation was found between the semantic similarity scores and the memorability of the videos ($r = 0.563$, $p < 0.003$). More memorable videos (high category) consistently had higher semantic similarity scores compared to less memorable videos (low category). In terms of mean semantic similarity scores, a statistically significant difference was noted between the high and low memorability videos. Specifically, high memorability videos demonstrated an average similarity score of 0.76 (SD = 0.07), which was significantly higher than that of low memorability videos, which had an average score of 0.68 (SD = 0.06), $t = 3.14$, $p < 0.011$. This stronger correlation and distinct difference in means in this method of comparison suggest a more evident relationship between video memorability and semantic alignment. The inclusion of the synthetic images representing the ground-truth videos potentially provides a more accurate gauge of the semantic consistency between recall and original video content.

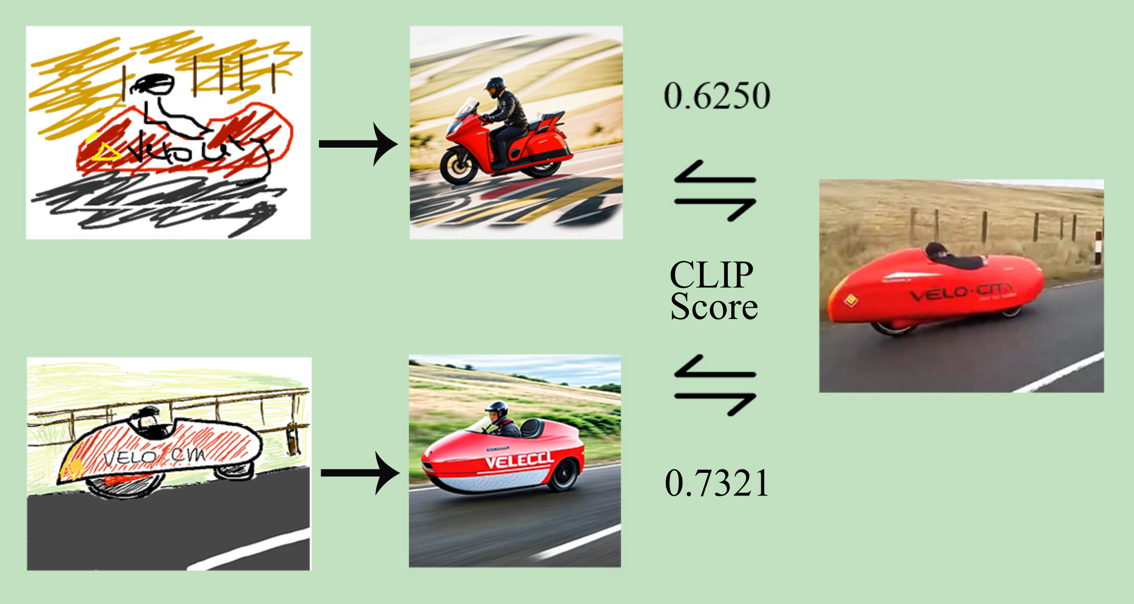


Figure 6.11: Example of synthetic images generated from recall drawings, and their CLIP scores to a ground-truth video frame.

6.3.3 Textual Measures

In the context of the drawing recall experiment, textual measures serve as an illuminating counterpart to our visual measures, capturing nuanced details of remembered stimuli that may not find expression in visual representations. The CLIP model, with its ability to evaluate the semantic similarity between text and image data, enables the additional assessment of textual representations of participant recall.

Three axes of comparison are considered:

Recall Captions vs Perceptual Captions: This comparison reflects the fidelity of recall, gauging the degree of correspondence between the semantics perceived and those recounted from memory. In the study, a marked distinction was observed between high and low recognition memorability videos. High memorability videos exhibited significantly greater semantic similarity between recall and perceptual captions ($\bar{x} = 0.833, SD = 0.064$) compared to low memorability videos ($\bar{x} = 0.674, SD = 0.051$; $t = 7.81, p < 0.0002$). These findings suggest that the recognition memorability of a video bears a strong influence on the semantic alignment between perceived and recalled stimuli.

Recall Captions vs Ground-Truth Captions: This comparison provides a more external assessment of recall accuracy, reflecting the degree of semantic congruence between the recalled content and the original video narrative. Interestingly, despite a difference in the mean similarity scores between high recognition memorability videos ($M = 0.67, SD = 0.05$) and low recognition memorability videos ($M = 0.64, SD = 0.06$), this difference did not reach statistical significance, $t = 1.90, p = 0.067$. This indicates that, although there is a trend for recalled captions of more memorable videos to align more with the ground-truth captions, this trend is not strong enough to yield a significant correlation with video recognition memorability. This could be attributed to the inherent variability in individual recall strategies and the natural transformation of information as it moves from perception to recall. Moreover, the baseline level of specificity in participants' caption writing, or their general descriptive ability, could heavily influence this comparison. This factor potentially muddies the waters and prevents any underlying effects from surfacing.

Normalised Recall-to-Ground-Truth Similarity: Accounting for individual differences in perception and descriptive ability, an adjusted measure of recall accuracy can be derived. This is achieved by normalising the recall-to-ground-truth similarity by the perception-to-ground-truth similarity. This ratio highlights how effectively a participant's recall aligns with the original video content, after factoring in their initial perceptual and descriptive ability. Examination of this normalised measure revealed a robust correlation between the normalised similarity score and video memorability ($r = 0.723, p < .0027$). High memorability videos yielded an average normalised similarity score of 0.819 ($SD = 0.038$), significantly higher than the average score of 0.667 ($SD = 0.053$) observed for low memorability videos ($p < .0154$). The emergence of this correlation after normalisation suggests that, while unadjusted recall may not consistently reflect the recognition memorability of a video, the extent to which recall preserves original perception appears to be strongly linked to the recognition memorability of the video content. The findings underscore the importance of considering individual differences in perception and descriptive abilities when evaluating recall performance.

Recall Caption Precision

A quantifiable measure of recall precision, the Caption Specificity Score (CSS), was introduced to assess the level of detail and specificity inherent within the captions produced during the recall phase of the experiment. The calculation of the CSS was predicated on the integration of two key measures: Average Term Frequency-Inverse Document Frequency (Avg TF-IDF) and Named Entity Count (NEC). The Avg TF-IDF was computed using standard natural language processing (NLP) procedures. This involved tokenization, case normalisation, and punctuation removal, applied to a corpus composed of the entire Google Conceptual Captions dataset [201]. Each unique term within a recall caption was subsequently assigned a score that was indicative of its relative significance within the caption and its rarity within the corpus, thereby facilitating the computation of Avg TF-IDF. The NEC complements the Avg TF-IDF by focusing on the level of detail of the caption. Named entities refer to definite nouns that correspond real-world objects, individuals, locations, etc. To compute the NEC, the default implementation of the Named Entity Recognition (NER) system in the Python spaCy library [313], was used. The final CSS assigned to each recall caption was computed by integrating the normalised Avg TF-IDF and NEC. Both components were scaled to lie within the range of 0 to 1, with the CSS for a given recall caption calculated as:

$$\text{CSS} = \text{normalised}(\text{Avg TF-IDF}) + \text{normalised}(\text{NEC}) \tag{6.1}$$

The careful examination of CCSs revealed more nuanced relationships with the variables of interest. This deeper dive uncovered the interactions between recall caption specificity, ground-truth caption specificity, and the video recognition memorability categories. To capture these complexities, a relative comparison measure was devised, based on the difference between the recall CSS and the ground-truth CSS, normalised by the difference between the perception CSS and the ground-truth CSS. This created a score that encapsulated the change in specificity from perception to recall, relative to the ground-truth. Analysing the normalised CSS yielded several notable outcomes. A moderate and statistically significant positive correlation was observed between the normalised CSS and high video recognition memorability, with $r = 0.36$, $p = 0.009$. This finding indicates a link between recall caption precision and the recognition memorability of the videos. Specifically, it suggests that for videos categorised as highly memorable, the recall caption specificity more closely matched the ground-truth caption specificity relative to the initial perception. Furthermore, an interesting trend emerged when comparing the recall and perceptual caption specificity within the recognition memorability categories. The average difference between recall caption specificity and perceptual caption specificity was smaller for high recognition memorability videos compared to low recognition memorability videos, $t = 2.18$, $p = 0.037$. These observations suggest that recognition memorability might be linked to the quality and detail of recall. However, it should be acknowledged that high recognition memorability videos might inherently contain more unique, detailed, or rich concepts, which could potentially influence the observed differences in caption recall precision.

6.3.4 Other Measures

Alongside the primary analysis measures, the recall drawing experiment incorporated two additional measures to enrich the understanding of video recall memorability and its influencing factors, the first of which addressed instances of forgotten or misremembered videos. During the experiment, participants who could not remember a particular video left the drawing canvas blank, and typically wrote a statement like “I don’t remember” in the caption. Videos that were misremembered were identified by comparing the recall drawing and captions with the corresponding perceptual drawing and captions. Interestingly, none of the videos in the high memorability category were forgotten or misremembered. For the low memorability category, there were nine instances of videos being forgotten or misremembered. Of these, one video was forgotten/misremembered by three participants, two were forgotten/misremebered by two participants. Notably, all misremembered instances involved a video from the encoding phase positions 2 or 3 being confused for a high memorability video in the corresponding 2$^{\text{nd}}$ or 3$^{\text{rd}}$ position. A subsequent Z-test for the difference in proportions of correctly recalled videos between high and low recognition memorability videos revealed a significant difference ($Z = 3.0542, p = .00228$). This difference in recall proportions between high and low recognition memorability videos provides further evidence for the existence of a relationship between recognition memorability and recall memorability.

The second measure involved participants completing a Vividness of Visual Imagery Questionnaire (VVIQ) following the experiment. As explained in section 6.2.1, this questionnaire provides an insight into individual differences in the ability to form mental visual images, which could potentially influence the quality of recall and the semantic similarity of recall drawings. The distribution of reported VVIQ scores largely aligned with what is expected in the general population. Only one participant reported a complete absence of mental visual imagery, scoring a 16 on the VVIQ. This specific participant’s drawings were not discernibly different to the average drawing, and did not result in any drawing score outliers. Additionally, a Pearson correlation analysis revealed no significant direct relationship between VVIQ score and any drawing recall score measures—recall vs perceptual drawings, synthetic recall images vs ground-truth images, and synthetic recall images vs synthetic ground-truth images, $r = -0.086$, $p = 0.182$ $r = 0.073$, $p = 0.235$ $r = 0.065$, $p = 0.275$, respectively. However, independent-samples t-tests showed a significant difference in mean VVIQ scores between participants in the top quartile (¿64) and those in the bottom three quartiles across the three measures of CLIP similarity scores. For recall versus perceptual drawings $t = 2.28$, $p = 0.026$; for synthetic recall images versus ground-truth images, $t = 2.15$, $p = 0.034$; and for synthetic recall images versus synthetic ground-truth images, $t = 1.98$, $p = 0.049$. These results suggest an interesting interplay between personal cognitive abilities, namely the capacity for vivid mental imagery, and the fidelity of memory recall. While there may not be a strong direct correlation between the ability to form mental images and recall drawing scores when examined across all participants, within the group that scored in the top quartile for VVIQ, there is a noticeable enhancement in the semantic precision of the recalled information in their drawings. It should be noted that the observed relationships might not entirely reflect the quality of recall, but could potentially be indicative of a participant's ability to accurately depict their mental representation through the medium of drawing. In other words, higher VVIQ scores might be more closely associated with better representational abilities rather than superior recall per se.

6.4 Conclusion

The hypothesis investigated in this chapter posited a measurable relationship between recognition and recall memorability (H3). The empirical evidence garnered from visual and textual measures robustly validates this hypothesis, underscoring the existence of a significant link between these two facets of memorability. Visual measures revealed marked disparities in the semantic alignment and precision of recall between videos categorised as high and low in terms of recognition memorability. Notably, a strong correlation was observed between the normalised measure of recall accuracy, which accounts for individual perceptual and descriptive abilities, and video memorability. This finding underscores the impact of individual differences in moulding recall performance, further substantiating H3.

In the textual domain, the Caption Specificity Score (CSS) was introduced as a novel, quantifiable measure of recall precision. Intriguingly, the analysis of CSS scores unveiled a nuanced relationship with recognition memorability. A statistically significant correlation was observed between the normalised CSS and high video recognition memorability. This lends credence to the notion that recall precision plays a crucial role in measuring the recognition memorability of video stimuli. Importantly, the absence of forgotten or misremembered videos in the high memorability category augments the body of evidence in favour of a correlation between recognition and recall memorability. This finding resonates with the research undertaken by Broers and Busch, providing converging evidence in support of H3. Further, an analysis of the Vividness of Visual Imagery Questionnaire (VVIQ) scores suggested an intricate interplay between individual cognitive abilities, specifically the capacity for vivid mental imagery, and the fidelity of memory recall. While this relationship was not universally observed across participants, a significant enhancement in recall precision was noted among participants in the top VVIQ quartile. Although this result is preliminary, it hints at the importance of accounting for personal cognitive abilities when examining recall performance. Interestingly, the results obtained call into question the measures of recall employed by Bainbridge et al. [152], as the findings contradict one another, ultimately underlining the necessity for more extensive investigation and the development of more robust, comprehensive measures of recall.

In summary, this chapter provides evidence that supports H3, illustrating a measurable correlation between recognition and recall memorability. The support of this hypothesis paves the way for further research exploring the cognitive and neural underpinnings of recognition and recall, thereby promising to unlock new insights in the realm of memory studies.

Chapter 7

The Conceptual Essence of Intrinsic Memorability

> *"Memory is not just the imprint of the past time upon us; it is the keeper of what is meaningful for our deepest hopes and fears."*
>
> — Rollo May, *Musicophilia*

In the unfathomable panorama of the cosmos, we find ourselves as intricately complex creatures on a serendipitous blue sphere, a miraculous haven teeming with life. Amidst the enthralling mystery of our existence, a fundamental truth emerges—we are living beings, bound by the immutable laws of biology, borne of the unyielding process of evolution. Our evolutionary history—profound, tumultuous, and enduring—sculpts the foundation of our biology, behaviour, and cognitive faculties, including our remarkable ability to remember [314].

In every waking moment, we find ourselves barraged by a relentless torrent of visual stimuli—vivid hues, intricate shapes, and compelling patterns. A select few breach the bulwarks of our consciousness, leaving an indelible imprint on our memory. What endows these chosen stimuli with their unforgettable memorability? One might suppose that it lies in their intrinsic attributes—their aesthetic allure, their captivating novelty, their emotive potency. Yet, we find that the reality is subtler and far more fascinating. As beings crafted by the relentless forge of evolution, our capacity to remember, to store and retrieve information, serves a fundamental purpose. It is not merely a backward-looking chronicle of our past experiences. Instead, it serves as a forward-facing oracle, allowing us to predict and prepare for the unfurling tapestry of the future [117], [315]. From an evolutionary perspective, this predictive capacity is invaluable—organisms that can anticipate danger or opportunity, navigate complex social landscapes, or detect changes in the environment, possess a critical edge in the fierce struggle for survival [316], [317].

Thus, the memorability of a stimulus is not anchored in the stimulus itself but rather in the profound mental representation it evokes within us. It lies in the capacity of the stimulus to resonate with our cognitive apparatus, to spark connections with our understanding of the world, and to contribute to our future predictive ability. A memorable stimulus is one that harmonises with our evolutionary imperative to anticipate, to forecast, and to prepare. The true essence of memorability, then, is found not in the external perceptual appeal of a stimulus but in the conceptual echo it creates within the chambers of our mind. This intimate relationship between memory and prediction illuminates why certain stimuli stand out in the colourful kaleidoscope of our perceptual experience.

Our minds serve not merely as passive observers but as active composers of a perceptual symphony, intricately transforming sensory data into elaborate mental representations [318]. This transformation embeds not only the physical attributes of stimuli but also an overarching conceptual essence. In this orchestration, the resonance of a stimulus with its conceptual connotation is the primary fuel for its memorability. Thus, the memorability of visual stimuli depends less on their physical features and more on their capacity to convey pertinent, utility-laden information—information that resonates with the perceptual apparatus of the evolved human observer. This perspective dovetails with insights drawn from diverse research domains. Linguistic evolution studies, for example, posit that our communication tools, including language, have evolved to deliver information efficiently [319]. In concert with this, cognitive neuroscience elucidates that our brains are inherently wired to prioritise stimuli that are rewarding [320], and abundant in information content [321]. These findings fortify the central premise that the cognitive value of stimuli, encompassing their memorability, is intimately entwined with their informational utility.

This understanding of memorability, steeped in the nuances of biology, cognition, and evolution, heralds new insights for computational memory research. It invites us to shape artificial systems that more faithfully reflect the profound structure of our biological memory, ultimately advancing our comprehension of both natural and artificial intelligences [322]. This discourse also imparts a salient lesson: it beckons us to view our perceptual experience not as a passive registration of sensory inputs, but as an active and dynamic engagement with the world around us—an engagement that is imbued with meaning and purpose and is intimately tied to our very existence. In this sense, every act of remembrance underscores our existence as biological entities, extracting meaningful information from the sensory deluge that envelops us. In this rich context, studying memorability emerges as more than an exploration of cognitive faculties—it unfurls as a journey into the essence of our shared humanity.

7.1 The Nature of Visual Memorability: Concepts over Visuals

The intricate relationship between visual content and memorability has been the subject of rigorous investigation over recent years, leading to a myriad of theories focusing on the intrinsic characteristics of visual stimuli that may influence their memorability. However, while the exploration of these physical, aesthetic elements have undoubtedly contributed to the depth of understanding within the field, they are predominantly rooted in the realm of static imagery [1], [141], [323], only very recently delving into the dynamic, rich, and complex realm of video content, which necessitates a recalibration of the theories of memorability to encapsulate this broader spectrum of stimuli.

Conceptually, the work described in this chapter takes a novel approach, focusing on the memorability of visual content as largely driven by the underlying concept it conveys, rather than its inherent visual features. This shift in perspective has hitherto not been explored or experimentally validated. This work does however align with the idea that content is memorable if it has a high utility of information. As posited by Bylinskii et al. [284], humans are more likely to remember that which surprises them, contradicts their current model of the world, or is likely to be relevant or useful in the future. This notion fundamentally shifts the focus from the visual features of an image or a video to the information it delivers, highlighting the crucial role of cognitive factors in shaping memorability. In essence, it is something akin to the distinctiveness of the concept relative to its context that makes a visual stimulus memorable. This suggests that our memory is primed to retain information that has a high utility, possibly because this type of information enables us to better prepare for future encounters and to adjust our worldview accordingly. The studies leveraging large, diverse memorability datasets, as described in [284], underscore this point further by revealing what universally tends to be memorable: emotional/affective stimuli, unexpected actions, social aspects, animate objects (human faces, gestures, interactions, etc.), and tangible objects.

In this light, it becomes clear that memorability transcends aesthetics or low-level visual features like colour or contrast. Instead, it encapsulates the higher-level properties of semantics (objects and actions) and composition (layout and clutter) in an image or video. In the context of video memorability, this proposition takes on an even greater importance. As we will explore in this chapter, the experiments reveal that machine learning models trained exclusively on synthetic data based on the underlying concept portrayed in the video, achieve state-of-the-art video memorability prediction. These models outperform their counterparts trained directly on frames extracted from the videos, thereby corroborating the central hypothesis of this chapter. Namely, that the memorability of visual content lies in the underlying concept and the mental representation it induces, not the specifics of its visual characteristics. As such, this chapter will further elucidate the cognitive and conceptual basis of video memorability, bringing to the fore the importance of conceptual information over visual specifics in shaping our visual memory. The implications of this for the broader field of memorability research are considerable, with the potential to redefine our understanding of what makes visual stimuli memorable.

