Effects of Human Cognition-Inspired Task Presentation on Interactive Video Retrieval

NINA WILLIS and ABRAHAM BERNSTEIN, Department of Informatics, University of Zurich, Zurich, Switzerland
LUCA ROSSETTO, Department of Informatics, University of Zurich, Zurich, Switzerland and School of Computing, Dublin City University, Dublin, Ireland

Interactive video retrieval is a cooperative process between humans and retrieval systems. Large-scale evaluation campaigns, however, often overlook human factors, such as the effects of perception, attention, and memory, when assessing media retrieval systems. Consequently, their setups fall short of emulating realistic retrieval scenarios. In this article, we study the effect of task target presentation on the rate of retrieval success in a large crowdsourced experiment. To do this, we design novel task presentation modes based on concepts in media memorability, implement the pipelines necessary for processing target video segments, and build a custom experimental platform for the final evaluation. Our findings demonstrate that the way in which the target of a video retrieval task is presented has a substantial influence on the difficulty of the retrieval task and that individuals can successfully retrieve a target video segment despite reducing or even altering the provided hints, opening up a discussion around future evaluation protocols in the domain of interactive media retrieval.

CCS Concepts: • **Information systems** → **Users and interactive retrieval**; **Retrieval tasks and goals**; **Evaluation of retrieval results**;

Additional Key Words and Phrases: Interactive Retrieval Evaluation, Video Retrieval, Human Perception and Memory

ACM Reference format:
Nina Willis, Abraham Bernstein, and Luca Rossetto. 2025. Effects of Human Cognition-Inspired Task Presentation on Interactive Video Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 21, 6, Article 159 (July 2025), 25 pages.
https://doi.org/10.1145/3727983

1 Introduction

In an era characterized by an ever-expanding wealth of information online, there is an increasing need for efficient methods of browsing and retrieving specific content. While retrieval systems have made impressive advancements, we still see that, more often than not, the retrieval process is

This work was partially supported by the Swiss National Science Foundation through project MediaGraph (contract no. 202125).
Authors' Contact Information: Nina Willis, Department of Informatics, University of Zurich, Zurich, Switzerland; e-mail: ninamari.willis@uzh.ch; Abraham Bernstein, Department of Informatics, University of Zurich, Zurich, Switzerland; e-mail: bernstein@ifi.uzh.ch; Luca Rossetto (corresponding author), Department of Informatics, University of Zurich, Zurich, Switzerland and School of Computing, Dublin City University, Dublin, Ireland; e-mail: rossetto@ifi.uzh.ch.

ACM 1551-6865/2025/7-ART159
https://doi.org/10.1145/3727983

an interactive one [58]. Retrieval systems present a set of results based on search queries, but these results may or may not provide what the searcher is looking for, whether it is due to limitations within the system or the quality of the queries themselves. In many cases, searchers scan the results and refine their queries in an iterative process, essentially engaging in a feedback loop with the retrieval system until they discover what they seek.

In order to enhance such interactive retrieval systems for optimal content retrieval, it is imperative to evaluate their performance and identify effective techniques and technologies. An inherent challenge in these evaluations is accounting for the human factor in the retrieval process and simulating realistic interactive retrieval scenarios. One such scenario is the **known-item search (KIS)** task, where the searcher aims to retrieve something they have encountered before from a database where they know the item exists. Video search evaluations, such as the **video browser showdown (VBS)**, attempt to approximate this scenario by repeatedly presenting the target segment and asking participants to find the video as quickly as possible using their retrieval system [65]. This, however, is not a realistic representation of the situation. In real life, certain details usually go unnoticed or are forgotten by the time people want to retrieve the video again. Repetitive exposure to the target video in a time-sensitive search environment may encourage heightened attention to elements typically unnoticed or unremembered, such as a watermark in the corner of the video or minuscule details in the landscape. Although media retrieval evaluations are designed to assess state-of-the-art systems for their efficiency and usability, none so far have considered the impact of human perception, attention, and memory on the interactive retrieval process. The central question then revolves around how to represent search tasks to replicate a KIS scenario for authentic video retrieval evaluations.

In this article, we study the effects of different ways of presenting the target of a KIS retrieval task on the task's success rate through a large-scale crowdsourced user experiment conducted on a dedicated platform. The different task target presentations are generated using custom preprocessing pipelines informed by the effects of perception, attention, and memory. We find that varying the representation of the target of a retrieval task has a clear effect on retrieval performance and that the type of modification applied to the target impacts end-to-end retrieval performance to different degrees. Specifically, filtered target images that approximate effects inspired by those caused by attention and memory show a slight decrease in retrieval performance. They do, however, perform better than purely textual descriptions, suggesting that processing even partial visual information is easier than imagining it based on a text. A synthetic, generated target seems even to aggravate the performance, possibly emulating misremembering a target. These findings shed new light on the effects of the design of retrieval benchmarks and the real-world scenarios they are trying to model. Future iterations of relevant benchmarks might be adjusted to reflect their target scenarios more accurately.

