Reconciling the Rift Between Recognition and Recall: Insights from a Video Memorability Drawing Experiment

Lorin Sweeney  
lorin.sweeney@dcu.ie  
Dublin City University  
Dublin, Ireland

Graham Healy  
graham.healy@dcu.ie  
Dublin City University  
Dublin, Ireland

Alan Smeaton  
alan.smeaton@dcu.ie  
Dublin City University  
Dublin, Ireland

ABSTRACT

Models of computational memorability have historically been predicated on "yes/no" recognition memory games, resultantly overlooking and obscuring the variability in how we remember—from unprompted intentional detail oriented retrieval to prompted feeling based familiarity. In this paper, we detail an innovative short-term video memorability experiment which leverages drawings as a measure of recollection to explore the relationship between recognition and recall memorability of a previously-viewed video, finding evidence to suggest a measurable interaction. Our findings highlight the need to refine how we currently quantify of remembrance to more faithfully reflect its true phenomenology, and accordingly adjust our current computational models of memorability so that their downstream application in multimedia retrieval may be of higher utility.

CCS CONCEPTS
• Computing methodologies → Computer vision problems.

KEYWORDS
memorability, recognition, recall, computer vision and language, neural networks, drawing

ACM Reference Format:

1 INTRODUCTION

Recognition and recall form the basic fabric of remembrance (the act of remembering) [40]. These mechanisms, though conceptually distinct, have been shown to be entangled on a neural level [29], but the full extent of their relationship has yet to be understood. Existing computational models of memorability—defined as the likelihood that something will be remembered—stand as impressive contributions to our understanding of human memory, allowing us to predict with near human consistency, the likelihood that an image or video will be subsequently recognised after a 24 to 72 hour period. This has been demonstrated at the predicting video memorability task as part of the annual MediaEval workshops in 2023 [6], 2022 [34] and 2021 [17].

The models developed in such benchmarks, however, focus on simple recognition tasks, where a binary response of "yes" or "no" indicates whether an individual believes they have encountered a particular stimulus before [5, 14, 16, 23]. This approach makes a silent presumption, one which sidesteps the known complexity within the remembering 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 memory. Though seemingly subtle, this distinction carries considerable weight, and has numerous implications for computational memorability models, and their downstream applications in the indexing of video content to support richer multimedia retrieval.

A review of three decades of memory research indicates that the recognition process first relies on the mechanisms of recollection, before defaulting to feelings of familiarity if explicit details fail to be recalled [40]. However, recent experimental work suggests an absence of connection between recognition and recall [2]. That 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 in an image and its recognition rate was found to be equally unremarkable. These counter-intuitive conclusions present a rift that deserves resolution. Bainbridge et al.’s chosen means of analysis offers potential insight into the etiology of this contradiction. While elegant in its simplicity, their approach arguably creates a false equivalence; relying on measures of comparison that, when examined closely, appear to be incongruous: straightforwardly comparing rates of "yes/no" recognition and free recall without accounting for innate differences in processing capacity (i.e., individuals can correctly recognise upwards of 10,000 images [3, 33], which is orders of magnitude greater than recall limits [7]), and the ensuing effects they have on experimental conditions. Providing a strong counterpoint to Bainbridge et al.’s conclusions, Broers and Busch [4] employed a more refined “remember/know” procedure—participants indicate directly, after an old/new statement, whether they recall specific episodic details about an item (recollection) or whether they only know that the item is old (familiarity) [11, 37]—finding evidence to suggest that an image’s memorability scales with a greater likelihood of 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 study highlights the limitations inherent with our current models of computational memorability and raises important questions, such as, what determines whether recollection or familiarity contributes most towards memorable content, and whether these findings translate to the video domain?

The nature and content of memories has historically been difficult to quantify. Representations of recollection, captured in the form of drawings, provide a window in ways that other methodologies may not permit [2]. They offer a tangible visual output that embodies the idiosyncrasies of individual cognition, revealing subtleties that may remain hidden within conventional measures of remembrance. This paper details an innovative video recall experiment which leverages drawings as a measure of recall to provide further clarity on the interplay between recognition and recall of a previously-viewed video. How we remember naturally impacts how videos may be found and why we may be searching for them in the first place, why they have been forgotten. After a section on background and related work we describe our experimental methodology, then we present results of our experiments, and finally a conclusion to the paper.