7.2 Distilling Intrinsic Memorability

Although much progress has been made thinning the query-saturated haze that conceals the landscape of answers mapped by the seminal question: "What makes an image memorable?" [141], [192], [211], [324], the summit remains out of sight, with 25% of the variance (i.e., the proportion of total variability in memorability not yet explained by the identified factors) still remaining unaccounted for [1]. The shortest path to understanding is through a hurricane of light: given that we are visually dominant creatures, with over half of the cortex involved in visual processing [325], we naturally expect visual sensory data to exert the greatest influence on memorability. However, it is important not to be lead awry by our brain's appetite for visual sensory soup, as semantic meaning is known to play a critical role in visual memorability. Richer and more conceptually distinctive events last longer in memory, and certain semantic categories are inherently more memorable than others [1], [326]. Even though visual memories are stored with an exceptional fidelity of detail (i.e., configurations and contexts of viewed objects[140]), our performance is poor when it comes to remembering random patterns unless they take on object-like qualities [327], suggesting that visual memory is not driven entirely by visual details.

Further evidence suggests that visual data is merely a means to conceptual understanding, which is in turn intimately tied to memory, with conceptual distinctiveness supporting higher fidelity visual long-term memory representations than perceptual distinctiveness, and influencing memory retention in a manner that cannot be accounted for by perceptual distinctiveness alone [326], [328]. Perceptual distinctiveness is typically measured within a given object category, and with reference to variations in low dimensional, knowledge agnostic, perceptual features (i.e., colour, and shape). Unfortunately, the line between perceptual and conceptual features begins to blur as we move into higher dimensional features (e.g., length of torso relative to head size), which become more category-specific and likely to be acquired through visual experience [329], making it difficult to probe the depth of connection between concept and memorability. However, with the recent explosion in progress in the image synthesis field, and the release of the open-source text-to-image diffusion model *Stable Diffusion*, we find ourselves uniquely positioned to assess the impact of conceptual features on video memorability independent of its perceptual features, with the exceptional ability to preserve the depth and richness of information inherent to the visual domain. Introducing synthetic images offers a distinct advantage over direct textual encoding of captions in this context. While text provides a linear and descriptive representation of a concept, it inherently lacks the nuanced, multidimensional interplay of visual features present in images. A simple caption might be able to convey the general gist or theme of a video, but it fails to capture the subtleties and visual relationships within stimuli that may contribute to memorability. Synthetic images, on the other hand, allow for a richer, more detailed representation, maintaining (as a product of having been trained on very large corpora of visual data), but not directly preserving, the intricacies of visual data that text simply cannot. By employing synthetic images, we ensure a robust and comprehensive exploration of the relationship between conceptual features and their impact on memorability, without the oversimplifications and limitations associated with textual representations.

As captured in (H5), if visual data truly is merely a means to conceptual understanding, and that it is the concept itself—which is conveyed/represented through the visual data—that holds the content's intrinsic memorability, then the inter-video relationship of memorability scores predicted with ground-truth video frames should be observable in the memorability scores predicted with synthetic images predicated on purely conceptual video data.

The investigation begins by leveraging state of the art image synthesis to facilitate the exploration of the aforementioned hypothesis, which can be concisely captured with the following question: can the intrinsic memorability of visual content be distilled to its underlying concept/meaning?

7.2.1 Overview of the Stable Diffusion Model

Latent Diffusion models, a novel addition to the generative modeling landscape, have emerged recently as a powerful tool for high-quality image synthesis tasks. Stable Diffusion, a meticulously crafted configuration of these models, offers an illustrative example of the capabilities of this new approach [311]. By intertwining three distinct and potent components—–a Variational Autoencoder (VAE) for perceptual image compression, a U-Net for diffusion, and a CLIP ViT-L/14 text encoder for text-to-image conditioning—Stable Diffusion achieves an impressive blend of performance and fidelity. The following sections delve into the intricacies of these components and their specific roles in the functioning of Stable Diffusion.

Perceptual Image Compression: The Power of Variational Autoencoders

One critical step in Stable Diffusion, and indeed, in any latent diffusion model, is the encoding of data into a compressed latent space. A latent space refers to a lower-dimensional representation of the training data, constructed such that it captures the essential, semantically significant features of the data. This encoding allows the model to focus on the most meaningful aspects of the data and improves computational efficiency by reducing the dimensionality of the training space. The task of encoding the data into a compressed latent space is entrusted to the Variational Autoencoder (VAE). A VAE is a type of generative model that performs a dual function: it not only encodes the input data into a compressed form but also is capable of generating new data that resemble the input [330]. The VAE accomplishes this through a two-part architecture. The first part, the encoder, is a neural network that takes as input the data and outputs a set of parameters defining a probability distribution in the latent space. This distribution is intended to capture the essential features of the input data. The second part, the decoder, is another neural network that takes as input a point sampled from the latent space distribution and outputs a reconstruction of the original data as shown in Figure 7.1.

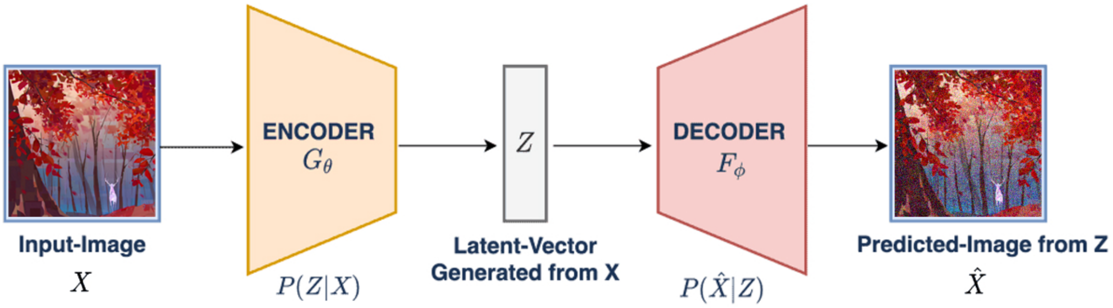


Figure 7.1: Variational Autoencoder architecture.

One of the distinguishing aspects of VAEs is the principle of "variational" inference, a method used to approximate complex probability distributions. In the context of the VAE, variational inference allows the model to encode the data into a standard Gaussian distribution in the latent space. This encoding provides a structured, normalized form for the latent space that facilitates the learning process of the diffusion model [330]. The Stable Diffusion model leverages the capabilities of the VAE to construct an efficient and semantically rich latent space. The use of VAEs brings a critical advantage: it decouples the learning of the latent space from the training of the diffusion model. This separation results in a more controlled learning process, avoiding a delicate balancing act often needed when simultaneously optimising reconstruction quality and learning the prior over the latent space [311]. Consequently, Stable Diffusion can concentrate on the distribution learning task, with a carefully curated, VAE-defined latent space serving as its starting point. This latent space, rich in perceptually significant features and lean in dimensionality, forms an ideal platform for the diffusion model to build upon, leading to more nuanced and high-quality image synthesis.

Diffusion in Latent Space: The U-Net Architecture

The latent diffusion model's core premise is the generation of a forward diffusion process, and its counterpart, the reverse process. This dichotomy of processes gradually transforms the original data into Gaussian noise and recovers the original data from the noise, respectively. The heart of this undertaking is the U-Net architecture, a robust convolutional network variant, integral to the successful image synthesis in the Stable Diffusion model [331]. Named for its characteristic "U" shape, the U-Net architecture shown in Figure 7.2, was originally developed for biomedical image segmentation. The structure of the U-Net is an innovation in itself: it is a symmetric architecture that comprises an encoder (the "contracting" path) and a decoder (the "expanding" path), connected by a bottleneck. Each stage in the contracting path is typically composed of two convolutional layers followed by a max pooling operation for downsampling, while each stage in the expanding path consists of an upsampling of the feature map followed by a convolution ("up-convolution") [331].

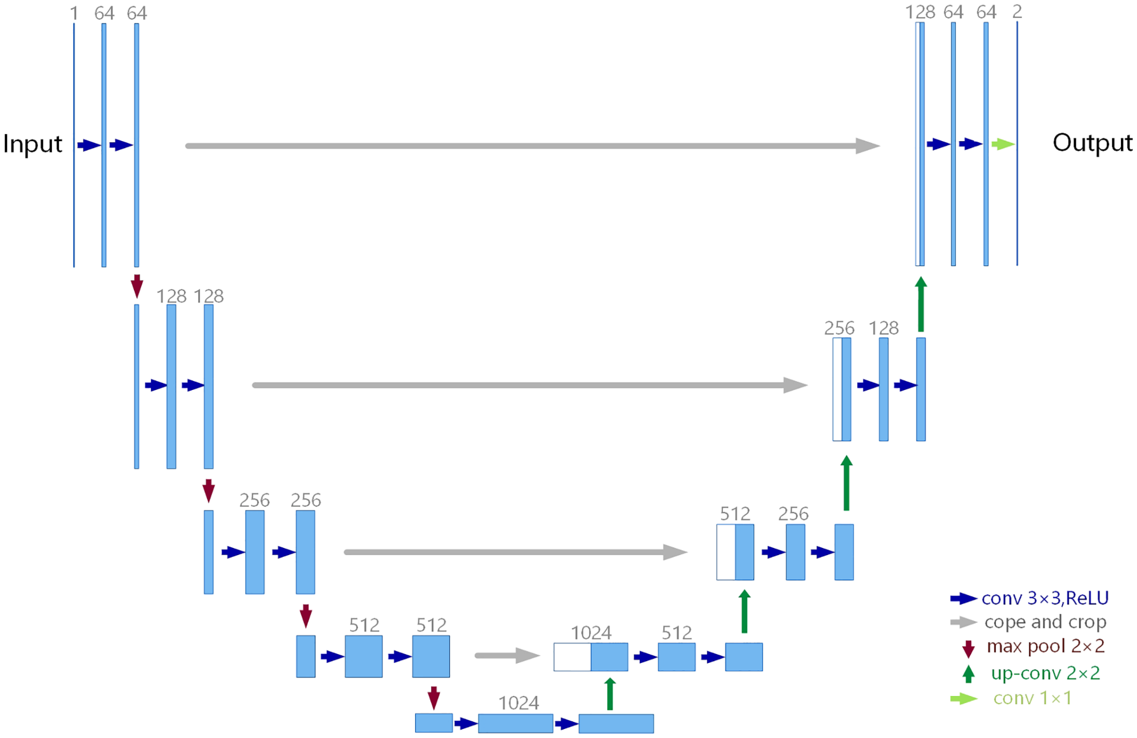


Figure 7.2: U-Net architecture. Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Reproduced from [331].

A critical feature of U-Net architecture is its series of long-range connections linking layers in the contracting path to the corresponding layers in the expanding path. These connections allow the network to leverage localized, detailed features along with the global, abstract ones, thereby synthesizing images with precise localization and rich context. The Stable Diffusion model adopts a time-conditional variant of the U-Net architecture. This variant enables the U-Net to learn and replicate the distribution of training images effectively. Coupled with a new optimisation target—the "reweighted bound"—that focuses on perceptually significant parts of the data, the U-Net in Stable Diffusion has a more balanced and refined optimisation target. This adjustment enhances the model's ability to create high-quality and diverse images that are perceptually close to the training data [311]. The role of the U-Net in the Stable Diffusion framework is, therefore, to enable effective diffusion in the latent space by establishing an intricate balance between data detail and abstraction. This balance facilitates the creation of images that are not only visually appealing but also retain the semantically significant features of the original data, thereby ensuring the quality of image synthesis.

Conditioning Mechanisms: Text-to-Image

Conditioning mechanisms are a crucial facet of generative models, providing control over the generation process to ensure that outputs adhere to specific requirements or guidelines. In Stable Diffusion, this conditioning process is achieved through a cross-attention mechanism [281] that interacts with a specific text encoder: the CLIP ViT-L/14 model. This encoder serves as a bridge between text prompts and their corresponding image outputs, empowering the Stable Diffusion model to synthesise high-fidelity images from textual descriptions [309]. The CLIP (Contrastive Language-Image Pretraining) model marries the strengths of vision transformers (ViT) and large language models, training on a vast array of image-text pairs to learn the intricate connections between words and their visual representations shown in Figure 7.3. [309], [332].

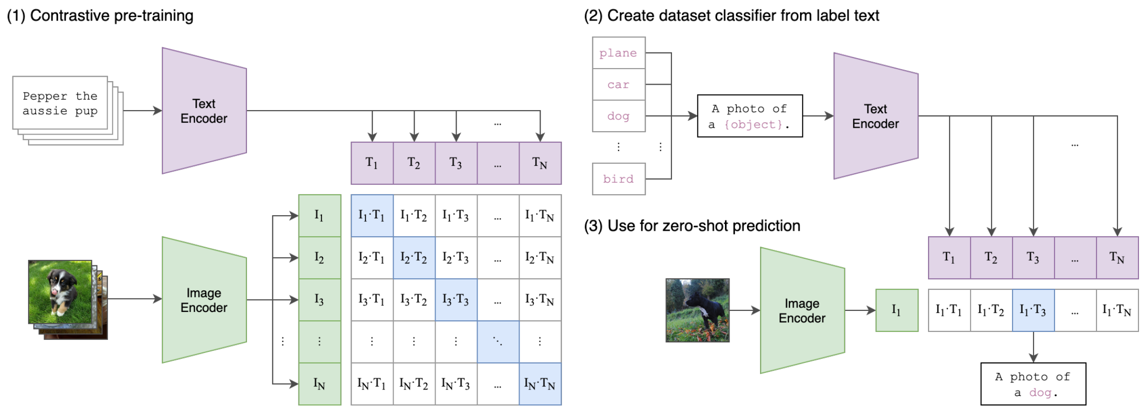

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Figure 7.3: CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time, the learned text encoder synthesises a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes. Reproduced from [309]

In the context of Stable Diffusion, the CLIP ViT-L/14 variant is utilised in this work. This particular model is a more substantial configuration of CLIP, offering higher performance at the expense of increased computational requirements. The ViT-L/14 text encoder within CLIP maps textual descriptions to an embedding space where semantically similar concepts, regardless of their format (textual or visual), cluster together. This conditioned encoding of textual prompts into a shared embedding space paves the way for the Stable Diffusion model to synthesise images that not only visually embody the text but also maintain semantic consistency [309]. The bi-directional interaction between the text embeddings and the latent code in the Stable Diffusion model (see Figure 7.4) is realised through the cross-attention mechanism, which in practice manifests as a set of blocks within the U-Net architecture. It allows the global context, derived from the text encoder, to interact with the local features of the latent code, thereby ensuring that the global structure of the data is considered during the image synthesis process [311]. The net result is the creation of images that adhere to the semantics of the original text prompts, making Stable Diffusion a potent tool for text-to-image synthesis tasks.

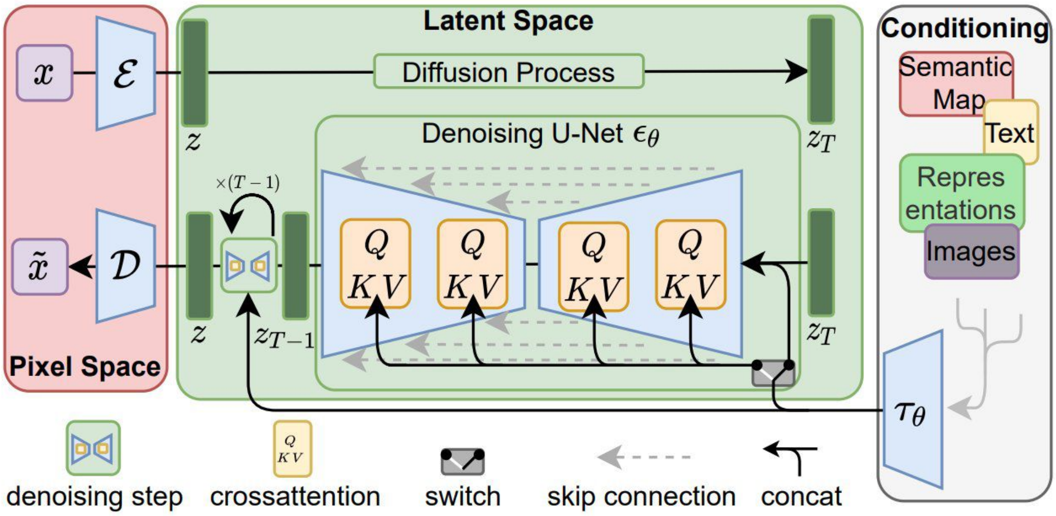


Figure 7.4: Stable Diffusion architecture. Reproduced from [309].

7.3 Visual Abstraction and Conceptual Distillation

The rapid emergence of high-fidelity open-source image synthesis technologies has ushered us into a remarkably novel landscape of visual analytics and understanding. By leveraging this innovation, we can strategically peel back the perceptual layers of a visual stimulus to focus our exploration on its underlying conceptual fabric. This perspective offers an innovative approach to dissect the memorable aspects of visual content, shifting the lens from merely its physical characteristics to its foundational concepts. Such an approach is further facilitated by the intricate interplay of the visual modality's richness and the prowess of advanced deep learning visual prediction techniques. This unique position is the stage for our exploration into the conceptual essence of video memorability.

Using Stable Diffusion—the open-source state-of-the-art image synthesis model—large swaths of synthetic images that transcend the domain of perceptual features are generated and used to delve into the subject of conceptual representation. This investigation is bifurcated into two parts: first, an initial foray into the efficacy of using synthetically generated images to discard perceptual features while preserving the underlying conceptual essence, involving the generation of four aesthetically distinct synthetic datasets, and the assessment of their comparative predictive capacity; second, the creation and benchmarking of a fully automated synthetic image predicated conceptual video memorability prediction framework.

7.3.1 Diffusing Surrogate Dreams of Video Scenes

The first strand of this investigation is unfurled with the generation of a diverse assortment of synthetic images. These images, while embodying distinct aesthetic styles, share a common purpose: to push the boundaries of perceptual traits while anchoring the underlying conceptual essence. The chosen tableau of aesthetic styles—realistic depiction, monochrome photography, surrealist art, and minimalist art—is not arbitrary, but chosen with deliberate academic intent. These styles offer not only a spectrum of aesthetic variance but also set the stage for deeper insights into how Stable Diffusion negotiates the intricate dynamics of preserving the conceptual kernel amidst a shift in aesthetics.

The first style—termed *Real* for brevity—anchors itself in the familiar territory of the everyday, a stylistic sibling to the Memento10k videos, and serving as a touchstone for the subsequent styles. By utilising the very style that mirrors our daily visual reality, we establish a comprehensive baseline for the ensuing comparisons. A realistic style helps us understand the capacity of synthesised images to preserve conceptual features while meaningfully altering the perceptual features, and equally attempting to maintain a high degree of visual fidelity to the original image.

The second style—termed *Monochrome* for brevity—is grounded in the world of black and white. The absence of color in this landscape places emphasis on elements like contrast, texture, and form. This greyscale style, brings into sharp relief the potential synthesised images to hold the underlying concept intact when colour, a vital instrument in perceptual communication, is eliminated.

The third style—termed *Minimal* for brevity—steps into the realm of simplicity and conceptual essence concentration. This style prides itself in distilling visual content to its bare essentials. The analytic value of minimalism is derived from the desire to evaluate the performance of synthetic images when extraneous details are pared away and the representation is simplified to its core concept.

The final style—termed *Surreal* for brevity—firmly sits in the expressive and often enigmatic domain of surrealist abstract art. This style signifies a substantial deviation from realistic representations, favouring emotional, symbolic, or conceptual expression over literal depiction. This setting, marked by its visual unpredictability, tests the limits of perceptual and conceptual distortion, allowing us to examine whether core conceptual elements can still be effectively perceptually communicated while radically deviating natural depiction.