The contributions of this article are threefold: (1) We propose a novel aspect within the evaluation process of information retrieval system, informed by the effects of human attention and memory. (2) We present several experimental tools, including query processing pipelines introducing distortions inspired by memory and attention effects, and a platform for conducting crowdsourced interactive video retrieval experiments. (3) We perform a large-scale online retrieval experiment that substantially exceeds prior experiments in terms of the number of participants. The experiment shows that the target representation style/approach impacts overall retrieval performance, highlighting the need to consider additional aspects of real-world scenarios.

The remainder of this article is structured as follows: Section 2 discusses related work around multimedia retrieval evaluations and memorability estimation. Section 3 introduces our video processing pipelines that aim to simulate certain perception and memory effects. These pipelines are then tested in a preliminary evaluation in Section 4. Section 5 outlines the setup of our large

crowdsourced experiment, the results of which are presented in Section 6. Section 7 discusses the insights that we gain from the experiment before we offer some concluding remarks in Section 8.

2 Related Work

This section delves into works related to media retrieval evaluations, focusing on interactive video search, followed by a review of research on human attention and memory, focusing on media memorability.

2.1 Media Retrieval Evaluations

The evaluation of information retrieval techniques came about from a need for a controlled comparison of the proposed methods and, essentially, to consolidate findings that researchers can use to advance the field further. Spreading from textual search evaluations [24], evaluation campaigns have been established for image retrieval [8], as well as for video content search [69]. The current space of content-based video search evaluations consists mainly of KIS tasks—where the searcher knows of a target scene that exists in the collection and wishes to retrieve it again—and **ad hoc video search (AVS)** tasks, where an unknown number of scenes from the collection can match a given target [43]. The most prominent venue for such evaluations is the VBS, held annually as a live interactive video retrieval evaluation event [65]. It evaluates both KIS and AVS tasks, with KIS tasks being further divided into visual KIS, where the actual target scene is presented in a repetitive manner, and textual KIS, where only a textual description of the target scene is presented to the searchers. Search performance and usability are assessed through the participation of both expert and novice searchers.

In 2021, VBS held its first fully remote competition [26] by adopting the **distributed retrieval evaluation server (DRES)** [63]. Designed to facilitate the evaluation of interactive multimedia retrieval systems in both on-site and distributed settings, DRES supports all aspects of the evaluation, including task configuration and logging. It was later extended to enable asynchronous evaluations, relaxing both the locality and time constraints [62]. The remote setting of VBS 2021 demonstrated the feasibility of conducting a fully virtual evaluation of interactive systems. With the establishment of a flexible server, it is now possible to configure tasks in different ways and to more easily explore the task category space for video search evaluations, much of which is still underexplored [43].

The incorrectness or incompleteness of the searcher's understanding of their target has been a long-identified challenge of KIS [40], but little has been done to address this in retrieval system evaluations. The visual KIS task, for instance, does not consider the effects of human perception and memory. Also, the effect of different means of presenting a task's target to the evaluation participants has so far not been studied explicitly. An initial step toward mimicking such aspects uses a saliency mask based on eye-gaze information to predict and visually degrade information that would not have been attended to in the target video [57]. No substantial difference in retrieval performance was found between using the original unfiltered videos and the filtered ones, indicating that some information can be removed without negatively impacting the solubility of the video retrieval task. While this prior work simulated the effects of attention, it did not explicitly look into the effects of human memory and the memorability of media content, nor did it simulate the effects of auditory attention or memory on the audio track other than a content-independent filter, which leaves room for further exploration.

2.2 Media Memorability

After the initial sensory memory stage, in which perceived information is briefly stored, attention facilitates the advancement of certain information into short-term or working memory and eventually into long-term memory [3]. The short-term memory component has a very limited capacity

but high recognition accuracy, while the long-term memory component has lower recognition accuracy but greater capacity and stability [52]. Working memory and attention have a very close relationship with object representation in visual working memory appearing to be overwritten when attention is directed to a new object within the same category [51]. In contrast, the long-term memory component can store many objects, their overall categorical information, and details of those objects [5]. These item-specific details, however, cannot be stored in long-term memory without the support of preexisting conceptual knowledge. Visual information capacity in long-term memory, hence, depends more on conceptual structure than perceptual distinctiveness between presented objects [38], emphasizing the importance of semantics in human memory. Scene representation in memory has also been found to have a high level of fidelity, suggesting that scenes and objects are represented at the same level of abstraction in visual long-term memory [39].