2 BACKGROUND & RELATED WORK

Research into visual recall has been primarily foundational, focusing on basic effects to support memory system theories, and resulting in few insights into the visual attributes influencing recall performance. Probability of recall is generally regarded as a function of position in a serial presentation, with two basic effects emerging in serial-position curves—a primacy effect, increasing the recall probability of items near the start of a presentation list, and a recency effect, increasing the recall probability for items near the end of a presentation list [22, 32, 36]. This primacy effect can be attributed to the increased rehearsal of the first few items in a list, resulting in better long-term storage for these items and can be eliminated by ensuring all items receive equal amounts of rehearsal [1]. 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 [25]. The degree of vividness with which a person reports being able to visualise imagery has been found to be predictive of their recall performance [20, 31]. The strongest determinants of recall are list length and the complexity of items, with short lists of low complexity items exhibiting the greatest recall [22, 25, 31, 32, 36].

Given that more complex stimuli also eliminate the primacy and recency effects [36], many past studies’ use of simple stimuli—line drawings [8, 19, 20], or images with simple depictions of objects [13, 21]—and low resolution verbal metrics—a single word [8, 19, 21], or a brief verbal description [20, 32, 36], has resulted in little insight into the content and contributing factors of memory formation. Bainbridge et al.’s aforementioned study explicitly set out to address many of these past limitations, aiming to provide more direct insight into recalled memories and assess the relationship between “recall memorability” and “recognition memorability” [2]. They found that drawings from delayed free 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.

2.1 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. 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? This widespread intuition was formally assessed for the first time by Sir Francis Galton, who pioneered the study of mental imagery with his “breakfast-table survey”, reporting a wide variation in reported mental vividness, and some participants describing “no power of visualising” [10]. Even though surveys of mental imagery abilities [20] have consistently suggested that 2-5% of people are non-imaging/imaging impaired, the contemporary mental imaging literature still largely views non-imaging/imaging impaired individuals as ‘repressive’/neurotic’, or outright denies their existence [9]. However, with the phenomenon’s recent acquisition of a name—aphantasia: a condition of reduced or absent voluntary mental imagery [41]—the subject of inter-individual variability in internal mental representations has garnered more serious attention.

The relationship between the ability to generate vivid mental imagery and memory recall has been expressed in several notable theories. Paivio’s dual coding theory [24], 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 (or “mind palace”), a mnemonic strategy based on the creation of detailed, spatially structured mental images to facilitate information retrieval [39]. Echoing this, [18] reinforces the correlation between vivid mental imagery and recall, indicating 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 [38], are generally characterised by higher detail, thus aiding recall [28]. Taken together, these studies suggest a deep-seated nexus between vividness of mental imagery and memory recall. However, with the recognition of aphantasia, recent research examining the phenomena presents a paradox: despite the lack of vivid (or any) mental imagery, individuals with aphantasia often exhibit recall abilities akin to those with typical mental imaging capacities [15]. These seemingly contradictory findings imply that the mechanisms of recall are resilient to the absence of vivid mental imagery, and that vividness of mental imagery—in
those that can conjure it—may simply be a proxy for another quality of the stimulus which facilitates better recall.

3 METHODOLOGY

With Bainbridge et al.’s findings in mind—demonstrating that object and spatial details of images can be captured with a drawing-based visual memory experiment [2]—a novel drawing based video recall experiment was devised and carried out to investigate the nature of the relationship between recognition and recall memorability, in videos. 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 drawing experiments. A total of 32 videos were selected from the Memento10k dataset [23]—a short-term “yes/no” recognition video memorability dataset. Half of these videos were selected from the top 100 memorable videos (Figure 2), and the other half were selected from the bottom 100 (Figure 3). Videos were selected with “drawability” in mind, and representing a broad array of depiction categories.

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; videos chosen had distinct visual elements that set them apart from others, 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 contexts.

The experiments were 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 participants’ abilities to draw.

3.1 Encoding Phase

A video selection algorithm created 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 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 ensured that all videos used would produce data, and that we could account for serial position and recency effects.

The algorithm commences by shuffling both lists of selected videos (high and low memorability), 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 randomised 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 balanced and unbiased experiments, while the random elements keep the experiments challenging and engaging for the participants. A total of 52 participants, with the 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, took part in the experiments. 17 participants were dropped for failing to complete all phases of the experiment, leaving a final 35 used in the analysis phase. The median participant age was 25 and gender was not recorded.

Figure 1: Encoding phase in online drawing experiment.

During the encoding phase, videos were displayed at their native resolution, ranging from 200px by 600px to 600px by 200px. Each video was 3 seconds in duration, and they were presented following an anchoring screen lasting 5 seconds, which was consistently coloured, with a countdown. The display procedure is shown in Figure 1.
3.2 Recall Drawing Phase

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 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 which enabled participants to change the colour of their brush, resize it, undo or redo an action, and clear the canvas; a caption bar for participants to describe their drawing; and a submit button to move onto the next phase.

3.3 Perceptual Baseline

After submitting their recall drawing and caption, participants were redirected to a perceptual drawing page where they were instructed to draw a scene from the target video which was depicted on screen, and then caption their drawing before submitting. As shown in Figure 5, the interface was the same as the previous phase, but with a video scene image added.