Each of these styles holds a unique academic motivation, and provides a distinct setting from which to examine the ability of synthesised to preserve/encapsulate the conceptual essence of ground visual stimuli. Together, they form a diverse aesthetic panorama from which we can broaden our understanding. The generation of synthetic images was carried out within the purview of predicting video memorability, with the Memento10k dataset—comprised of 7,000 training videos, 1,500 validation videos, and 1,500 test videos—acting as the data landscape. This landscape was terraformed into four distinct synthetic datasets—based on the four aforementioned styles, and consisting of 20,000 images, two per video—collectively termed "Memento10k surrogate dream". A common base of the stable-diffusion-v1-5' checkpoint was used to generate each dataset, with the exception of the model used for the *Real* style, which was fine-tuned on 20 real-world photographs (see Figure 7.5), which encapsulate the heterogeneity and "in the wild" nature of Memento10k videos, and

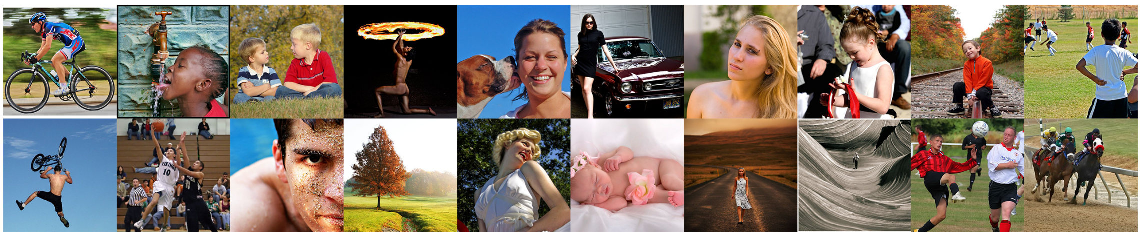


Figure 7.5: Images used to fine-tune the Stable Diffusion model and create the mem10kstyle token.

provides a stylistic grounding for the nature of realistic images desired. The input prompts to the Stable Diffusion models combine the Memento10k video's ground-truth first caption, action labels, custom prompt modifiers which specify the stylistic nature and emphasis of the models' generation, and in the exclusive case of the *Real* style, a"mem10kstyle" token (see Figure 7.7). This fusion aims to guide the generation process towards images that echo the conceptual substance of the original videos while jettisoning their perceptual specifics. The prompt modifiers play an integral role in generating high-quality, style-specific images that align with the desired output characteristics. Leveraging the power of language, these modifiers—which could be in the form of famous artists, renowned photographers, specific mediums (e.g., paint, photography, sketch), art styles, and compositional keywords (portrait, landscape, wide-angle, and more)—provide Stable Diffusion with precise instructions on the aesthetic attributes desired in the final output.

A caption categorisation algorithm acts as the backbone of this approach, ensuring that hand-crafted, style and category-specific prompt modifiers are appropriately assigned. By classifying the video captions into categories such as People, Animals, Landscapes, or Interiors, not only does this ensure the preservation of the inherent category characteristics in the ground-truth videos, but it aids in emphasising certain features or inducing one of the four—*Real*,*Monochrome*, *Minimalist*, and *Surreal*—image generation styles. For instance, using a prompt modifier like "Picasso" might result in images with strong Cubist influences, whereas "Ansel Adams" might produce high contrast monochrome landscapes. The caption classification algorithm is built upon a careful selection of indicative keywords for each category: People, Animals, Landscapes, and Interiors. These keywords, essentially specific nouns, serve as our classifiers, enabling us to assign a corresponding category label to a caption if it contains a word from these sets. The classification begins with the tokenization of the caption, transforming it into individual words which are then cross-checked against the dictionaries of each category. A match leads to the immediate labeling of the caption and halts the search. In case of no matches, the caption is classified

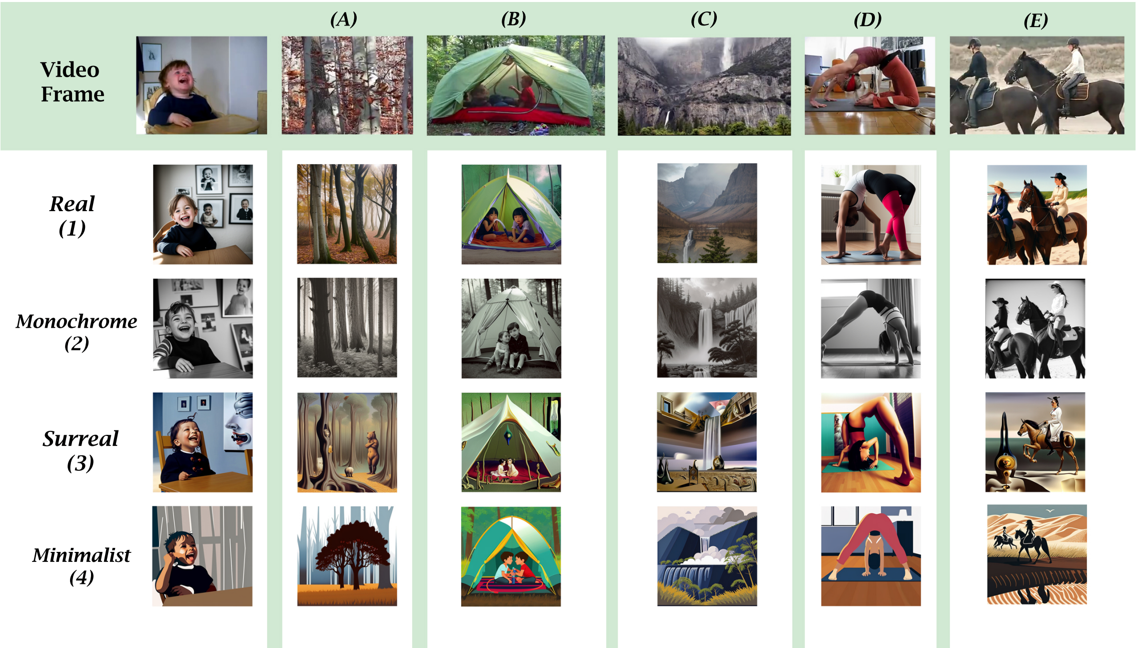


Figure 7.6: Example surrogate images generated for each of the styles.

as "Misc". This approach, as simple as it is, was chosen due to a lack of readily available labeled caption data for our specific categories, and perhaps more importantly, the fact that it met the requirements. The corresponding pseudo-code for this method is shown below:

Algorithm 3 Caption categorisation algorithm for prompt modification

```
$people \leftarrow \{'man', 'woman', 'child', ..., 'person'\}$
$animals \leftarrow \{'dog', 'cat', ..., 'elephant'\}$
$landscapes \leftarrow \{'mountain', 'ocean', ..., 'canyon'\}$
$interiors \leftarrow \{'livingroom', 'kitchen', ..., 'attic'\}$
function CATEGORIZE_CAPTION($caption$)
    $tokens \leftarrow word\_tokenize(caption.lower())$
    for all $word$ in $tokens$ do
        if $word$ in $people$ then return 'People'
        else if $word$ in $animals$ then return 'Animals'
        else if $word$ in $landscapes$ then return 'Landscapes'
        else if $word$ in $interiors$ then return 'Interiors'
        end if
    end forreturn 'Misc'
end function
```

Each of the synthesised datasets were generated with a different set of category specific prompt modifiers unique to their style. Below are the prompt modifiers used to generate images in the *Real* style:

- **People:** “by Alasdair McLellan, by Jovana Rikalo by Alessio Albi, by Andrea Kowch, by Guy Aroch, detailed, sharp focus, cinematic, unsplash featured photograph 8k, mem10k style”
- **Animals:** “by Frans Lanting, by Steve McCurry, by Tim Flach, macro, sharp focus, national geographic style, trending on Instagram, featured photograph 8k, mem10k style”
- **Landscapes:** “by Alvar Aalto, by Christophe Jacrot, by Wayne Thom, by David Muench, 35mm, stunning environment, sharp focus, landscape photograph, cinematic, featured photograph 8k, mem10k style”
- **Interiors:** “by Valeria Lazareva, by Julius Shulman, by Vincent Van Duysen, architectural digest, sharp focus, minimalist, vogue living, featured photograph 8k, mem10k style”
- **Misc:** “by Neil Leifer, by Frans Lanting, by Frank Lloyd Wright, by David Muench, 35mm, sharp focus, insanely detailed, trending on pixabay, cinematic, featured photograph 8k, mem10k style”

The *People*’s selection leans heavily into the human element, referencing photographers renowned for their distinctive capture of people. These include Alasdair McLellan, known for his authentic portraiture and Jovana Rikalo, a surreal and conceptual photographer, and Alessio Albi who specializes in emotive portraiture. Andrea Kowch, a painter recognized for her intense, narrative compositions, adds an artistic perspective. Guy Aroch’s modern, cinematic style further broadens the spectrum. Terms like ‘detailed’ and ‘sharp focus’ emphasise the intent for high-definition and clear imagery, with ‘cinematic’ suggesting a storytelling component. “Unsplash featured photograph 8k” and “mem10k style” hint at seeking top-rated and popular aesthetics in contemporary photography.

The *Animals*’ prompt modifiers reference esteemed wildlife photographers Frans Lanting, Steve McCurry, and Tim Flach, signalling a desire for high-quality, professional animal imagery. ‘Macro’ alludes to the often-used technique in wildlife photography that captures intricate details of animals, while ‘sharp focus’ underscores the importance of clarity. By including ‘national geographic style’, we denote a wish for impactful, action oriented wildlife photography. The prompts ‘trending on Instagram’ and ‘featured photograph 8k’ indicate the need for popular, high-resolution images, and distinctive animal photography.

The *Landscape* category incorporates a blend of urban architects, urban photographers, and landscape photographers. Alvar Aalto, a renowned architect, and Christophe Jacrot, a photographer with a contemporary, contrast focused style, adds a dimension of precise architectural detail and compositional structure typical for cityscapes. The inclusion of photographers like Wayne Thom, and David Muench shows a focus on capturing the natural world’s awe-inspiring vistas. ‘35mm’ invokes a classic, wide field of view often associated with landscape photography. ‘Stunning environment’, ‘sharp focus’, ‘landscape photograph’, and ‘cinematic’ all point to a desire for immersive, breathtaking sceneries in high detail. The prompt ‘featured photograph 8k’ aims for popular, high-resolution and unique landscape photography.

The *Interiors* array draws from the works of interior architectural photographers Valeria Lazareva and Julius Shulman, creating a focus on professional, aesthetically pleasing interior shots. Vincent Van Duysen, a renowned interior designer, adds an additional design perspective that focuses on detail. ‘Architectural Digest’ and ‘Vogue Living’ evoke high-end, modern interior designs. ‘Sharp focus’ and ‘minimalist’ imply a preference for crisp, uncluttered images, while ‘featured photograph 8k’ for aims popular, high-resolution, and distinctive interior photography.

Finally, the *Misc* prompt modifiers is a versatile and category agnostic set to cover everything not covered by the previous categories and references a wide array of photographers with different specialities. This includes Neil Leifer capturing intense sports shots, Frans Lanting capturing wildlife shots, Frank Lloyd Wright capturing modern architecture, and David Muench capturing landscape photography. ‘35mm’ and ‘sharp focus’ emphasise the image’s overall clarity and depth. ‘Insanely detailed’ and ‘cinematic’ imply a desire for rich, narrative images. ‘Trending on pixabay’ indicates popular appeal, while ‘featured photograph 8k’ aim for high-resolution, popular, and unique photography regardless of the category.

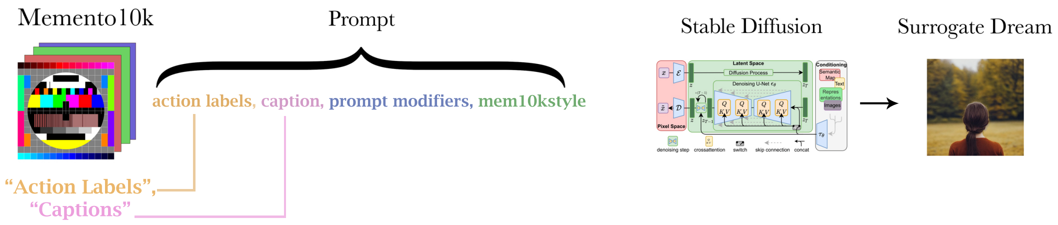


Figure 7.7: Surrogate Dream Pipeline to synthesise images (the “mem10kstyle” token is only included in the generation of the *Real* dataset images).

7.3.2 Predicting Video Memorability

Evaluating the efficacy of the synthesised datasets[1] in downstream tasks is critical in gauging their ability to capture and preserve the conceptual essence inherent within the original video frames. To this end, five ImageNet pre-trained DenseNet121 [196] neural networks—one for each synthesised dataset, and one for extracted ground-truth video frames—were trained to predict video memorability. Each of these models was put through a standardised training regime, fine-tuned for 50 epochs with a maximum learning rate of 1e-3 and a weight decay of 1e-2. The intent behind this methodology was to ascertain the predictive capacity of each synthesised dataset relative to a control model. This control model, referred to as the *Mem10k* model, was trained on unaltered, ground-truth video frames extracted directly from the original Memento10k videos. At test time, a video’s memorability score is calculated by averaging predictions over the first, middle, and last frame.

The testing strategy can be split into one of two categories, namely *Genesis* and *Surrogate Dream*. Approaches trained on vanilla Memento10k data—the control—are considered to be *Genesis*, and serve to establish a baseline memorability prediction performance, and a baseline for how well the synthesised images preserve the videos’ underlying memorability. Approaches trained on the synthetic image datasets are considered to be *Surrogate Dream*, and with the exception of memo-

[1] available at `https://figshare.com/projects/Memento10k_Conceptual_Dream/177663`

rability scores, are trained exclusively on surrogate visual data. By comparing the predictive performance of these two categories of approach, the extent to which the synthesised datasets maintain conceptual features pertinent to memorability through a transformation in perceptual features. Performance was quantified using Spearman rank correlation, a non-parametric measure that assesses how well the relationship between two variables can be described using a monotonic function.

Table 7.1: Results on the test sets for each of our approaches.The syntax for each approach communicates the datasets used for training and testing in the following manner: *trainedOn_ModelArchitecture_testedOn*.

Approach	**Spearman**
Genesis	
Mem10k_DenseNet121_Dream_Real	0.583
Mem10k_DenseNet121_Dream_Monochrome	0.501
Mem10k_DenseNet121_Dream_Minimal	0.407
Mem10k_DenseNet121_Dream_Surreal	0.438
Mem10k_DenseNet121_Mem10k	0.645
Surrogate Dream	
Dream_Real_DenseNet121_Mem10k	0.625
Dream_Monochrome_DenseNet121_Mem10k	0.567
Dream_Minimal_DenseNet121_Mem10k	0.431
Dream_Surreal_DenseNet121_Mem10k	0.489
Dream_Real_DenseNet121_Dream_Real	**0.664**
Dream_Monochrome_DenseNet121_Dream_Monochrome	0.601
Dream_Minimal_DenseNet121_Dream_Minimal	0.458
Dream_Surreal_DenseNet121_Dream_Surreal	0.512
State-f-the-art	
SemanticMemNet [192]	0.663

The results, shown in Table 7.1, elucidate insights into the nature of visual memorability and the efficacy of synthetic datasets in memorability prediction. The Genesis approach, which involves training a DenseNet121 model on ground-truth Memento10k video frames, yielded a robust Spearman correlation of 0.645 when tested on the same Mem10k test set.

Fascinating observations arise when the Genesis model is tested on synthetic datasets (*Real*, *Monochrome*,*Minimal*, and *Surreal*). Despite the notable perceptual differences between the ground-truth and synthetic images, the model was still able to predict memorability with reasonable accuracy, with Spearman correlations ranging from 0.407 to 0.583. This suggests that memorability is not strongly tied to the specific perceptual features of the visual stimuli. A clear divergence in model performance was observed when it was tested on the *Minimal* and *Surreal* datasets, where Spearman correlations dropped to 0.407 and 0.438 respectively, indicating a significant divergence in the perceptual features used by the model to predict memorability. Interestingly, despite the *Monochrome* dataset's significant stylistic transformation, it still retained enough visual cues that the model linked to memorability to attain a respectable correlation of 0.501.

Shifting attention to the Surrogate Dream approaches, the "Dream_Real" model, trained on the *Real* synthetic dataset, surprisingly outperformed the control Genesis model when tested on its own test set, achieving the highest Spearman correlation of 0.664. This noteworthy performance suggests that the dataset does not simply retain a high degree of perceptual alignment with the original data, but instead displays an improved conveyance of the underlying memorability, which accordingly suggests that memorability is not merely a perceptual attribute. Additionally, the "Dream_Monochrome" model, trained and tested on its test set, showed a lower but still competitive correlation of 0.601, corroborating the fact that the synthetic datasets, despite their perceptual mutation, can retain sufficient conceptual information for accurate memorability prediction. The lower correlation scores of the "Dream_Minimal" and "Dream_Surreal" models, when compared to their counterparts, can be interpreted as these synthetic datasets being more conceptually divergent from the original video frames due to radical nature of their stylistic transformations. Yet, even these models managed to perform decently, demonstrating that even with significant alteration in perceptual features, they capture enough of the videos' conceptual essence to reasonably predict memorability. Taken together, these results strongly suggest that memorability is less about specific perceptual attributes and more about the underlying conceptual essence of the visual stimuli. Even synthetic surrogate images, bearing no perceptual resemblance to the original video, can effectively predict memorability. This has profound implications for understanding visual memorability and opens exciting avenues for future research in this field.

The distributions of memorability score predictions for the best performing Genesis and Surrogate dream approaches, shown in Figure 7.8, indicates a strong overlap between predictions made with the Genesis control, the Dream_Real model tested on the control test set, and the Dream_Real model tested on its own test set. This combined with the evaluation scores, provides the first-of-its-kind strong evidence that visual data is merely a means to conceptual understanding, and that it is the concepts themselves—which are conveyed/represented through the visual data–that hold the content's intrinsic memorability. Additionally, graph B in Figure 7.8 tentatively suggests that surrogate dream images can be more memorable than ground-truth video frames by virtue of the left skew in predicted scores from the Genesis model tested on the surrogate *Real* test set.

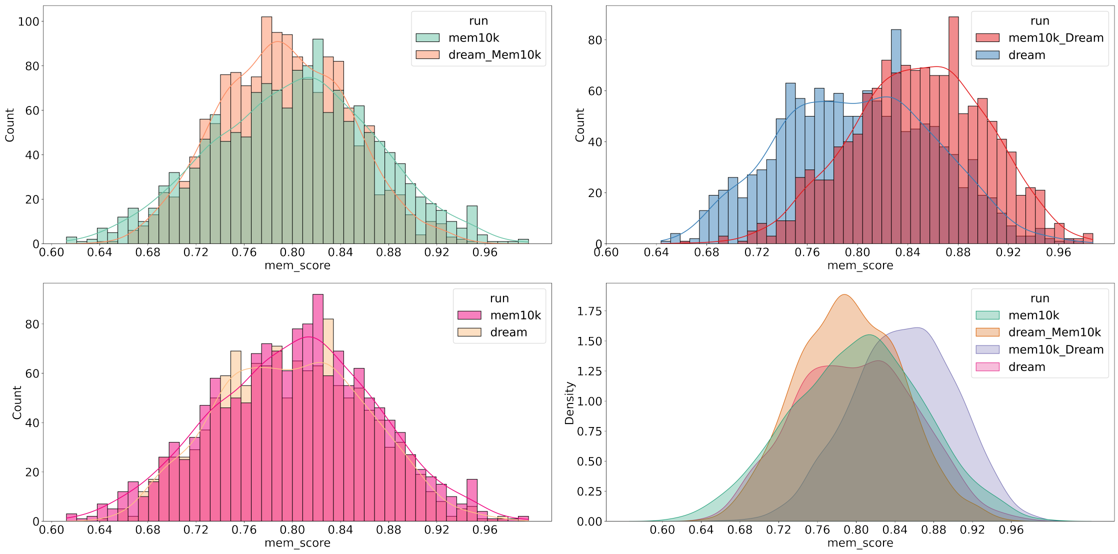


Figure 7.8: Distribution of memorability predictions.

7.4 ConceptualDream: A Video Memorability Prediction Framework

The second part of this chapter's investigation details the creation and benchmarking of a fully automated framework for prediction of conceptual video memorability for synthetic images. This framework, named ConceptualDream, represents an approach to prediction of short-term recognition video memorability, merging cutting-edge techniques in synthetic image generation, and state-of-the-art memorability prediction. It is an entirely automated framework, requiring only a video as input, and yielding a memorability prediction based on the conceptual essence of the video. ConceptualDream can be broken down into its two underlying core processes: image synthesis, and short-term recognition memorability prediction. The first process generates a set of four synthetic images based on the conceptual essence of the input video, and the second process uses these synthetic images to predict the video's short-term recognition memorability. In the initial stage of the first process, an input video is parsed at a rate of one key frame per second. For videos of

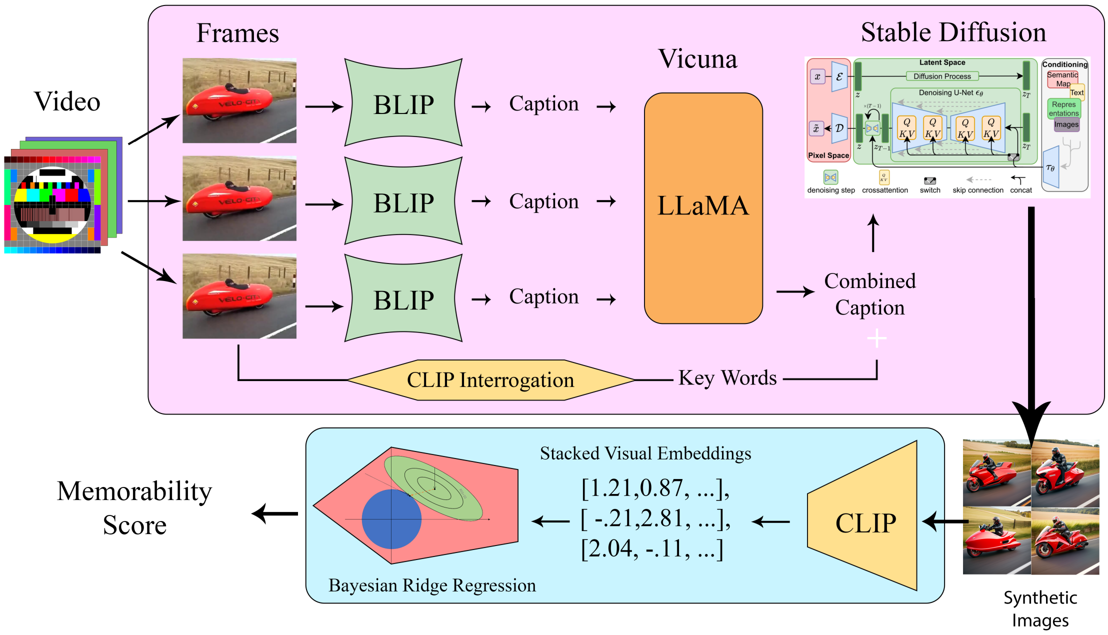


Figure 7.9: ConceptualDream framework. The lilac block represents the image synthesis process, and the blue block represents the memorability prediction process.

a three-second duration, such as those in the Memento10k dataset, this results in three key frames—one from the start, middle, and end of the video. Each frame is then processed through the BLIP model [333], a robust image captioning neural network. The BLIP model employs a unique architecture known as the Multimodal mixture of Encoder-Decoder (MED), where a visual transformer (ViT) is used as the image encoder. This structure allows the model to effectively understand the content of each frame and generate a high-quality summary. As a result, this procedure yields three separate captions that accurately summarise the content of the start, middle, and end of the video. To synthesise a single coherent caption that encapsulates the full narrative of the video, the individual captions are then passed to an open-source large language model, namely Vicuna 13B [334], which is a fine-tuned LLaMA-13B model [335]. The LLaMA model is an auto-regressive language model based on the transformer architecture [281], but leverages various improvements (pre-normalisation, the SwiGLU activation function, and rotary positional embeddings) that were subsequently proposed and used in other language models. This advanced language model is instructed to integrate the information from the three captions into a single, unified and comprehensive caption that effectively captures the essence of the video. To further enrich the resultant caption, and ultimately transform it into a Stable-Diffusion prompt, the top-k keywords from a pre-existing bank of keywords are appended to the caption. The selection of keywords is guided by their CLIP scores, which measure the semantic similarity between the keywords and the extracted images. The bank of keywords is a robust compilation sourced from multiple resources, including the names of 5,265 artists, and 100,970 phrases drawn from prompt engineering exercises. The keywords are greedily sampled until the prompt reaches CLIP's token length limit of 77. This process is termed *CLIP Interrogation*. The final prompt, is sent to a Stable Diffusion model fine-tuned on high quality real-world photographs. The model generates four distinct synthetic images, each visually representing perceptual variations of the conceptual essence encapsulated in the prompt. This part of the framework was applied to all 10,000

Memento10k videos to produce a synthetic dataset called MementoDream, which consists of 40,000 synthetic images, each ground-truth video corresponding to four synthetic images. The second part of the framework focuses on prediction of short-term recognition memorability. Here, visual embeddings from the synthetic images, processed via the CLIP model, are extracted and then stacked together to form a structured representation. This structured input serves as a basis for a Bayesian Ridge Regressor (BRR), which delivers a memorability score. The BRR model, using the sklearn library's implementation and the Grid Search algorithm [195], is trained on stacked CLIP visual embeddings from the MementoDream training and validation sets. To demonstrate the efficacy of the ConceptualDream framework, it was compared against the Memento10k benchmark Spearman scores [192]. As illustrated in Table 7.2, ConceptualDream surpasses the performance of the previous state-of-the-art model, SemanticMemNet, by achieving a remarkable Spearman score of 0.724, marking a new high in the realm of video memorability prediction. Furthermore, the magnitude of this achievement becomes even more significant when considering the innovative methodology ConceptualDream adopts.

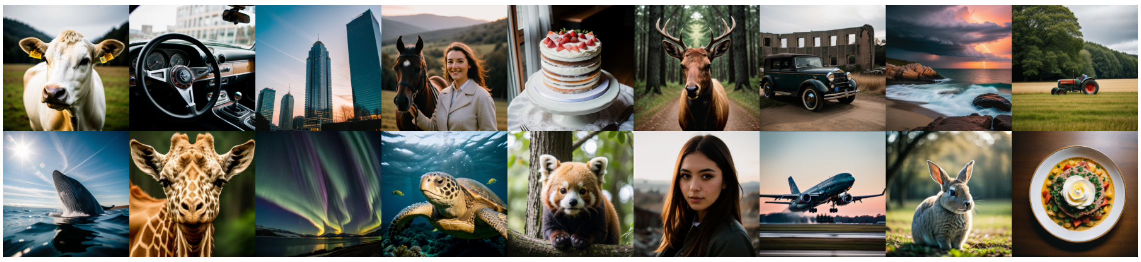


Figure 7.10: Example images generated with the ConceptualDream framework.

7.5 Conclusion

This chapter has presented a rigorous investigation of Hypothesis 5 from the thesis, which posits that it is the conceptual essence of a video, rather than its perceptual features, that holds the key to its memorability. This investigation employed a blend of advanced machine-learning methodologies, most notably the leading edge synthetic image generation technique of latent diffusion. In the initial phase of the

Table 7.2: Comparison of state-of-the-art on Memento10k test set.

	Memorability
Approach	**Spearman**
Human Consistency	0.730
MemNet Baseline [198]	0.485
Cohendet et al. (Semantic) [222]	0.552
Cohendet et al. (ResNet3D) [222]	0.574
Feature Extraction + Regression (as in [223])	0.615
MeMAD [212]	0.658
SemanticMemNet [192]	0.663
ConceptualDream	**0.724**

investigation, a novel experimental design was implemented to probe the relationship between visual perceptual features and conceptual representation in predicting video memorability. This was achieved by generating a diverse array of synthetic surrogate images that retained the conceptual essence of ground-truth videos, but drastically transformed the perceptual features. The experiment was designed to test models on four distinctive styles of synthetic images. The results painted a captivating narrative: despite perceptual disparities between ground-truth and synthetic variants, accurate memorability predictions remained achievable. This revelation indicates a diminished reliance on perceptual characteristics in the determination of memorability, spotlighting the importance of conceptual characteristics instead. The ConceptualDream framework, introduced in the subsequent phase, marks an evolution in the exploration of video memorability prediction. By harmoniously inter-twinning the essence-capturing capabilities of advanced synthetic image generation and the acuity of state-of-the-art memorability prediction techniques, ConceptualDream was able to achieve a remarkable Spearman score of 0.724, not only surpassing the previous state-of-the-art model by a significant margin, but approaching human consistency.

The findings in this chapter provide compelling evidence in support of Hypothesis 5. It is shown that even synthetic images bearing no perceptual resemblance to an original video can effectively predict its memorability. This firmly implies that memorability is less about specific perceptual attributes and more about the underlying conceptual essence conveyed by the visual stimuli. This chapter offers a novel understanding of the nature of visual memorability, while simultaneously opening new avenues for future research. The exciting potential of this work encourages subsequent inquiries to delve deeper into the intricate interplay of memorability and the conceptual essence of visual stimuli. These findings have implications not only within the academic realm of computational memorability research and cognitive psychology, but also in broader sectors like visual arts, advertising, and content creation, underscoring the broad relevance and significance of this work.

Chapter 8

Conclusions

This thesis set out to investigate the intricate complexities of a fundamental aspect of human cognition: memorability. The primary aim was to elucidate its true character, decipher its mechanisms, and examine its relationship with sensory experiences, its temporal aspects, and its interplay with recognition and recall. Intrinsic to human cognition, memorability broadly refers to the propensity of a stimulus to be remembered or forgotten. Yet, it possesses far-reaching implications that extend beyond this simple definition, its influence might not merely affect surface-level cognition; it could be closely aligned with preferences rooted in our evolutionary lineage, thus influencing what we instinctively prioritise. It's not merely about what is remembered but also how, where, and when. Understanding memorability could unveil layers of our cognitive processes, offering a scaffold to refine didactic strategies, shape neuropsychological interventions, and even discern the patterns of selective attention in various scenarios. This research advanced the hypothesis that the intrinsic memorability of a given stimulus is a complex construct, shaped by the dynamic and symbiotic interplay of various sensory modalities. It built on the fact that due to inherent cortical bias, visual stimuli act as the primary modality through which it is communicated. Furthermore, this thesis proposed that remembrance is a polymorphic process. It involves an intricate balance between recognition, detailed recollection, or a confluence of both. This nuanced process is proposed to be governed by biological storage limitations, implying that specific moments of representational compression exist, and these correspond to particular instances of recollection. Arguably the most significant part of this hypothesis was the assertion that memorability transcends the perceptual attributes of a stimulus. The conceptual understanding gleaned from a perception was suggested to be the key to memorability. This perspective signals a radical departure from traditional views of memorability, suggesting that it is not just a visual feature, but fundamentally a conceptual one. In alignment with these research objectives, this dissertation conducted a systematic exploration of the multifaceted phenomenon of memorability, delving into its sensory, temporal, and cognitive aspects, and ultimately attempting to decode the essence of what makes experiences memorable. The intention was to illuminate the intrinsic nature of memorability, providing fresh insights into why and how memories are formed, stored, and retrieved.

8.1 Summary of Findings

Chapter 6 delved into the intricacies of multimodal memorability, and in doing so, provided robust empirical support for the first hypothesis (H1), which emphasises the inherently multimodal nature of memorability. The audio modality was found to be a potent contextual element, assisting recognition when conveying high-level features. This finding substantiates the assertion that recognition memorability correlates strongly with conceptual properties of the content, extending beyond just the visual domain.

However, a complete understanding of the role of audio in short-term video recognition memorability remains elusive. The complex interaction between a video's auditory and visual content could have a more profound, and yet to be discovered, influence on its overall memorability. A comprehensive understanding of this interplay necessitates further rigorous investigation, potentially through the implementation of independent memorability metrics for each sensory modality—audio, visual, and textual.

The second hypothesis (H2) examined the part played by visual sensory data in multimodal memorability prediction. While the findings reaffirmed the substantial role of the visual domain, they simultaneously highlighted the powerful impact of textual features. The considerable importance of the textual modality in memorability underlines the proposition that semantics can encapsulate a significant proportion of a video's memorability, thereby reducing reliance on perceptual information. Consequently, these findings hint at the potentially amodal nature of memorability, revealing it as less of a direct product of a single modality, but a fine interplay of sensory inputs, and possibly even an abstraction beyond the direct grasp of senses.

Chapter 7 explored the use of electroencephalography (EEG) signals in concert with advanced deep learning techniques to predict subsequent recognition of previously seen videos. While the investigation did not identify a significant interaction between subject-dependent (SD) and subject-independent (SI) training approaches, it nevertheless illuminated the potential relevance of subject-specific EEG data for enhancing prediction accuracy. A cornerstone of this chapter was the introduction of the novel 'moment of memorability' hypothesis (H4). This transformative perspective reframes the consideration of video content and its memorability, proposing that memorability is not a static property of the video, but rather, dynamically linked to the point at which observers typically form a compressed mental representation of the conceptual content conveyed, and assign it an information-utility value—memorability score. Chapter 7, therefore, paved the way for further detailed investigation into the facets of memory encoding and retrieval, focusing not only on what content is remembered, but also explored the critical timing of content comprehension, and the underlying reasons why specific content garners higher memorability. This fresh perspective, built on the integration of neurophysiological data and machine learning techniques, opens new avenues of exploration in the nuanced domain of memorability.

Chapter 8 validated the hypothesis that there is a measurable relationship between recognition and recall memorability (H3). Both visual and textual measures provided robust empirical support for this hypothesis. There were noticeable disparities in the semantic alignment and precision of recall between videos categorised as high and low in recognition memorability. A strong correlation was found between the normalised measure of recall accuracy and video memorability, highlighting the impact of individual differences on recall performance. In the textual domain, the introduction of Caption Specificity Score (CSS) as a novel measure of recall precision further validated H3. There was a significant correlation between the normalised CSS and high video recognition memorability. These findings emphasised the importance of recall precision in determining the recognition memorability of video stimuli. The absence of forgotten or misremembered videos in the high recognition memorability category bolsters the body of evidence supporting a correlation between recognition and recall memorability.

Chapter 9 conducted a detailed investigation of Hypothesis 5, which posited that memorability of a stimulus is essentially a reflection of its underlying conceptual representation. According to this hypothesis, perception serves as a conduit to conceptual understanding, and memorability acts as a gauge of information utility. This supposition was scrutinised using cutting-edge machine learning techniques, including the novel synthetic image generation technique known as latent diffusion. The initial stage of the investigation involved the creation of synthetic surrogate images. These images were a transformative departure from the original videos in terms of perceptual features, but preserved the conceptual essence of the source. The potency of these surrogates lay in their utility as test entities for analysing the interplay between visual perceptual features and conceptual representation in predicting video memorability. Remarkably, despite the pronounced perceptual variations between the original and synthetic variants, memorability prediction accuracy remained high. The chapter further unfolded to introduce the ConceptualDream framework, a leap forward in the field of video memorability prediction. By integrating the synthetic image generation's essence-capturing capabilities with state-of-the-art memorability prediction techniques, ConceptualDream achieved a laudable Spearman score of 0.724—surpassing previous state-of-the-art, and approaching human consistency. In essence, the chapter's findings provided robust support for Hypothesis 5 by demonstrating that the recognition memorability of a stimulus is more strongly tied to its conceptual representation than its perceptual attributes.

8.2 Implications of Findings

Plunging into the neural undercurrents of cognition, this thesis has surfaced with an enlightened view of memorability, altering its standing within the cognitive architecture. Memorability, often treated as a mere result of a specific stimulus's distinctive or peculiar aspects, emerges, in reality, as an intricate cognitive phenomenon that is as complex as it is fascinating. This thesis heralds a shift from a narrow, limiting perspective of memorability to a broader, more comprehensive conceptualisation.

The notion of memorability as a proxy measure of human importance or information utility unfurls in this exploration. The essence of memorability, it appears, is tied less to the perceptual distinctiveness of the stimuli and more to their underlying conceptual representation, which resonates with the hypothesis (H5) that the memorability of a stimulus can be reduced to its underlying conceptual representation. Herein, the crucial point of departure is the emphasis on the significance of the 'conceptual essence' of stimuli and its impact on human cognition and its operation. It is the content that carries meaning, relevance, and utility to the viewer that indeed becomes memorable, positing memorability as a measure of the informational value a stimulus holds for an individual.

This approach invites us to perceive memorability not as a static property inherent in the stimulus, but rather as a dynamic process intricately tied to the timing and comprehension of the stimulus's content. This temporal aspect of memorability—captured in the notion of a 'moment of memorability'—augments the traditional view, encouraging us to delve deeper into not only 'what' is remembered but also 'when' and 'why' certain content is remembered. This temporal dynamism extends the dimensions of memorability, adding a depth of complexity to our understanding of it.

Within the complex milieu of this investigation, the intertwined roles of auditory, visual, and textual modalities emerge as critical in forming a robust prediction of memorability in multimodal contexts such as video. Their collective influence, either amplifying or tempering the memorability of a given stimulus. This intricate relationship extends the understanding of memorability beyond the confines of the visual realm, revealing its deeply multimodal nature. Memorability, thus, emerges not as a one-dimensional perceptual response, but as a rich cognitive process mediated by the dynamic interplay of multiple sensory modalities.

Additionally, the intertwining of recall and recognition memorability in this exploration enhances the understanding of memorability. The interconnected nature of recognition and recall, far from being isolated phenomena, form an integral cornerstone of our understanding of remembrance. This sheds light on a key misconception (see chapter 6), elevating memorability from being a solitary phenomenon, tethered merely to recognition, to a broader cognitive tapestry, intricately woven with the threads of both recognition and recall. This nuanced interpretation, no longer confined within the rigid walls of a binary recognition task, but expanded to incorporate the more profound facets of recollection, offers a more comprehensive representation of human information utility.

In conclusion, this thesis illuminates the nuanced landscape of memorability, revealing it as a measure of information's human importance or utility. This innovative interpretation invites a broader, deeper exploration of memorability, transcending its traditionally narrow confines and resonating across various domains of cognition and expressions of remembering. It opens up new pathways for future research, elucidating the rich mosaic of cognitive processes that constitute the complex experiences of remembering. Through this lens, we are invited to look beyond the traditional boundaries of memorability and embrace a more expansive, more nuanced understanding of this fascinating cognitive phenomenon.

8.3 Future Work

Influence of Modalities

The complex interplay of modalities in determining video memorability represents a promising yet largely unexplored domain. While our current studies provide foundational insights into the synergistic effects of audio and visuals, a holistic understanding demands more granular investigations. Different combinations of modalities – such as visual-audio, visual-textual, and audio-textual – could potentially produce effects more pronounced than the mere sum of their individual contributions.

To navigate this intricate landscape, we propose a series of comprehensive, large-scale experiments. The objective is to evaluate memorability metrics across various modality combinations systematically. This can be achieved by implementing successive iterations of the 'Memorability Game' across diverse participant cohorts. In one proposed iteration, participants would experience a video in its entirety – encompassing visuals, audio, and on-screen captions. Successive iterations would then refine the modalities: a pairing of visuals with audio, visuals combined with captions, and an amalgamation of audio with captions. To distill the essence of each modality's contribution, subsequent versions would present each in isolation – focusing solely on visuals, audio, or captions.