3.4 Vividness of Mental Imagery

Upon successful completion of eight rounds of the experiments, participants were redirected to a page with the Vividness of Visual Imagery Questionnaire (VVIQ), a widely recognised self-report measure that gauges the vividness of an individual's visual imagery.
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. 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. Accordingly, the total possible score ranges from 16 (e.g., captions).

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 challenge arises when attempting to calculate the semantic similarity between a drawing and the ground-truth video frames. This comparison is naturally not straightforward as 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 [42] conditioned Stable Diffusion [27] model. Stable Diffusion is a state-of-the-art open-source text-to-image model 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 can be synthesised. These synthesised images served as a “semantic bridge”, allowing for a more valid calculation of semantic similarity between the drawings and the ground-truth video frames.

### 4 RESULTS

An essential aspect of assessing the efficacy of the drawing recall experiment lies in effectively quantifying recall. Traditional metrics: recall accuracy (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), fail to capture the complexity and nuance inherent in the process of remembering, providing a relatively artificial view of memory processes. Human cognition, in contrast, is more about comprehending the world in a meaningful, interconnected manner, rather than just cataloguing discrete details. These measures, while providing easy-to-quantify metrics, strip memories of the narrative and contextual associations that imbue them with value. An effective measure of recall, especially of complex stimuli like videos, should therefore encapsulate the essence, meaning, or narrative of the perceived stimuli. Semantic similarity, is one such measure, offering a more ecologically valid and holistic comparison between the perception of a stimulus and its reconstruction from memory. Semantic similarity can be quantified by calculating the cosine angle between two vectors in a multi-dimensional space—a smaller angle indicating greater similarity. Due to its training on image and text paired data, the Contrastive Language–Image Pretraining (CLIP) model [26] can project visual and textual embeddings into a common latent space, making it suitable for extracting and comparing semantic information from both images (e.g., video frames, and drawings) and text (e.g., captions).

#### 4.1 Drawing Based Measures

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 challenge arises when attempting to calculate the semantic similarity between a drawing and the ground-truth video frames. This comparison is naturally not straightforward as 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 [42] conditioned Stable Diffusion [27] model. Stable Diffusion is a state-of-the-art open-source text-to-image model 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 can be synthesised. These synthesised images served as a "semantic bridge", allowing for a more valid calculation of semantic similarity between the drawings and the ground-truth video frames.

#### 4.1.1 Recall Drawings vs Perceptual Drawings

Consider two drawing samples produced by a participant (Figure 6). Both consist of black scribbles. Despite a lack of discernible semantic features in these drawings, the CLIP similarity score between them rounds to one. However, this actually makes a lot of sense if we consider what the model is doing, and the fact that the dominant 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 identical colour characteristics. This highlights the importance of a minimum level 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. 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 7.

![Figure 6: Example Similarity score between a participant’s recall (A) and perceptual (B) drawings.](image-url)
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 ($t = 2.09, p = 0.051$) and the weak correlation could be explained by videos with high recognition memorability—due to their distinct and memorable content—stimulating more comprehensive and precise drawings. However, given the small effect size and marginal statistical significance, this finding should be interpreted cautiously.

Once the synthetic image has been generated, it is then compared against the 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 ground-truth 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 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. The final
similarity score between a recall drawing and a video was chosen from the highest comparison score between the synthetic recall image and each of the synthetic video-frame-based images. A 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 similarity score of 0.68 ($SD = 0.06$), $t = 3.14$, $p < 0.011$. This stronger correlation and distinct difference in means suggests a more evident relationship between video memorability and semantic alignment. The inclusion of synthetic images representing the ground-truth videos potentially provides a more accurate gauge of the semantic consistency between recall and original video content.

4.2 Textual Measures

In the context of the drawing recall experiments, textual measures serve as an illuminating counterpart to visual measures, capturing nuanced details of remembered stimuli that may not find expression in visual representations. Three axes of comparison are considered:

Recall Captions vs Perceptual Captions reflects the fidelity of recall, gauging the degree of correspondence between the semantics perceived and those recounted from memory. 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 ($\bar{x} = 0.674$, $SD = 0.051$; $t = 7.81$, $p < 0.0002$), suggesting 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 provides an external assessment of recall accuracy, reflecting the degree of semantic congruence between the recalled content and the original video narrative. A difference in mean similarity scores between high recognition memorability videos ($M = 0.67$, $SD = 0.05$) and low recognition memorability videos ($M = 0.64$, $SD = 0.06$), suggests that recalled captions of more memorable videos tend to align more with ground-truth captions, but not to a degree sufficient to yield a significant correlation with video recognition 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 the extent to which recall preserves original perception appears to be strongly linked to the recognition memorability of the video content.