This multifaceted approach promises a dual advantage. Firstly, it enables the dissection of the unique memorability footprint imprinted by each modality. Secondly, it uncovers any emergent properties arising from their combined presentation. Such revelations could significantly augment our comprehension of multimedia memorability. As we look ahead, these findings have the potential to guide content creators, allowing them to craft more resonant and memorable multimedia experiences.

EEG Video Memorability

The research outline in chapter 5 underscores the potential significance of theta band oscillations over the right temporal lobe in memory encoding, and introduces the intuitive notion of a ‘moment of memorability’, where a moment of representational compression (“biological understanding”) of the video content corresponds to the moment in the video at which the content is recognised. These findings, while insightful, are limited, and signal the need for more extensive inquiries in the field. A large-scale EEG study, with a proposed participant sample of approximately four times the typical study minimum (typical is 30, suggested is 120) is a logical next step. Such a study would permit a more comprehensive investigation of recognition memorability scores, enabling us to move beyond the limiting binary “remembered”/”forgotten” categorisations. With this refined approach, subtle variations in recognition memorability and its EEG correlations can be more accurately discerned. Additionally, the ‘moment of memorability’ warrants further validation. By pinpointing exact instances within video content where comprehension peaks, we can attempt to elucidate the specific moments when memory encoding is most potent. Such understanding would offer a more refined perspective on the temporal dynamics of memory formation in relation to video content. Moreover, the subtle distinction in performance between subject-dependent and subject-independent models raises questions about the role of individualised EEG patterns in memory related predictions. Expanding the study to include a larger sample could more definitively determine whether models trained on specific individual EEG data or generalised EEG data are more effective in forecasting memorability.

Recall Video Memorability

The relationship this research (see chapter 6) suggests between recognition and recall memorability underscores an novel area of exploration in memory studies with much to offer. As we strive to gain a more comprehensive understanding of memory and its related areas, there’s an evident need to refine our approaches, especially those concerning the measurement of recall in the context of video stimuli.

Leveraging drawings as a metric for recall offers a unique lens through which to view and analyse memory processes. However, the subjective nature of such a measure demands a meticulous, standardised criterion for its analysis. Future endeavors should focus on establishing a rigorous framework, both qualitative and quantitative, which can encompass the multifarious nuances of recall. This method, though rich in potential, also calls for a universally accepted set of guidelines, ensuring consistent interpretations despite the inherent subjectivity.

The marked correlation between the normalised measures of recall accuracy and video recognition memorability underscores that recall, much like recognition memorability, exhibits a degree of consistency across the population. This consistency hints at a symbiotic relationship between recognition and recall memorability, with one potentially informing or influencing the other. The nuances revealed by the VVIQ scores offer an intriguing dimension to this discourse. The absence of a broad correlation between VVIQ scores and recall accuracy, coupled with enhanced recall among those with the highest VVIQ scores, suggests that the drawing-based measurement might inherently favour those with particularly vivid mental imagery. Consequently, while individual cognitive differences can certainly shape recall outcomes, the chosen measure—drawing in this instance—may play a pivotal role in the manifestation of these outcomes. This underlines the imperative to further refine and broaden recall measures in future investigations. Moreover, the correlation between the Caption Specificity Score (CSS) and recognition memorability advocates for its potential as a reliable metric. Yet, its full adoption mandates an in-depth evaluation, potentially leading to its enhancement to ensure that it remains robust across different contexts and applications.

Conceptual Dream

The empirical findings presented in Chapter 7 underscore the primacy of a video's conceptual essence in determining its recognition memorability, rendering perceptual features secondary in this regard. This observation naturally begets several forward-looking research trajectories.

To begin with, the term 'conceptual essence' in video content, as employed within the present research context, necessitates rigorous academic unwrapping. While the significance of conceptual attributes in driving memorability is strongly suggested, the next logical step mandates a detailed dissection of these attributes. A methodical exploration into which conceptual elements—or a synergy thereof—hold the most sway in the memorability paradigm remains as a potentially useful undertaking.

Latent diffusion has been used effectively in our research for generating synthetic images. However, with machine learning techniques constantly evolving, it's important for future research to consider and assess other emerging methods. Two key questions arise: Do other techniques provide different insights into the conceptual core of videos? And do some methods emphasize certain aspects of video memorability more effectively?

Additionally, a next logical step in research would be to expand our findings to longer video formats. Videos are more than just a series of frames; they tell stories over time. Therefore, understanding memorability in the context of full-length videos, which can span minutes or even hours, is crucial. This would help confirm if our conclusions about the conceptual core of videos remain consistent across different video lengths and types.

From a practical perspective, the concept of memorability has potential applications in areas like visual arts, advertising, and digital content creation. It would be beneficial to investigate how these industries can utilise our findings. Specifically, it's worth exploring if content designed with our memorability insights in mind is more impactful or resonant with its intended audience.

Appendix A

Data Quality Concerns with the TRECVid Video Memorability Dataset

The TRECVid Video Memorability Dataset, although a valuable resource, has prompted several queries related to its data quality. Here we elaborate on the primary concerns:

- **Annotation Variability:** There are significant disparities in the number of annotations across different subsets. For instance, the VideoMem and Memento10k collections have an average of 40 and 90 annotations for each video in terms of short-term memorability respectively. In contrast, the TRECVid 2019 Video-to-Text dataset settles for an average of 22 annotations per video. Such inconsistencies can lead to variations in memorability score computations, thereby questioning the comparative robustness and reliability of the datasets.

- **Annotator Authenticity:** A notable anomaly in the dataset is the emergence of memorability scores that are significantly higher than what one would expect. This aberration raises concerns about the authenticity of annotators. It's plausible that some participants might have engaged with the memorability game on multiple occasions, leading to potential biases in the scores due to their familiarity with the video content.

- **Video Quality:** Another influential variable in this context is the inherent quality of videos present in the dataset. Factors such as video resolution, content clarity, and even the thematic uniqueness could inadvertently influence memorability scores. As such, it becomes imperative to account for these attributes and discern their respective contributions to the memorability quotient.

- **Limited Long-term Annotations:** When it comes to long-term memorability assessments, there's a noticeable paucity in annotations, especially in the TRECVid 2019 Video-to-Text dataset. This dataset, in particular, demonstrates a reduced average of annotations for long-term memorability relative to its short-term counterpart. Such imbalances can potentially skew the insights derived, thereby underscoring the need for a more balanced annotation approach to ensure comprehensive long-term memorability evaluations.

To rectify these concerns and shed light on the inherent ambiguities, a meticulous investigation is currently underway. The outcomes of this investigation are eagerly awaited, as they promise to enhance our confidence in the dataset and refine the way we interpret and leverage its findings.

Bibliography

[1] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, “What makes a photograph memorable,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1469–1482, 2013.

[2] S. Herculano-Houzel, “The human brain in numbers: A linearly scaled-up primate brain,” *Frontiers in Human Neuroscience*, vol. 3, p. 31, 2009.

[3] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, A. J. Hudspeth, and S. Mack, *Principles of Neural Science, Fifth Edition*. McGraw Hill, 2013. [Online]. Available: `https://neurology.mhmedical.com/content.aspx?bookid=1049&sectionid=59138139`.

[4] H. H. Jasper, L. D. Proctor, R. S. Knighton, W. C. Noshay, and R. T. Costello, “Reticular formation of the brain,” *Academic Medicine*, vol. 33, no. 11, p. xviii, 1958.

[5] M. F. Bear, B. W. Connors, and M. A. Paradiso, *Neuroscience: Exploring the Brain*. Wolters Kluwer, 2016.

[6] V. B. Mountcastle, “Perceptual neuroscience: The cerebral cortex,” *Harvard University Press*, 1998.

[7] D. Purves, G. J. Augustine, D. Fitzpatrick, W. Hall, A.-S. LaMantia, and L. White, *Neurosciences*. De Boeck Supérieur, 2019.

[8] J. M. Fuster, *The Prefrontal Cortex*. Academic Press, 2015.

[9] J. C. Culham and N. G. Kanwisher, "Cortical fMRI activation produced by attentive tracking of moving targets," *Journal of Neurophysiology*, vol. 86, no. 5, pp. 2657–2662, 2001.

[10] D. H. Hubel and T. N. Wiesel, "Brain and visual perception: The story of a 25-year collaboration," *Brain and Visual Perception: The Story of a 25-Year Collaboration*, pp. 1–448, 2005.

[11] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.

[12] L. R. Squire and J. T. Wixted, "The storage and recall of memories in the hippocampus and cortex," *The Oxford Handbook of Memory*, pp. 417–428, 2011.

[13] G. M. Shepherd, *The Synaptic Organization of the Brain*, 5th ed. Oxford University Press, 2004.

[14] G. Rizzolatti and M. A. Arbib, "Language within our grasp," *Trends in Neurosciences*, vol. 21, no. 5, pp. 188–194, 1998.

[15] S. G. Martin, *Synaptic Plasticity and the Mechanism of Alzheimer's Disease.* Springer, 2000.

[16] J. L. McGaugh, "Memory–a century of consolidation," *Science*, vol. 287, no. 5451, pp. 248–251, 2000.

[17] Khan Academy, *The synapse*, 2023. [Online]. Available: `https://www.khanacademy.org/science/biology/human-biology/neuron-nervous-system/a/the-synapse`.

[18] T. Splettstoesser, *Synapse schematic*, 2023. [Online]. Available: `https://commons.wikimedia.org/w/index.php?curid=41349083`.

[19] J. Panksepp, *Affective Neuroscience: The Foundations of Human and Animal Emotions.* Oxford University Press, 1998.

[20] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. MIT Press, 2014.

[21] G. G. Márquez, *One hundred years of solitude*. Penguin UK, 2014.

[22] J. L. Borges, “Funes the memorious,” in *Labyrinths*, New York: New Dir, 1964, pp. 65–71.

[23] S. Finger, *Origins of neuroscience: A history of explorations into brain function*. Oxford University Press, 2001.

[24] H. Eichenbaum, “The cognitive neuroscience of memory: An introduction,” *Oxford University Press*, 2002.

[25] L. Squire, “Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans,” *Psychological Review*, vol. 99, no. 2, p. 195, 1992.

[26] Y. Dudai, “The neurobiology of consolidations, or, how stable is the engram?” *Annual Review of Psychology*, vol. 55, pp. 51–86, 2004.

[27] E. Tulving, “Episodic memory: From mind to brain,” *Annual Review of Psychology*, vol. 53, no. 1, pp. 1–25, 2002.

[28] D. Schacter, “The seven sins of memory: Insights from psychology and cognitive neuroscience,” *American Psychologist*, vol. 54, no. 3, p. 182, 1999.

[29] S. Zola-Morgan and L. R. Squire, “Neuroanatomy of memory,” *Annual Review of Neuroscience*, vol. 16, no. 1, pp. 547–563, 1993.

[30] L. Squire, P. Slater, and P. Chace, “Retrograde amnesia for recent events: Clinical and experimental observations,” *Cortex*, vol. 12, no. 4, pp. 258–274, 1976.

[31] A. D. Baddeley, “The development of the concept of working memory: Implications and contributions of neuropsychology.,” 1990.

[32] E. Tulving, “Memory and consciousness.,” *Canadian Psychology/Psychologie Canadienne*, vol. 26, no. 1, p. 1, 1985.

[33] M. S. Gazzaniga and G. R. Mangun, Eds., *The Cognitive Neurosciences*, 5th. Boston: MIT Press, 2014.

[34] C. E. Curtis and M. D'Esposito, "Persistent activity in the prefrontal cortex during working memory," *Trends in Cognitive Sciences*, vol. 7, pp. 415–423, 9 2003.

[35] M. D'Esposito and B. R. Postle, "The dependence of span and delayed-response performance on prefrontal cortex," *Neuropsychologia*, vol. 37, no. 11, pp. 1303–1315, 1999.

[36] J. E. Lisman and M. A. Idiart, "Storage of 7±2 short-term memories in oscillatory subcycles," *Science*, vol. 267, no. 5203, pp. 1512–1515, 1995.

[37] R. S. Zucker and W. G. Regehr, "Short-term synaptic plasticity," *Annual Review of Physiology*, vol. 64, no. 1, pp. 355–405, 2002.

[38] G. V. Williams and P. S. Goldman-Rakic, "Modulation of memory fields by dopamine Dl receptors in prefrontal cortex," *Nature*, vol. 376, no. 6541, pp. 572–575, 1995.

[39] J. Lisman, A. A. Grace, and E. Duzel, "A neohebbian framework for episodic memory; role of dopamine-dependent late ltp," *Trends in Neurosciences*, vol. 34, no. 10, pp. 536–547, 2011.

[40] M. E. Hasselmo and M. Sarter, "Modes and models of forebrain cholinergic neuromodulation of cognition," *Neuropsychopharmacology*, vol. 36, pp. 52–73, 1 2011.

[41] D. Hannula *et al.*, "Relational memory impairments at short lags," *J. Neurosci.*, vol. 26, pp. 8352–8359, 2006.

[42] T. Hartley *et al.*, "Learning in a geometric model of place cell firing," *Hippocampus*, vol. 17, pp. 786–800, 2007.

[43] I. R. Olson, K. Page, K. S. Moore, A. Chatterjee, and M. Verfaellie, "Working memory for conjunctions relies on the medial temporal lobe," *Journal of Neuroscience*, vol. 26, no. 17, pp. 4596–4601, 2006.

[44] K. J. Mitchell, M. K. Johnson, C. L. Raye, and M. D'Esposito, "fMRI evidence of age-related hippocampal dysfunction in feature binding in working memory," *Cognitive Brain Research*, vol. 10, no. 1-2, pp. 197–206, 2000.

[45] N. Axmacher, S. Haupt, M. X. Cohen, C. E. Elger, and J. Fell, "Interference of working memory load with long-term memory formation," *European Journal of Neuroscience*, vol. 29, no. 7, pp. 1501–1513, 2009.

[46] K. Schon, Y. T. Quiroz, M. E. Hasselmo, and C. E. Stern, "Greater working memory load results in greater medial temporal activity at retrieval," *Cerebral Cortex*, vol. 19, no. 11, pp. 2561–2571, 2009.

[47] K. Schon, R. S. Ross, M. E. Hasselmo, and C. E. Stern, "Complementary roles of medial temporal lobes and mid-dorsolateral prefrontal cortex for working memory for novel and familiar trial-unique visual stimuli," *European Journal of Neuroscience*, vol. 37, no. 4, pp. 668–678, 2013.

[48] N. Axmacher, M. M. Henseler, O. Jensen, I. Weinreich, C. E. Elger, and J. Fell, "Cross-frequency coupling supports multi-item working memory in the human hippocampus," *Proceedings of the National Academy of Sciences*, vol. 107, no. 7, pp. 3228–3233, 2010.

[49] J. Rissman, A. Gazzaley, and M. D'Esposito, "Dynamic adjustments in prefrontal, hippocampal, and inferior temporal interactions with increasing visual working memory load," *Cerebral Cortex*, vol. 18, no. 7, pp. 1618–1629, 2008.

[50] E. Pastalkova, V. Itskov, A. Amarasingham, and G. Buzsaki, "Internally generated cell assembly sequences in the rat hippocampus," *Science*, vol. 321, no. 5894, pp. 1322–1327, 2008.

[51] B. J. Kraus, R. J. Robinson II, J. A. White, H. Eichenbaum, and M. E. Hasselmo, "Hippocampal "time cells": Time versus path integration," *Neuron*, vol. 78, no. 6, pp. 1090–1101, 2013.

[52] C. J. MacDonald, K. Q. Lepage, U. T. Eden, and H. Eichenbaum, "Hippocampal "time cells" bridge the gap in memory for discontiguous events," *Neuron*, vol. 71, no. 4, pp. 737–749, 2011.

[53] M. E. Hasselmo, *How we remember: Brain mechanisms of episodic memory.* MIT Press, 2011.

[54] E. A. Phelps, "Human emotion and memory: Interactions of the amygdala and hippocampal complex," *Current Opinion in Neurobiology*, vol. 14, no. 2, pp. 198–202, 2004.

[55] L. E. Jarrard, "What does the hippocampus really do?" *Behavioural Brain Research*, vol. 71, no. 1-2, pp. 1–10, 1995.

[56] S. C. Bir, S. Ambekar, S. Kukreja, and A. Nanda, "Julius caesar arantius (giulio cesare aranzi, 1530–1589) and the hippocampus of the human brain: History behind the discovery," *Journal of Neurosurgery*, vol. 122, no. 4, pp. 971–975, 2015.

[57] P. Broca, "Anatomie comparée des circonvolutions cérébrales. le grand lobe limbique et la scissure limbique dans la série des mammifères," *Rev Anthrop*, vol. 1, pp. 385–498, 1978.

[58] J. W. Papez, "A proposed mechanism of emotion. 1937.," *The Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 7, no. 1, pp. 103–112, 1937.

[59] P. D. MacLean, "Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain).," *Electroencephalography & Clinical Neurophysiology*, 1952.

[60] W. B. Scoville and B. Milner, "Loss of recent memory after bilateral hippocampal lesions," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 20, no. 1, p. 11, 1957.

[61] S. Corkin, *Permanent present tense: The unforgettable life of the amnesic patient, HM.* Basic Books (AZ), 2013.

[62] H. Eichenbaum, "What hm taught us," *Journal of Cognitive Neuroscience*, vol. 25, no. 1, pp. 14–21, 2013.

[63] C. M. Bird and N. Burgess, "The hippocampus and memory: Insights from spatial processing," *Nature Reviews Neuroscience*, vol. 9, no. 3, pp. 182–194, 2008.

[64] E. Tulving, *Elements of episodic memory.* Oxford University Press, 1983.

[65] L. Davachi, "Item, context and relational episodic encoding in humans," *Current Opinion in Neurobiology*, vol. 16, no. 6, pp. 693–700, 2006.

[66] J. B. Brewer, Z. Zhao, J. E. Desmond, G. H. Glover, and J. D. Gabrieli, "Making memories: Brain activity that predicts how well visual experience will be remembered," *Science*, vol. 281, no. 5380, pp. 1185–1187, 1998.

[67] A. D. Wagner, R. A. Poldrack, L. L. Eldridge, J. E. Desmond, G. H. Glover, and J. D. Gabrieli, "Material-specific lateralization of prefrontal activation during episodic encoding and retrieval," *Neuroreport*, vol. 9, no. 16, pp. 3711–3717, 1998.

[68] B. P. Staresina and L. Davachi, "Selective and shared contributions of the hippocampus and perirhinal cortex to episodic item and associative encoding," *Journal of Cognitive Neuroscience*, vol. 20, no. 8, pp. 1478–1489, 2008.

[69] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory.," *Psychological Review*, vol. 102, no. 3, p. 419, 1995.

[70] B. Milivojevic, M. Varadinov, A. V. Grabovetsky, S. H. Collin, and C. F. Doeller, "Coding of event nodes and narrative context in the hippocampus," *Journal of Neuroscience*, vol. 36, no. 49, pp. 12 412–12 424, 2016.

[71] R. M. Mok and B. C. Love, "A non-spatial account of place and grid cells based on clustering models of concept learning," *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[72] N. A. Herweg and M. J. Kahana, “Spatial representations in the human brain,” *Frontiers in Human Neuroscience*, vol. 12, p. 297, 2018.

[73] G. Umbach, P. Kantak, J. Jacobs, *et al.*, “Time cells in the human hippocampus and entorhinal cortex support episodic memory,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 45, pp. 28 463–28 474, 2020.

[74] H. Eichenbaum, “Time cells in the hippocampus: A new dimension for mapping memories,” *Nature Reviews Neuroscience*, vol. 15, no. 11, pp. 732–744, 2014.

[75] S. Thavabalasingam, E. B. O’Neil, J. Tay, A. Nestor, and A. C. Lee, “Evidence for the incorporation of temporal duration information in human hippocampal long-term memory sequence representations,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 13, pp. 6407–6414, 2019.

[76] A. Tsao, J. Sugar, L. Lu, *et al.*, “Integrating time from experience in the lateral entorhinal cortex,” *Nature*, vol. 561, no. 7721, pp. 57–62, 2018.

[77] C. J. MacDonald, S. Carrow, R. Place, and H. Eichenbaum, “Distinct hippocampal time cell sequences represent odor memories in immobilized rats,” *Journal of Neuroscience*, vol. 33, no. 36, pp. 14 607–14 616, 2013.

[78] J. J. Sakon, Y. Naya, S. Wirth, and W. A. Suzuki, “Context-dependent incremental timing cells in the primate hippocampus,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 51, pp. 18 351–18 356, 2014.

[79] G. A. Radvansky and D. E. Copeland, “Walking through doorways causes forgetting: Situation models and experienced space,” *Memory & Cognition*, vol. 34, no. 5, pp. 1150–1156, 2006.

[80] J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds, “Event perception: A mind-brain perspective.,” *Psychological Bulletin*, vol. 133, no. 2, p. 273, 2007.

[81] G. A. Radvansky and J. M. Zacks, “Event boundaries in memory and cognition,” *Current Opinion in Behavioral Sciences*, vol. 17, pp. 133–140, 2017.

[82] A. S. Bangert, C. A. Kurby, A. S. Hughes, and O. Carrasco, “Crossing event boundaries changes prospective perceptions of temporal length and proximity,” *Attention, Perception, & Psychophysics*, vol. 82, no. 3, pp. 1459–1472, 2020.

[83] S. Folkerts, U. Rutishauser, and M. W. Howard, “Human episodic memory retrieval is accompanied by a neural contiguity effect,” *Journal of Neuroscience*, vol. 38, no. 17, pp. 4200–4211, 2018.

[84] M. W. Howard, I. V. Viskontas, K. H. Shankar, and I. Fried, “Ensembles of human MTL neurons “jump back in time” in response to a repeated stimulus,” *Hippocampus*, vol. 22, no. 9, pp. 1833–1847, 2012.

[85] J. R. Manns, M. W. Howard, and H. Eichenbaum, “Gradual changes in hippocampal activity support remembering the order of events,” *Neuron*, vol. 56, no. 3, pp. 530–540, 2007.

[86] W. Mau, D. W. Sullivan, N. R. Kinsky, M. E. Hasselmo, M. W. Howard, and H. Eichenbaum, “The same hippocampal CA1 population simultaneously codes temporal information over multiple timescales,” *Current Biology*, vol. 28, no. 10, pp. 1499–1508, 2018.

[87] J. Sugar and M.-B. Moser, “Episodic memory: Neuronal codes for what, where, and when,” *Hippocampus*, vol. 29, no. 12, pp. 1190–1205, 2019.

[88] K. Ghandour, N. Ohkawa, C. C. A. Fung, *et al.*, “Orchestrated ensemble activities constitute a hippocampal memory engram,” *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.

[89] G. G. Knyazev, “EEG+ oscillatory systems and personality: A study of individual differences in healthy adults,” *Personality and Individual Differences*, vol. 53, no. 8, pp. 972–977, 2012.

[90] G. Buzsáki, “Theta rhythm of navigation: Link between path integration and landmark navigation, episodic and semantic memory,” *Hippocampus*, vol. 15, no. 7, pp. 827–840, 2005.

[91] P. Fries, J. H. Reynolds, A. E. Rorie, and R. Desimone, “Modulation of oscillatory neuronal synchronization by selective visual attention,” *Science*, vol. 291, no. 5508, pp. 1560–1563, 2001.

[92] W. Klimesch, “Alpha-band oscillations, attention, and controlled access to stored information,” *Trends in Cognitive Sciences*, vol. 16, no. 12, pp. 606–617, 2012.

[93] R. S. Blumenfeld and C. Ranganath, “Prefrontal cortex and long-term memory encoding: An integrative review of findings from neuropsychology and neuroimaging,” *The Neuroscientist*, vol. 13, no. 3, pp. 280–291, 2007.

[94] H. Eichenbaum, “A cortical-hippocampal system for declarative memory,” *Nature Reviews Neuroscience*, vol. 1, no. 1, pp. 41–50, 2000.

[95] D. Godden and A. Baddeley, “Context-dependent memory in two natural environments: On land and underwater,” *British Journal of Psychology*, vol. 66, no. 3, pp. 325–331, 1975.

[96] M. W. Howard and M. J. Kahana, “A distributed representation of temporal context,” *Journal of Mathematical Psychology*, vol. 46, no. 3, pp. 269–299, 2002.

[97] M. Moscovitch, R. Cabeza, G. Winocur, and L. Nadel, “The cognitive neuroscience of remote episodic, semantic and spatial memory,” *Current Opinion in Neurobiology*, vol. 18, no. 2, pp. 179–193, 2008.

[98] H. Eichenbaum, “Hippocampus: Cognitive processes and neural representations that underlie declarative memory,” *Neuron*, vol. 44, no. 1, pp. 109–120, 2004.

[99] R. C. O’Reilly and J. W. Rudy, “Conjunctive representations in learning and memory: Principles of cortical and hippocampal function,” *Psychological Review*, vol. 108, no. 2, p. 311, 2001.

[100] M. Carr and T. Nielsen, "The role of sleep in directed forgetting and remembering of human memories," *Cerebral Cortex*, vol. 21, no. 11, pp. 2534–2541, 2011.

[101] L. Jenkins and C. Ranganath, "Neural mechanisms of context effects on face recognition: Automatic binding and context shift decrements," *Journal of Cognitive Neuroscience*, vol. 22, no. 11, pp. 2541–2554, 2010.

[102] C. J. MacDonald, "Prospective and retrospective duration memory in the hippocampus: Is time in the foreground or background?" *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1637, p. 20 120 463, 2014.

[103] M. E. Montchal, Z. M. Reagh, and M. A. Yassa, "Precise temporal memories are supported by the lateral entorhinal cortex in humans," *Nature Neuroscience*, vol. 22, no. 2, pp. 284–288, 2019.

[104] M. Faber and S. P. Gennari, "In search of lost time: Reconstructing the unfolding of events from memory," *Cognition*, vol. 143, pp. 193–202, 2015.

[105] ——, "Effects of learned episodic event structure on prospective duration judgments.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 43, no. 8, p. 1203, 2017.

[106] Y. Ezzyat and L. Davachi, "Similarity breeds proximity: Pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity," *Neuron*, vol. 81, no. 5, pp. 1179–1189, 2014.

[107] O. Lositsky, J. Chen, D. Toker, *et al.*, "Neural pattern change during encoding of a narrative predicts retrospective duration estimates," *Elife*, vol. 5, e16070, 2016.

[108] L. Deuker, J. L. Bellmund, T. N. Schröder, and C. F. Doeller, "An event map of memory space in the hippocampus," *Elife*, vol. 5, e16534, 2016.

[109] L. Jacoby, “A process dissociation framework: Separating automatic from intentional uses of memory,” *Journal of Memory and Language*, vol. 30, no. 5, pp. 513–541, 1991.

[110] D. Maurer, R. Le Grand, and C. J. Mondloch, “The many faces of configural processing,” *Trends in Cognitive Sciences*, vol. 12, no. 6, pp. 255–260, 2007.

[111] L. R. Squire, “Memory systems of the brain: A brief history and current perspective,” *Neurobiology of Learning and Memory*, vol. 82, no. 3, pp. 171–177, 2004.

[112] T. F. Heatherton, “The neuroscience of self and self-regulation,” *Annual Review of Psychology*, vol. 57, pp. 573–598, 2006.

[113] J. P. Keenan, P. Long, G. Rhodes, and E. Ryan, “The cognitive neuroscience of self-enhancement and self-denigration,” *Annals of the New York Academy of Sciences*, vol. 1001, no. 1, pp. 210–212, 2003.

[114] R. N. Henson, “A mini-review of fMRI studies of human medial temporal lobe activity associated with recognition memory,” *Quarterly Journal of Experimental Psychology Section B*, vol. 58, no. 3-4, pp. 340–360, 2005.

[115] K. A. Ericsson and A. C. Lehmann, “Expert and exceptional performance: Evidence of maximal adaptation to task constraints,” *Annual Review of Psychology*, vol. 47, no. 1, pp. 273–305, 1996.

[116] D. L. Schacter, N. M. Alpert, C. R. Savage, S. L. Rauch, and M. S. Albert, “Recollective experience in recognition memory: Functional neuroanatomy and cognitive characteristics,” *Neuropsychology*, vol. 9, no. 4, p. 514, 1995.

[117] D. L. Schacter, D. R. Addis, and R. L. Buckner, “Remembering the past to imagine the future: The prospective brain,” *Nature reviews neuroscience*, vol. 8, no. 9, pp. 657–661, 2007.

[118] M. A. Conway, “Memory and the self,” *Journal of Memory and Language*, vol. 53, no. 4, pp. 594–628, 2005.

[119] J. T. Wixted, “Dual-process theory and signal-detection theory of recognition memory,” *Psychological Review*, vol. 114, no. 1, p. 152, 2007.

[120] A. P. Yonelinas, “Receiver-operating characteristics in recognition memory: Evidence for a dual-process model,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 20, no. 6, p. 1341, 1994.

[121] B. B. Murdock, “An analysis of the strength-latency relationship,” *Memory & Cognition*, vol. 13, no. 6, pp. 486–493, 1985.

[122] J. M. Gardiner, C. Ramponi, and A. Richardson-Klavehn, “Remembering and knowing: Two different expressions of declarative memory,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 24, no. 3, p. 789, 1998.

[123] J. R. Quamme, A. P. Yonelinas, and K. A. Norman, “Effect of unitization on associative recognition in amnesia,” *Hippocampus*, vol. 17, no. 3, pp. 192–200, 2007.

[124] M. K. Johnson, S. Hashtroudi, and D. S. Lindsay, “Source monitoring,” *Psychological Bulletin*, vol. 114, no. 1, p. 3, 1993.

[125] B. B. Murdock Jr, “The serial position effect of free recall.,” *Journal of Experimental Psychology*, vol. 64, no. 5, p. 482, 1962.

[126] M. Glanzer and N. Bowles, “Analysis of the recall of unorganized lists,” *Journal of Experimental Psychology*, vol. 71, no. 6, p. 872, 1966.

[127] E. Tulving and D. M. Thomson, “Encoding specificity and retrieval processes in episodic memory,” *Psychological Review*, vol. 80, no. 5, p. 352, 1977.

[128] R. N. Henson, “Short-term memory for serial order: The start-end model,” *Cognitive Psychology*, vol. 36, no. 2, pp. 73–137, 1998.

[129] M. A. McDaniel and G. O. Einstein, “Remembering to perform actions: A different type of memory?” *Memory for action: A distinct form of episodic memory*, pp. 25–44, 1995.

[130] C. Baldwin and J. Runkle, "Biohazard symbol design," *Archives of Environmental Health*, vol. 15, pp. 388–392, 1967.

[131] T. Valentine, "A unified account of the effects of distinctiveness, inversion, and race in face recognition," *The Quarterly Journal of Experimental Psychology*, vol. 43, pp. 161–204, 1991.

[132] T. Valentine, M. Lewis, and P. Hills, "Face-space: A unifying concept in face recognition research," *The Quarterly Journal of Experimental Psychology*, vol. 69, pp. 1996–2019, 2016.

[133] L. Light, S. Hollander, and F. Kayra-Stuart, "Why attractive people are harder to remember," *Personality and Social Psychology Bulletin*, vol. 5, pp. 269–276, 1979.

[134] E. Winograd, "Elaboration and distinctiveness in memory for faces," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 7, pp. 181–190, 1981.

[135] J. Bartlett, S. Hurry, and W. Thorley, "Typicality and familiarity of faces," *Memory & Cognition*, vol. 12, pp. 219–228, 1984.

[136] J. Vokey and J. Read, "Familiarity, memorability, and the effect of typicality on the recognition of faces," *Memory & Cognition*, vol. 20, pp. 291–302, 1992.

[137] V. Bruce, M. Burton, and N. Dench, "What's distinctive about a distinctive face?" *The Quarterly Journal of Experimental Psychology Section A*, vol. 47, pp. 119–141, 1994.

[138] T. Busey and J. Tunnicliff, "Accounts of blending, distinctiveness, and typicality in the false recognition of faces," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 27, pp. 1618–1632, 2001.

[139] L. Standing, "Learning 10000 pictures," *Quarterly Journal of Experimental Psychology*, vol. 25, no. 2, pp. 207–222, 1973.

[140] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva, "Visual long-term memory has a massive storage capacity for object details," *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, pp. 14 325–14 329, 2008.

[141] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2011, pp. 145–152.

[142] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, "Sun database: Exploring a large collection of scene categories," *International Journal of Computer Vision*, vol. 119, pp. 3–22, 2016.

[143] W. A. Bainbridge, P. Isola, and A. Oliva, "The intrinsic memorability of face photographs.," *Journal of Experimental Psychology: General*, vol. 142, no. 4, p. 1323, 2013.

[144] M. A. Borkin, A. A. Vo, Z. Bylinskii, *et al.*, "What makes a visualization memorable?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2306–2315, 2013.

[145] K. Mahowald, P. Isola, E. Fedorenko, E. Gibson, and A. Oliva, *Memorable words are monogamous: The role of synonymy and homonymy in word recognition memory*, PsyArXiv, 2018.

[146] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem, "What makes an object memorable?" In *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1089–1097.

[147] R. Cohendet, K. Yadati, N. Q. Duong, and C.-H. Demarty, "Annotating, understanding, and predicting long-term video memorability," in *Proceedings of the 2018 ACM International Conference on Multimedia Retrieval*, 2018, pp. 178–186.

[148] W. A. Bainbridge, "The memorability of people: Intrinsic memorability across transformations of a person's face.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 43, no. 5, p. 706, 2017.

[149] Q. Lin, S. R. Yousif, M. M. Chun, and B. J. Scholl, "Visual memorability in the absence of semantic content," *Cognition*, vol. 212, p. 104 714, 2021.

[150] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision research*, vol. 116, pp. 165–178, 2015.

[151] L. Goetschalckx, P. Moors, and J. Wagemans, "Image memorability across longer time intervals," *Memory*, vol. 26, no. 5, pp. 581–588, 2018.

[152] W. A. Bainbridge, E. H. Hall, and C. I. Baker, "Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory," *Nature Communications*, vol. 10, no. 1, pp. 1–13, 2019.

[153] W. A. Bainbridge, "The resiliency of image memorability: A predictor of memory separate from attention and priming," *Neuropsychologia*, vol. 141, p. 107 408, 2020.

[154] W. A. Bainbridge, D. D. Dilks, and A. Oliva, "Memorability: A stimulus-driven perceptual neural signature distinctive from memory," *NeuroImage*, vol. 149, pp. 141–152, 2017.

[155] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: A module in human extrastriate cortex specialized for face perception," *Journal of Neuroscience*, vol. 17, no. 11, pp. 4302–4311, 1997.

[156] R. Epstein and N. Kanwisher, "A cortical representation of the local visual environment," *Nature*, vol. 392, no. 6676, pp. 598–601, 1998.

[157] K. Grill-Spector, T. Kushnir, S. Edelman, G. Avidan, Y. Itzchak, and R. Malach, "Differential processing of objects under various viewing conditions in the human lateral occipital complex," *Neuron*, vol. 24, no. 1, pp. 187–203, 1999.

[158] M. W. Brown and J. P. Aggleton, "Recognition memory: What are the roles of the perirhinal cortex and hippocampus?" *Nature Reviews Neuroscience*, vol. 2, no. 1, pp. 51–61, 2001.

[159] W. Xie, W. A. Bainbridge, S. K. Inati, C. I. Baker, and K. A. Zaghloul, "Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe," *Nature Human Behaviour*, vol. 4, no. 9, pp. 937–948, 2020.

[160] W. A. Bainbridge and J. Rissman, "Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval," *Scientific Reports*, vol. 8, no. 1, p. 8679, 2018.

[161] A. Jaegle, V. Mehrpour, Y. Mohsenzadeh, T. Meyer, A. Oliva, and N. Rust, "Population response magnitude variation in inferotemporal cortex predicts image memorability," *Elife*, vol. 8, e47596, 2019.

[162] Y. Mohsenzadeh, C. Mullin, A. Oliva, and D. Pantazis, "The perceptual neural trace of memorable unseen scenes," *Scientific Reports*, vol. 9, no. 1, p. 6033, 2019.

[163] N. C. Rust and V. Mehrpour, "Understanding image memorability," *Trends in Cognitive Sciences*, vol. 24, no. 7, pp. 557–568, 2020.

[164] M. N. Hebart, A. H. Dickter, A. Kidder, *et al.*, "Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images," *PLoS ONE*, vol. 14, no. 10, e0223792, 2019.

[165] M. Kramer, M. Hebart, C. Baker, and W. Bainbridge, "Revealing the relative contributions of conceptual and perceptual information to visual memorability," *Journal of Vision*, vol. 21, no. 9, pp. 2048–2048, 2021.

[166] J. Deese and R. A. Kaufman, "Serial effects in recall of unorganized and sequentially organized verbal material.," *Journal of Experimental Psychology*, vol. 54, no. 3, p. 180, 1957.

[167] M. H. Erdelyi and J. Becker, "Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials," *Cognitive Psychology*, vol. 6, no. 1, pp. 159–171, 1974.

[168] M. Bock, “The influence of emotional meaning on the recall of words processed for form or self-reference,” *Psychological Research*, vol. 48, no. 2, pp. 107–112, 1986.

[169] D. L. Nelson and T. A. Schreiber, “Word concreteness and word structure as independent determinants of recall,” *Journal of Memory and Language*, vol. 31, no. 2, pp. 237–260, 1992.

[170] M. A. Upala, L. O. Gonce, R. D. Tweney, and D. J. Slone, “Contextualizing counterintuitiveness: How context affects comprehension and memorability of counterintuitive concepts,” *Cognitive Science*, vol. 31, no. 3, pp. 415–439, 2007.

[171] A. Paivio, R. Philipchalk, and E. J. Rowe, “Free and serial recall of pictures, sounds, and words,” *Memory & Cognition*, vol. 3, no. 6, pp. 586–590, 1975.

[172] D. C. Rubin and M. Friendly, “Predicting which words get recalled: Measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns,” *Memory & Cognition*, vol. 14, no. 1, pp. 79–94, 1986.

[173] V. M. Garlock, A. C. Walley, and J. L. Metsala, “Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults,” *Journal of Memory and Language*, vol. 45, no. 3, pp. 468–492, 2001.

[174] P. Klaver, J. Fell, T. Dietl, *et al.*, “Word imageability affects the hippocampus in recognition memory,” *Hippocampus*, vol. 15, no. 6, pp. 704–712, 2005.

[175] E. A. Kensinger and S. Corkin, “Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words?” *Memory & Cognition*, vol. 31, no. 8, pp. 1169–1180, 2003.

[176] L. L. Jacoby and M. Dallas, “On the relationship between autobiographical memory and perceptual learning.,” *Journal of Experimental Psychology: General*, vol. 110, no. 3, p. 306, 1981.

[177] I. Begg and W. A. Wickelgren, “Retention functions for syntactic and lexical vs semantic information in sentence recognition memory,” *Memory & Cognition*, vol. 2, no. 2, pp. 353–359, 1974.

[178] J. C. Bartlett, “Remembering environmental sounds: The role of verbalization at input,” *Memory & Cognition*, vol. 5, no. 4, pp. 404–414, 1977.