4.2.1 Recall Caption Precision. A quantifiable measure of recall precision, Caption Specificity (CS), was introduced to assess the level of detail and specificity of the captions produced during the recall phase of the experiments. CS was predicated on 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: tokenization, case normalisation, and punctuation removal, applied to a corpus composed of the entire Google Conceptual Captions dataset [30]. Each unique term within a recall caption was assigned a score that was indicative of its relative significance within the caption and its rarity within the corpus, facilitating the computation of Avg TF-IDF. NEC complements Avg TF-IDF by focusing on the level of detail of the caption. The default implementation of Named Entity Recognition (NER) in the Python spaCy library [12], was used. The final CS assigned to each recall caption was computed by summing the normalised Avg TF-IDF and NEC. CS values from the experiments revealed interactions between recall caption specificity and video recognition memorability categories. A relative comparison measure was devised based on the difference between the recall CS and the ground-truth CS, normalised by the difference between the perception CS and the ground-truth CS. This created a score that encapsulated the change in specificity from perception to recall, relative to the ground-truth. A statistically significant positive correlation was observed between the normalised CS 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, when comparing
recall and perceptual caption specificity within recognition memorability categories, the average differences between the was smaller for high memorability videos compared to low 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 noted that high recognition memorability videos might inherently contain more unique, detailed, or rich concepts, which could influence the observed differences in caption recall precision.

4.3 Other Measures

Alongside the primary analysis, the recall drawing experiments incorporated two additional measures to enrich understanding of video recall memorability. The first of these addressed instances of forgotten or misremembered videos. Participants who could not remember a particular video typically left the drawing canvas blank, and wrote a statement akin to “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 (out of 140) of videos being forgotten or misremembered. Of these, one video was forgotten/misremembered by three participants and two were forgotten/misremembered 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 2nd or 3rd 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 and recall.

A second measure involved participants completing a VVIQ following the experiments, the distribution of which largely aligned with what is expected in the general population. Only one participant reported a complete absence of mental visual imagery, scoring a 16. 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 indicate an interaction between vivid mental imagery capacity and memory recall fidelity. While a strong direct correlation between mental imagery ability and recall accuracy isn’t consistent across participants, those in the top VVIQ quartile show enhanced semantic precision in their recall drawings. This enhancement may not solely indicate recall quality but could also suggest that higher VVIQ scores relate to superior drawing representation abilities, rather than enhanced recall ability alone.

5 CONCLUSION

The results in this paper suggest that there is, at the very least, not an absence of a relationship between recognition and recall memorability for videos, as indicated by Bainbridge et al.’s previous work on images [2]. Rather, there is likely a notable relationship obscured by their amalgamation in “yes/no” paradigms, which only becomes evident with more refined measures, and at extremes of the memorability spectrum. This relationship, underscored by disparities in semantic alignment and recall precision between videos of high and low recognition memorability, emphasises the need for future experiments to isolate familiarity-based and recollection-based recognition.

Our introduction of the Caption Specificity (CS) metric represents a step forward in quantifying recall precision, revealing a significant correlation between normalised scores and high video recognition memorability. Additionally, the absence of forgotten or misremembered videos in the high memorability category augments the body of evidence in favour of a general correlation between recognition and recall memorability. Incidentally, a significant enhancement in recall precision was noted among participants in the top VVIQ quartile, warranting further investigation into the relationship between imageability and memorability. However, our exploration also reveals the nascent state of our methodology and the need for refinement in both experimental design and analytical tools. The drawing-based recall experiment, while providing valuable insights, underscores the challenge of comparing recall with traditional measures of recognition, and points toward the need to evolve beyond current paradigms.

Future endeavours should focus on refining drawing analysis techniques, potentially through the development of a standardised scoring rubric that integrates novel metrics for assessing qualitative aspects of memory, such as emotional resonance or narrative coherence. Furthermore, while we focus on visual input there are other modalities such as the audio channel [35] that impact video memorability. The creation of multimodal datasets that can independently and comprehensively capture auditory, visual, and conceptual dimensions may offer a more holistic understanding of memorability across various sensory modalities. Ultimately, the imperative for a more exhaustive exploration into memorability is clear. The development of a refined computational framework that distinguishes between familiarity- and recollection-based memorability could impact how we interact with and retrieve digital content. Only by meticulously accounting for the subtleties of human memory can we pave the way for advancements in multimedia retrieval systems—enhancing our ability to remember and rediscover valuable content.

Acknowledgements: This research was conducted with financial support of Science Foundation Ireland 12/RD/2289_F2 at Insight the SFI Research Centre for Data Analytics at Dublin City University.