[179] I. Ananthabhotla, D. B. Ramsay, and J. A. Paradiso, “HCU400: An annotated dataset for exploring aural phenomenology through causal uncertainty,” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 920–924.

[180] D. Dubois, C. Guastavino, and M. Raimbault, “A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories,” *Acta Acustica United with Acustica*, vol. 92, no. 6, pp. 865–874, 2006.

[181] L. Jäncke, “Music, memory and emotion,” *Journal of Biology*, vol. 7, no. 6, pp. 1–5, 2008.

[182] J. Bigelow and A. Poremba, “Achilles’ ear? inferior human short-term and recognition memory in the auditory modality,” *PloS One*, vol. 9, no. 2, e89914, 2014.

[183] A. Thelen, D. Talsma, and M. M. Murray, “Single-trial multisensory memories affect later auditory and visual object discrimination,” *Cognition*, vol. 138, pp. 148–160, 2015.

[184] A. Schirmer, Y. H. Soh, T. B. Penney, and L. Wyse, “Perceptual and conceptual priming of environmental sounds,” *Journal of Cognitive Neuroscience*, vol. 23, no. 11, pp. 3241–3253, 2011.

[185] D. Ramsay, I. Ananthabhotla, and J. Paradiso, “The intrinsic memorability of everyday sounds,” in *Audio Engineering Society Conference: 2019 AES Intnl. Conference on Immersive and Interactive Audio*, 2019.

[186] C. Spence and J. Driver, *Crossmodal space and crossmodal attention.* Oxford University Press, 2004.

[187] M. I. Posner, M. J. Nissen, and R. M. Klein, "Visual dominance: An information-processing account of its origins and significance.," *Psychological Review*, vol. 83, no. 2, p. 157, 1976.

[188] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 1226–1233.

[189] A. Zadbood, J. Chen, Y. C. Leong, K. A. Norman, and U. Hasson, "How we transmit memories to other brains: Constructing shared neural representations via communication," *Cerebral Cortex*, vol. 27, no. 10, pp. 4988–5000, 2017.

[190] M. G. Boltz, "The cognitive processing of film and musical soundtracks," *Memory & Cognition*, vol. 32, pp. 1194–1205, 2004.

[191] E. Perego, F. Del Missier, M. Porta, and M. Mosconi, "The cognitive effectiveness of subtitle processing," *Media Psychology*, vol. 13, no. 3, pp. 243–272, 2010.

[192] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, and A. Oliva, "Multimodal memorability: Modeling effects of semantics and decay on video memorability," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 223–240, ISBN: 978-3-030-58517-4.

[193] G. Awad, A. A. Butt, K. Curtis, *et al.*, "TRECVID 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval," 2019.

[194] R. S. Kiziltepe, M. G. Constantin, C. H. Demarty, *et al.*, "Overview of the mediaeval 2021 predicting media memorability task," in *MediaEval Multimedia Benchmark Workshop Working Notes*, 2021.

[195] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of Machine Learning Rf esearch*, vol. 12, pp. 2825–2830, 2011.

[196] G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, and K. Weinberger, "Convolutional networks with dense connectivity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[197] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567.

[198] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proc. IEEE International Conference on Computer Vision*, 2015, pp. 2390–2398.

[199] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing lstm language models," in *International Conference on Learning Representations*, 2018.

[200] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 328–339.

[201] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[202] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[203] S. Hershey, S. Chaudhuri, D. P. Ellis, *et al.*, “Cnn architectures for large-scale audio classification,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 131–135.

[204] L. Sweeney, M. G. Constantin, C. H. Demarty, *et al.*, “Overview of the mediaeval 2022 predicting video memorability task,” in *MediaEval Multimedia Benchmark Workshop Working Notes*, 2022.

[205] M. G. Constantin, B. Ionescu, C.-H. Demarty, N. Q. Duong, X. Alameda-Pineda, and M. Sjöberg, “The predicting media memorability task at mediaeval 2019.,” in *Working Notes Proceedings of the MediaEval 2019 Workshop*, 2019.

[206] R. Cohendet, C.-H. Demarty, N. Duong, M. Sjöberg, B. Ionescu, and T.-T. Do, “Mediaeval 2018: Predicting media memorability task,” 2018.

[207] R. Savran Kiziltepe, M. G. Constantin, C.-H. Demarty, *et al.*, “Overview of The MediaEval 2021 Predicting Media Memorability Task,” in *CEUR Workshop Proceedings*, vol. 3181, 2021.

[208] A. G. S. de Herrera, R. S. Kiziltepe, J. Chamberlain, *et al.*, “Overview of MediaEval 2020 Predicting Media Memorability Task: What Makes a Video Memorable?” In *MediaEval Multimedia Benchmark Workshop Working Notes*, 2020.

[209] D. Azcona, E. Moreu, F. Hu, T. Ward, and A. F. Smeaton, “Predicting media memorability using ensemble models,” in *Proceedings of MediaEval 2019, Sophia Antipolis, France*, CEUR Workshop Proceedings, Oct. 2019. [Online]. Available: `http://ceur-ws.org/Vol-2670/`.

[210] A. Reboud, I. Harrando, J. Laaksonen, R. Troncy, *et al.*, “Predicting media memorability with audio, video, and text representations,” in *Proceedings of the MediaEval 2020 Workshop*, Dec. 2020. [Online]. Available: `http://ceur-ws.org/Vol-2882/`.

[211] L. Sweeney, G. Healy, and A. F. Smeaton, "Leveraging audio gestalt to predict media memorability," in *MediaEval Multimedia Benchmark Workshop Working Notes, arXiv preprint arXiv:2012.15635*, 2020.

[212] A. Reboud, I. Harrando, J. Laaksonen, R. Troncy, *et al.*, "Exploring multimodality, perplexity and explainability for memorability prediction," in *MediaEval Multimedia Benchmark Workshop Working Notes*, 2021.

[213] R. Kleinlein, C. Luna-Jiménez, and F. Fernández-Martínez, "Thau-upm at mediaeval 2021: From video semantics to memorability using pretrained transformers," in *MediaEval 2021 workshop*, 2021.

[214] Y. Lu and X. Wuez, "Cross-modal interaction for video memorability predictions," in *MediaEval 2021 workshop*, 2021.

[215] E.-R. Nguyen, H.-D. Huynh-Lam, H.-D. Nguyen, and M.-T. Tran, "Hcmus at mediaeval2021: Attention-based hierarchical fusion network for predicting media memorability," in *MediaEval 2021 workshop*, 2021.

[216] M. G. Constantin and B. Ionescu, "Using vision transformers and memorable moments for the prediction of video memorability," in *MediaEval 2021 workshop*, 2021.

[217] M. Wertheimer, "Laws of organization in perceptual forms," in *A Source Book of Gestalt Psychology*, W. Ellis, Ed., Kegan Paul, Trench, Trubner & Company, 1938, ch. 5, pp. 71–88.

[218] D. J. Peterson and M. E. Berryhill, "The gestalt principle of similarity benefits visual working memory," *Psychonomic Bulletin & Review*, vol. 20, no. 6, pp. 1282–1289, 2013.

[219] L. Goetschalckx, P. Moors, S. Vanmarcke, and J. Wagemans, "Get the picture? Goodness of image organization contributes to image memorability," *Journal of Cognition*, vol. 2, no. 1, 2019.

[220] C. v. Ehrenfels, "Über gestaltqualitäten," *Vierteljahrsschrift für wissenschaftliche Philosophie*, vol. 14, no. 3, pp. 249–292, 1890.

[221] A. R. Bowles, C. B. Chang, and V. P. Karuzis, "Pitch ability as an aptitude for tone learning," *Language Learning*, vol. 66, no. 4, pp. 774–808, 2016.

[222] R. Cohendet, C.-H. Demarty, N. Q. Duong, and M. Engilberge, "Videomem: Constructing, analyzing, predicting short-term and long-term video memorability," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2531–2540.

[223] S. Shekhar, D. Singal, H. Singh, M. Kedia, and A. Shetty, "Show and recall: Learning what makes videos memorable," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2730–2739.

[224] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis," *Brain Research Reviews*, vol. 29, no. 2-3, pp. 169–195, 1999.

[225] T. F. Sanquist, J. W. Rohrbaugh, K. Syndulko, and D. B. Lindsley, "Electrocortical signs of levels of processing: Perceptual analysis and recognition memory," *Psychophysiology*, vol. 17, no. 6, pp. 568–576, 1980.

[226] D. Karis, M. Fabiani, and E. Donchin, "P300 and memory: Individual differences in the von Restorff effect," *Cognitive Psychology*, vol. 16, no. 2, pp. 177–216, 1984.

[227] E. Noh, G. Herzmann, T. Curran, and V. R. de Sa, "Using single-trial EEG to predict and analyze subsequent memory," *NeuroImage*, vol. 84, pp. 712–723, 2014.

[228] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.

[229] N.-Y. Liang, P. Saratchandran, G.-B. Huang, and N. Sundararajan, "Classification of mental tasks from EEG signals using extreme learning machine," *International journal of neural systems*, vol. 16, no. 01, pp. 29–38, 2006.

[230] F. Ebrahimi, M. Mikaeili, E. Estrada, and H. Nazeran, "Automatic sleep stage classification based on EEG signals by using neural networks and wavelet packet coefficients," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2008, pp. 1151–1154.

[231] C. Lehmann, T. Koenig, V. Jelic, *et al.*, "Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG)," *Journal of Neuroscience Methods*, vol. 161, no. 2, pp. 342–350, 2007.

[232] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito, "A novel multimodal machine learning based approach for automatic classification of EEG recordings in dementia," *Neural Networks*, vol. 123, pp. 176–190, 2020.

[233] B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal," *Computer methods and programs in biomedicine*, vol. 109, no. 3, pp. 339–345, 2013.

[234] D. A. Engemann, F. Raimondo, J.-R. King, *et al.*, "Robust EEG-based cross-site and cross-protocol classification of states of consciousness," *Brain*, vol. 141, no. 11, pp. 3179–3192, 2018.

[235] S.-Y. Jo and J.-W. Jeong, "Prediction of visual memorability with EEG signals: A comparative study," *Sensors*, vol. 20, no. 9, p. 2694, 2020.

[236] W. A. Bainbridge, D. D. Dilks, and A. Oliva, "Memorability: A stimulus-driven perceptual neural signature distinctive from memory," *NeuroImage*, vol. 149, pp. 141–152, 2017.

[237] J. Polich, "Updating P300: an integrative theory of P3a and P3b," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.

[238] S. Sutton, M. Braren, J. Zubin, and E. John, "Evoked potential correlates of stimulus uncertainty," *Science*, vol. 150, no. 3700, pp. 1187–1188, 1965.

[239] M. Soltani and R. Knight, “P3a and P3b in task and disctraction conditions: An event-related potential study of response inhibition,” *Experimental Brain Research*, vol. 176, no. 1, pp. 116–123, 2007.

[240] E. Donchin and M. Coles, “The P300 complex and the measurement of information processing,” *Cognitive Psychophysiology: Event-related Potentials and the Study of Cognition*, vol. 3, pp. 255–307, 1988.

[241] D. Friedman, “The late positive component (LPC) in visual and auditory environments,” *International Journal of Psychophysiology*, vol. 52, no. 2, pp. 221–230, 2004.

[242] T. Curran, “The electrophysiology of incidental and intentional retrieval: Erp old/new effects in lexical decision and recognition memory,” *Neuropsychologia*, vol. 38, no. 5, pp. 677–690, 2000.

[243] M. D. Rugg and T. Curran, “Event-related potentials and recognition memory,” *Trends in Cognitive Sciences*, vol. 11, no. 6, pp. 251–257, 2007.

[244] T. Curran, “Brain potentials of recollection and familiarity,” *Memory & Cognition*, vol. 28, pp. 923–938, 2000.

[245] R. N. Henson, M. Rugg, T. Shallice, O. Josephs, and R. J. Dolan, “Recollection and familiarity in recognition memory: An event-related functional magnetic resonance imaging study,” *Journal of Neuroscience*, vol. 19, no. 10, pp. 3962–3972, 1999.

[246] W. E. Hockley and A. Consoli, “Familiarity and recollection in item and associative recognition,” *Memory & Cognition*, vol. 27, no. 4, pp. 657–664, 1999.

[247] X. Lei, K. Liao, D. Yao, and P. Xu, “The impact of the reference choice on scalp EEG connectivity estimation,” *Journal of Neural Engineering*, vol. 11, no. 3, p. 036 005, 2014.

[248] F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier, “Spherical splines for scalp potential and current density mapping,” *Electroencephalography and C linical Neurophysiology*, vol. 72, no. 2, pp. 184–187, 1989.

[249] M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort, “Automated rejection and repair of bad trials in MEG/EEG,” *In Proceedings of the 6th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2017.

[250] A. Delorme, S. Makeig, and T. J. Sejnowski, “Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis,” *Neuroimage*, vol. 34, no. 4, pp. 1443–1449, 2007.

[251] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, “Adjust: An automatic EEG artifact detector based on the joint use of spatial and temporal features,” *Psychophysiology*, vol. 48, no. 2, pp. 229–240, 2011.

[252] H. He and E. A. Garcia, “Learning from imbalanced data,” in *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, IEEE, 2009, pp. 1263–1284.

[253] O. Al Zoubi, C. Ki Wong, R. T. Kuplicki, *et al.*, “Predicting age from brain EEG signals—A machine learning approach,” *Frontiers in Aging Neuroscience*, vol. 10, p. 184, 2018.

[254] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (roc) curve,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[255] D. M. W. Powers, “Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[256] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The balanced accuracy and its posterior distribution,” in *2010 20th International Conference on Pattern Recognition*, IEEE, 2010, pp. 3121–3124.

[257] M. X. Cohen, *Analyzing neural time series data: theory and practice.* MIT Press, 2014.

[258] S. Stober, "Deep feature learning for EEG recordings," in *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2016, pp. 3000–3007.

[259] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[260] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[261] S. Makeig, A. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," *Advances in Neural Information Processing Systems*, vol. 8, 1995.

[262] S. Makeig, "Auditory event-related dynamics of the EEG spectrum and effects of exposure to tones," *Electroencephalography and Clinical Neurophysiology*, vol. 86, no. 4, pp. 283–293, 1993.

[263] C. Tallon-Baudry, O. Bertrand, C. Delpuech, and J. Permier, "Stimulus time-locking of EEG can bias the event-related spectral perturbation (ERSP) measure," *Electroencephalography and Clinical Neurophysiology*, vol. 91, no. 4, pp. 260–263, 1994.

[264] F. Lotte, L. Bougrain, A. Cichocki, *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces: A 10-year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031 005, 2018.

[265] M. D. Rugg, "ERP studies of memory," 1995.

[266] S. Makeig, S. Debener, J. Onton, and A. Delorme, "Mining event-related brain dynamics," *Trends in Cognitive Sciences*, vol. 8, no. 5, pp. 204–210, 2004.

[267] G. Pfurtscheller, "Event-related synchronization (ers): An electrophysiological correlate of cortical areas at rest," *Electroencephalography and Clinical Neurophysiology*, vol. 83, no. 1, pp. 62–69, 1992.

[268] W. Klimesch, “EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis,” *Brain Research Reviews*, vol. 29, no. 2-3, pp. 169–195, 1999.

[269] J. Allen, “Short term spectral analysis, synthesis, and modification by discrete fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[270] S. Mallat, *A wavelet tour of signal processing: The Sparse Way*. Academic Press, 2008.

[271] R. S. Kiziltepe, M. G. Constantin, C.-H. Demarty, *et al.*, “Overview of the MediaEval 2021 predicting media memorability task,” in *MediaEval Multimedia Benchmark Workshop Working Notes*, Dec. 2021. [Online]. Available: `http://ceur-ws.org/Vol-3181/`.

[272] A. Widmann, E. Schröger, and B. Maess, “Digital filter design for electrophysiological data–a practical approach,” *Journal of Neuroscience Methods*, vol. 250, pp. 34–46, 2015.

[273] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, “Efficient backprop,” *Neural Networks: Tricks of the trade*, pp. 9–48, 2012.

[274] L. Sweeney, M. G. Constantin, C.-H. Demarty, *et al.*, “Overview of the MediaEval 2022 predicting video memorability task,” in *MediaEval Multimedia Benchmark Workshop Working Notes*, Jan. 2022.

[275] R. Hamelink, *Visual recognition of the memento10k dataset: A p300 component erp study*, 2022.

[276] J. Polich and E. Donchin, “P300 and the word frequency effect,” *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 1, pp. 33–45, 1988.

[277] R. Kleinlein, E. R. Sebastián, and F. Fernández-Martínez, “Understanding media memorability from event-related potential records and visual semantics,” 2022.

[278] D. Osipova, A. Takashima, R. Oostenveld, G. Fernández, E. Maris, and O. Jensen, "Theta and gamma oscillations predict encoding and retrieval of declarative memory," *Journal of Neuroscience*, vol. 26, no. 28, pp. 7523–7531, 2006.

[279] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[280] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, 2022.

[281] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[282] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[283] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[284] Z. Bylinskii, L. Goetschalckx, A. Newman, and A. Oliva, "Memorability: An image-computable measure of information utility," *Human Perception of Visual Information: Psychological and Computational Perspectives*, pp. 207–239, 2022.

[285] A. P. Yonelinas, "The nature of recollection and familiarity: A review of 30 years of research," *Journal of Memory and Language*, vol. 46, no. 3, pp. 441–517, 2002.

[286] M. D. Rugg and K. L. Vilberg, "Brain networks underlying episodic memory retrieval," *Current Opinion in Neurobiology*, vol. 23, no. 2, pp. 255–260, 2013.

[287] N. Cowan, "The magical mystery four: How is working memory capacity limited, and why?" *Current directions in psychological science*, vol. 19, no. 1, pp. 51–57, 2010.

[288] N. Broers and N. Busch, "The effect of intrinsic image memorability on recollection and familiarity," *Memory & Cognition*, vol. 49, pp. 998–1018, 2021.

[289] J. M. Gardiner, C. Ramponi, and A. Richardson-Klavehn, "Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses," *Memory*, vol. 10, no. 2, pp. 83–98, 2002.

[290] R. M. Shiffrin, "Visual free recall," *Science*, vol. 180, no. 4089, pp. 980–982, 1973.

[291] B. Tabachnick and S. J. Brotsky, "Free recall and complexity of pictorial stimuli," *Memory & Cognition*, vol. 4, no. 5, pp. 466–470, 1976.

[292] R. C. Atkinson and R. M. Shiffrin, "The control of short-term memory," *Scientific American*, vol. 225, no. 2, pp. 82–91, 1971.

[293] L. Postman and L. W. Phillips, "Short-term temporal changes in free recall," *Quarterly Journal of Experimental Psychology*, vol. 17, no. 2, pp. 132–138, 1965.

[294] D. F. Marks, "Visual imagery differences in the recall of pictures," *British Journal of Psychology*, vol. 64, no. 1, pp. 17–24, 1973.

[295] P. W. Sheehan and U. Neisser, "Some variables affecting the vividness of imagery in recall," *British Journal of Psychology*, vol. 60, no. 1, pp. 71–80, 1969.

[296] S. Madigan, "Representational storage in picture memory," *Bulletin of the Psychonomic Society*, vol. 4, no. 6, pp. 567–568, 1974.

[297] D. M. McBride and B. A. Dosher, "A comparison of conscious and automatic memory processes for picture and word stimuli: A process dissociation analysis," *Consciousness and cognition*, vol. 11, no. 3, pp. 423–460, 2002.

[298] H. Intraub and M. Richardson, "Wide-angle memories of close-up scenes.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 15, no. 2, p. 179, 1989.

[299] F. Galton, “Statistics of mental imagery,” *Mind*, vol. 5, no. 19, pp. 301–318, 1880.

[300] B. Faw, “Conflicting intuitions may be based on differing abilities: Evidence from mental imaging research,” *Journal of Consciousness Studies*, vol. 16, no. 4, pp. 45–68, 2009.

[301] A. Z. Zeman, M. Dewar, and S. Della Sala, “Lives without imagery-Congenital aphantasia,” *Cortex*, vol. 73, pp. 378–380, 2015.

[302] X. Cui, C. B. Jeter, D. Yang, P. R. Montague, and D. M. Eagleman, “Vividness of mental imagery: Individual variability can be measured objectively,” *Vision Research*, vol. 47, no. 4, pp. 474–478, 2007.

[303] N. Dijkstra, S. E. Bosch, and M. A. van Gerven, “Vividness of visual imagery depends on the neural overlap with perception in visual areas,” *Journal of Neuroscience*, vol. 37, no. 5, pp. 1367–1373, 2017.

[304] A. Paivio, *Mental representations: A dual coding approach*. Oxford University Press, 1990.

[305] F. A. Yates, “The art of memory,” *Chicago IL*, 1966.

[306] C. R. Madan, M. G. Glaholt, and J. B. Caplan, “The influence of item properties on association-memory,” *Journal of Memory and Language*, vol. 63, no. 1, pp. 46–63, 2010.

[307] D. C. Rubin and S. Umanath, “Event memory: A theory of memory for laboratory, autobiographical, and fictional events.,” *Psychological Review*, vol. 122, no. 1, p. 1, 2015.

[308] R. Keogh and J. Pearson, “The blind mind: No sensory visual imagery in aphantasia,” *Cortex*, vol. 105, pp. 53–60, 2018.

[309] A. Radford, J. W. Kim, C. Hallacy, *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[310] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[311] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[312] R. K. Merton, "The matthew effect in science: The reward and communication systems of science are considered.," *Science*, vol. 159, no. 3810, pp. 56–63, 1968.

[313] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," To appear, 2017.

[314] D. M. Buss, *The handbook of evolutionary psychology, volume 1: Foundation.* John Wiley & Sons, 2015, vol. 1.

[315] D. R. Addis, A. T. Wong, and D. L. Schacter, "Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration," *Neuropsychologia*, vol. 45, no. 7, pp. 1363–1377, 2007.

[316] J. L. Bellmund, P. Gärdenfors, E. I. Moser, and C. F. Doeller, "Navigating cognition: Spatial codes for human thinking," *Science*, vol. 362, no. 6415, eaat6766, 2018.

[317] R. I. Dunbar and S. Shultz, "Evolution in the social brain," *Science*, vol. 317, no. 5843, pp. 1344–1347, 2007.

[318] S. Dehaene, *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin, 2014.

[319] M. H. Christiansen and N. Chater, "Language as shaped by the brain," *Behavioral and Brain Sciences*, vol. 31, no. 5, pp. 489–509, 2008.

[320] M. J. Gruber, M. Ritchey, S.-F. Wang, M. K. Doss, and C. Ranganath, “Post-learning hippocampal dynamics promote preferential retention of rewarding events,” *Neuron*, vol. 89, no. 5, pp. 1110–1120, 2016.

[321] J. M. Wolfe, “Visual attention,” *Seeing*, pp. 335–386, 2000.

[322] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, “Neuroscience-inspired artificial intelligence,” *Neuron*, vol. 95, no. 2, pp. 245–258, 2017.

[323] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, “Memorability of image regions,” *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[324] L. Sweeney, G. Healy, and A. F. Smeaton, “The influence of audio on video memorability with an audio gestalt regulated video memorability system,” in *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, IEEE, Jun. 2021, pp. 1–6.

[325] R. Snowden, R. J. Snowden, P. Thompson, and T. Troscianko, *Basic vision: an introduction to visual perception.* Oxford University Press, 2012.

[326] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva, “Conceptual distinctiveness supports detailed visual long-term memory for real-world objects.,” *Journal of Experimental Psychology: General*, vol. 139, no. 3, p. 558, 2010.

[327] S. Wiseman and U. Neisser, “Perceptual organization as a determinant of visual recognition memory,” *The American Journal of Psychology*, pp. 675–681, 1974.

[328] G. M. Huebner and K. R. Gegenfurtner, “Conceptual and visual features contribute to visual memory for natural images,” *PLoS ONE*, vol. 7, no. 6, e37575, 2012.

[329] P. G. Schyns, R. L. Goldstone, and J.-P. Thibaut, “The development of features in object concepts,” *Behavioral and Brain Sciences*, vol. 21, no. 1, pp. 1–17, 1998.

[330] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[331] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[332] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: `https://openreview.net/forum?id=YicbFdNTTy`.

[333] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, PMLR, 2022, pp. 12 888–12 900.

[334] W.-L. Chiang, Z. Li, Z. Lin, *et al.*, *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality*, Mar. 2023. [Online]. Available: `https://lmsys.org/blog/2023-03-30-vicuna/`.

[335] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[336] L. Sweeney, G. Healy, and A. F. Smeaton, "Diffusing surrogate dreams of video scenes to predict video memorability," in *MediaEval Multimedia Benchmark Workshop Working Notes*, 2023.

[337] A. García Seco de Herrera, M. G. Constantin, C. H. Demarty, *et al.*, "Experiences from the mediaeval predicting media memorability task," 2022.

[338] S. Cummins, L. Sweeney, and A. Smeaton, "Analysing the memorability of a procedural crime-drama tv series, CSI," in *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*, Sep. 2022, pp. 174–180.

[339] Y. Jiang, I. R. Olson, and M. M. Chun, "Organization of visual short-term memory.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 26, no. 3, p. 683, 2000.

[340] G. F. Woodman, S. P. Vecera, and S. J. Luck, "Perceptual organization influences visual working memory," *Psychonomic Bulletin & Review*, vol. 10, no. 1, pp. 80–87, 2003.

[341] Y. Xu, "Encoding color and shape from different parts of an object in visual short-term memory," *Perception & Psychophysics*, vol. 64, no. 8, pp. 1260–1280, 2002.

[342] ——, "Understanding the object benefit in visual short-term memory: The roles of feature proximity and connectedness," *Perception & Psychophysics*, vol. 68, no. 5, pp. 815–828, 2006.

[343] Y. Xu and M. M. Chun, "Visual grouping in human parietal cortex," *Proceedings of the National Academy of Sciences*, vol. 104, no. 47, pp. 18 766–18 771, 2007.

[344] R. S. Kiziltepe, L. Sweeney, M. G. Constantin, *et al.*, "An annotated video dataset for computing video memorability," *Data in Brief*, vol. 39, p. 107 671, 2021.

[345] L. Sweeney, G. Healy, and A. F. Smeaton, "Memories in the making: Predicting video memorability with encoding phase EEG," in *2023 International Conference on Content-Based Multimedia Indexing (CBMI)*, Accepted, yet to be published, IEEE, Jun. 2023.

[346] V. Mudgal, Q. Wang, L. Sweeney, G. Healy, and A. F. Smeaton, "Using saliency and cropping to improve video memorability," in *2024 International Conference on Multimedia Modelling (MMM)*, Accepted, yet to be published, Springer, Feb. 2024.

[347] L. Sweeney, G. Healy, and A. F. Smeaton, "The conceptual essence of intrinsic memorability," In Progress, Feb. 2024.

[348] A. García Seco de Herrera, R. Savran Kiziltepe, J. Chamberlain, *et al.*, "Overview of MediaEval 2020 predicting media memorability task: What makes a video memorable?" In *Working Notes Proceedings of the MediaEval 2020 Workshop*, 2020.

[349] P. L. Nunez and R. Srinivasan, *Electric fields of the brain: The neurophysics of EEG.* Oxford University Press, 2006.

[350] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[351] M. Wolfe Jeremy, K. R. Kluender, and D. M. Levi, *Sensation and Perception.* Oxford University Press, 2017.

[352] F. Hutmacher and C. Kuhbandner, "Long-term memory for haptically explored objects: Fidelity, durability, incidental encoding, and cross-modal transfer," *Psychological Science*, vol. 29, no. 12, pp. 2031–2038, 2018.

[353] V. Santangelo, "Forced to remember: When memory is biased by salient information," *Behavioural Brain Research*, vol. 283, pp. 1–10, 2015.

[354] S. M. Kosslyn, G. Ganis, and W. L. Thompson, "Neural foundations of imagery," *Nature Reviews Neuroscience*, vol. 2, no. 9, pp. 635–642, 2001.

[355] M. Fabiani, G. Gratton, and D. Karis, "Event-related brain potentials: Methods, theory, and applications," *Handbook of Psychophysiology*, vol. 2, pp. 85–119, 2006.

[356] F. Rosenberg and C. Ranganath, "Event-related potentials and recognition memory for pictures," *Cognitive Brain Research*, vol. 6, no. 4, pp. 351–362, 1997.

[357] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[358] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[359] C. Tallon-Baudry, O. Bertrand, C. Delpuech, and J. Pernier, “Oscillatory γ-band (30–70 hz) activity induced by a visual search task in humans,” *Journal of Neuroscience*, vol. 17, no. 2, pp. 722–734, 1997.

[360] S. Kaplan, “Environmental preference in a knowledge-seeking, knowledge-using organism.,” 1992.

[361] L. Cosmides and J. Tooby, *Origins of domain specificity: The evolution of functional organization*. Cambridge University Press, 1994, pp. 85–116.

[362] J. Simons and H. Spiers, “Functional organization of the human medial temporal lobe,” *Hippocampus*, vol. 13, no. 6, pp. 809–824, 2003.

[363] A. Delorme and S. Makeig, “EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[364] T. Naselaris, C. A. Olman, D. E. Stansbury, K. Ugurbil, and J. L. Gallant, “A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes,” *Neuroimage*, vol. 105, pp. 215–228, 2015.

[365] T. Horikawa and Y. Kamitani, “Generic decoding of seen and imagined objects using hierarchical visual features,” *Nature Communications*, vol. 8, no. 1, pp. 1–15, 2017.

[366] D. L. Schacter, “Implicit memory: History and current status.,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 13, no. 3, p. 501, 1987.

[367] C. M. Bird, “The role of the hippocampus in recognition memory,” *Cortex*, vol. 93, pp. 155–165, 2017.

[368] A. D. Wagner, D. L. Schacter, M. Rotte, *et al.*, “Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity,” *Science*, vol. 281, no. 5380, pp. 1188–1191, 1998.

[369] H. Kim, "Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies," *Neuroimage*, vol. 54, no. 3, pp. 2446–2461, 2011.

[370] A. Kafkas and D. Montaldi, "Two separate, but interacting, neural systems for familiarity and novelty detection: A dual-route mechanism," *Hippocampus*, vol. 24, no. 5, pp. 516–527, 2014.

[371] T. F. Brady, G. A. Alvarez, and V. S. Störmer, "The role of meaning in visual memory: Face-selective brain activity predicts memory for ambiguous face stimuli," *Journal of Neuroscience*, vol. 39, no. 6, pp. 1100–1108, 2019.

[372] J. S. Nairne, "Modeling distinctiveness: Implications for general memory theory," *Distinctiveness and Memory*, pp. 27–46, 2006.

[373] L. Sweeney, G. Healy, and A. F. Smeaton, "Predicting media memorability: Comparing visual, textual and auditory features," in *MediaEval Multimedia Benchmark Workshop Working Notes*, 2021.

[374] W. A. Bainbridge, P. Isola, and A. Oliva, "The intrinsic memorability of face photographs," *Journal of Experimental Psychology: General*, vol. 142, no. 4, p. 1323, 2013.

[375] C. Schuhmann, R. Beaumont, R. Vencu, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv preprint arXiv:2210.08402*, 2022.

[376] M. Bock and E. Klinger, "Interaction of emotion and cognition in word recall," *Psychological Research*, vol. 48, no. 2, pp. 99–106, 1986.

[377] I. Walker and C. Hulme, "Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 25, no. 5, p. 1256, 1999.

[378] D. L. Nelson and M. A. Friedrich, "Encoding and cuing sounds and senses.," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 6, no. 6, p. 717, 1980.

[379] E. Tulving and D. M. Thomson, "Retrieval processes in recognition memory: Effects of associative context.," *Journal of Experimental Psychology*, vol. 87, no. 1, p. 116, 1971.

[380] J. D. Bransford and J. J. Franks, "The abstraction of linguistic ideas," *Cognitive Psychology*, vol. 2, no. 4, pp. 331–350, 1971.

[381] M. Kinsbourne and J. George, "The mechanism of the word-frequency effect on recognition memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 13, no. 1, pp. 63–69, 1974.

[382] T. Zhao, I. Fang, J. Kim, and G. Friedland, "Multi-modal ensemble models for predicting video memorability," in *Proceedings of the MediaEval 2020 Workshop, CEUR Workshop Proceedings*, Dec. 2020. [Online]. Available: `http://ceur-ws.org/Vol-2882/`.

[383] A. F. Smeaton, O. Corrigan, P. Dockree, *et al.*, "Dublin's participation in the predicting media memorability task at MediaEval 2018," in *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*, 2018. [Online]. Available: `http://ceur-ws.org/Vol-2283/`.

[384] R. Gupta and K. Motwani, "Linear models for video memorability prediction using visual and semantic features," in *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*, 2018. [Online]. Available: `http://ceur-ws.org/Vol-2283/`.

[385] R. Cohendet, C.-H. Demarty, and N. Q. Duong, "Transfer learning for video memorability prediction.," in *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*, 2018. [Online]. Available: `http://ceur-ws.org/Vol-2283/`.

[386] R. Chaudhry, M. Kilaru, and S. Shekhar, "Show and Recall@ MediaEval 2018 ViMemNet: Predicting Video Memorability," in *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*, 2018. [Online]. Available: `http://ceur-ws.org/Vol-2283/`.

[387] W. Sun and X. Zhang, "Video memorability prediction with recurrent neural networks and video titles at the 2018 mediaeval predicting media memorability task.," in *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*, 2018. [Online]. Available: `http://ceur-ws.org/Vol-2283/`.

[388] S. Wang, W. Wang, S. Chen, and Q. Jin, "RUC at MediaEval 2018: Visual and Textual Features Exploration for Predicting Media Memorability," in *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*, 2018. [Online]. Available: `http://ceur-ws.org/Vol-2283/`.

[389] D. Azcona, E. Moreu, F. Hu, T. Ward, and A. F. Smeaton, "Predicting media memorability using ensemble models," in *Proceedings of MediaEval 2019, Sophia Antipolis, France*, CEUR Workshop Proceedings, Oct. 2019. [Online]. Available: `http://ceur-ws.org/Vol-2670/`.

[390] D.-T. Tran-Van, L.-V. Tran, and M.-T. Tran, "Predicting media memorability using deep features and recurrent network.," in *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*, 2018. [Online]. Available: `http://ceur-ws.org/Vol-2283/`.

[391] R. Leyva, F. Doctor, A. Garcia Seco De Herrera, and S. Sahab, "Multimodal deep features fusion for video memorability prediction," 2019. [Online]. Available: `http://ceur-ws.org/Vol-2670/`.

[392] D.-T. Tran-Van, L.-V. Tran, and M.-T. Tran, "Predicting media memorability using deep features and recurrent network," in *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*, 2019. [Online]. Available: `http://ceur-ws.org/Vol-2670/`.

[393] E. A. Kirkpatrick, "An experimental study of memory.," *Psychological Review*, vol. 1, no. 6, p. 602, 1894.

[394] M. A. Cohen, T. S. Horowitz, and J. M. Wolfe, "Auditory recognition memory is inferior to visual recognition memory," *Proceedings of the National Academy of Sciences*, vol. 106, no. 14, pp. 6008–6010, 2009.

[395] M. Larsson and L. Bäckman, "Modality memory across the adult life span: Evidence for selective age-related olfactory deficits," *Experimental Aging Research*, vol. 24, no. 1, pp. 63–82, 1998.

[396] A. Thelen and M. M. Murray, "The efficacy of single-trial multisensory memories," *Multisensory Research*, vol. 26, no. 5, pp. 483–502, 2013.

[397] A. Furnham, B. Gunter, and A. Green, "Remembering science: The recall of factual information as a function of the presentation mode," *Applied Cognitive Psychology*, vol. 4, no. 3, pp. 203–212, 1990.

[398] E. J. Rowe, "Ordered recall of sounds and words in short-term memory," *Bulletin of the Psychonomic Society*, vol. 4, no. 6, pp. 559–561, 1974.

[399] J. E. LeDoux, "Emotion, memory and the brain," *Scientific American*, vol. 270, no. 6, pp. 50–57, 1994.

[400] M. J. Chadwick, D. Hassabis, N. Weiskopf, and E. A. Maguire, "Decoding individual episodic memory traces in the human hippocampus," *Current Biology*, vol. 20, no. 6, pp. 544–547, 2010.

[401] A. Shimbo, E.-I. Izawa, and S. Fujisawa, "Scalable representation of time in the hippocampus," *Science Advances*, vol. 7, no. 6, eabd7013, 2021.

[402] L. Reddy, B. Zoefel, J. K. Possel, *et al.*, "Human hippocampal neurons track moments in a sequence of events," *Journal of Neuroscience*, vol. 41, no. 31, pp. 6714–6725, 2021.

[403] C. Vidaurre and B. Blankertz, "Towards a cure for BCI illiteracy," *Brain topography*, vol. 23, pp. 194–198, 2010.

[404] C. M. MacLeod, “Directed forgetting affects both direct and indirect tests of memory,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 15, no. 1, p. 13, 1989.

[405] M. M. Chun and N. B. Turk-Browne, “Interactions between attention and memory,” *Current Opinion in Neurobiology*, vol. 17, no. 2, pp. 177–184, 2007.

[406] M. Bywaters, J. Andrade, and G. Turpin, “Determinants of the vividness of visual imagery: The effects of delayed recall, stimulus affect and individual differences,” *Memory*, vol. 12, no. 4, pp. 479–488, 2004.

[407] A. P. Bradley, “The use of the area under the roc curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[408] M.-P. Hosseini, A. Hosseini, and K. Ahi, “A review on machine learning for EEG signal processing in bioengineering,” *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 204–218, 2020.

[409] M. R. Faller and A. D. Wagner, “Item- and task-level processes in the left inferior prefrontal cortex: Positive and negative correlates of encoding,” *NeuroImage*, 2002.

[410] A. D. Wagner, “Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity,” *Science*, 1998.

[411] A. M. Gordon, J. Rissman, R. Kiani, and A. D. Wagner, “Using spatial information as an implicit training tool in learned irrelevance,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2014.

[412] B. A. Kuhl, J. Rissman, and A. D. Wagner, “Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory,” *Neuropsychologia*, 2012.

[413] H. Lu, S. S. Chan, B. M. Lam, T. M. Lee, A. Raine, and C.-F. Lee, “Anodal transcranial direct current stimulation of the prefrontal cortex enhances complex verbal associative thought,” *Journal of Cognitive Neuroscience*, 2015.

[414] G. Xue, Q. Dong, C. Chen, Z. Lu, J. A. Mumford, and R. A. Poldrack, "Greater neural pattern similarity across repetitions is associated with better memory," *Science*, vol. 330, no. 6000, pp. 97–101, 2010.

[415] G. Xue, C. Chen, Z. Lu, and R. A. Poldrack, "Complementary role of frontoparietal activity and cortical pattern similarity in successful episodic memory encoding," *Cerebral Cortex*, 2013.

[416] I. C. Wagner, M. van Buuren, L. Bovy, and G. Fernández, "Schematic memory components converge within angular gyrus during retrieval," *eLife*, 2016.

[417] T. V. Bliss and G. L. Collingridge, "Synaptic plasticity in the hippocampus," *Nature*, vol. 361, no. 6410, pp. 31–39, 1993.

[418] R. C. Malenka, "Long-term potentiation–a decade of progress?" *Science*, vol. 260, no. 5116, pp. 1591–1594, 1993.