Memory has been found to decay deterministically, making it possible to predict the effects of memory decay with some forgetting function [20]. The degradation of memory, however, does not occur uniformly. Certain kinds of information, such as the objects in an image and their relative locations, stay in long-term memory. Others, such as details of objects within a scene and overall spatial composition, are not as well retained [46]. More recently, memory representations of visual scenes have been found to lose not only their high-level details but also their low-level visual qualities, such as color saturation and luminance [13].

Memorability was found to be an intrinsic property of images, which means that certain types of visual content are more memorable than others, regardless of viewer demographics, context, or other high-level qualities such as aesthetics or interestingness [28, 29]. This sparked research on media memorability, including image memorability and, more recently, video memorability. Video memorability data collection began with the collection of brain activity data [23], which demonstrated useful findings but is difficult to scale and generalize. Designed based on the semantic storing of information in human memory, a more scalable protocol was proposed, in which participants were asked textual recall questions about the videos they remember [66]. Textual recall prompts, however, do not fully represent visual recall and pose several limitations, such as exclusion by language barrier. To address these issues, another protocol was proposed in which long-term video memorability annotations were collected through participants' prior memory of well-known movies [11], but this, in turn, had its own limitations, such as limited content choice and reliance on participants' subjective judgments. Collected by measuring memorability within a few minutes and 24–72 hours after memorization, VideoMem [10] was introduced as a large-scale dataset composed of videos with both short-term and long-term memorability scores. Finally, Memento10k was introduced as a dynamic video memorability dataset, enabling the prediction of not only the memorability score of a video but also its rate of decay over time [50]. The VideoMem and Memento10k datasets are often used in testing media memorability estimation models, such as in the MediaEval video memorability prediction task [71], which continues to challenge researchers today. In the following, we focus on visual and audio memorability.

2.2.1 Visual Memorability. Ever since memorability was identified as a stable and intrinsic property of an image [29], many approaches have been made toward predicting image memorability, which transfers well into the multimodal realm of video. Since visual memory is more dominant than auditory memory [4, 9], features related to the visual track are more strongly predictive of overall video memorability than the audio track. While many different features have been explored within the visual domain, most features and their relationships to memory can perhaps be explained through the lenses of *saliency* and *semantics*.

Saliency. The role of attention in visual perception has been studied extensively, with the consensus being that humans are guided by a combination of bottom-up pre-attentive processes,

which direct perception toward salient stimuli, and top-down processes, which are driven by the viewer's state of mind, including context, task, goals, and memory [25]. Without attention, significant changes to a scene [56] or even the presence of visually salient objects can go completely unperceived [67] and thus not encoded into memory. Attention and memory, in fact, have an interdependent relationship. While attention determines what will be encoded, memory guides what should be attended [7]. Since attention guides what information gets stored in memory, saliency is often explored as a feature for media memorability prediction. Within the realm of video memorability prediction in particular, Kar et al. [31] and Shekhar et al. [66] identified visual saliency early on as a key feature contributing to overall video memorability. Although not discussed extensively here, salient aspects in video may include features such as motion onset, which can attract attention [1].

Guided by theories of human attention, several studies combine bottom-up visual saliency prediction with top-down object-level approaches to estimate attentional mechanisms for the prediction of image memorability. One method identifies bottom-up saliency map coverage and object contrast level as two attention-related features that can complement and/or replace other low-level features for memorability prediction [45]. Another uses spatial histograms based on object-saliency maps of images, generated by replacing each detected object with its average bottom-up visual saliency [76]. These studies confirm that although it is able to improve predictions of memorability when taken with other features, saliency alone is not a perfect predictor of memorability [6, 16, 45]. While saliency can help predict object memorability in simple scenes, it is not as strong of a predictor when there are many points of interest [14], suggesting that memorability and salience are affected by different factors. Visually salient regions may not be the most memorable ones. Likewise, memorable regions are not necessarily the most salient ones.

However, the role of attention in memorability should not be overlooked. Again, objects must be attended to in order to be perceived and remembered, and memorable image regions correlate well with real visual fixations. Highly memorable images tend to have more consistent human fixations, with only one or a few points of focus [34, 44, 45, 76]. Memorable regions correlate with longer visual fixation durations [45]. Unlike bottom-up saliency maps, real gaze fixation maps actually correlate quite well with memorability maps [33]. The limited predictive power of saliency so far may be due to the fact that bottom-up saliency maps do not perfectly predict visual attention in real-world scenes and that top-down attentional mechanisms are complex to model accurately [78]. AMNet [16], an attention-based memorability estimation network, produces three visual attention maps, each conditioned on the one before it, to predict image memorability. Attention appears to move from pre-attentive visual saliency, generally more influenced by center bias, toward regions more responsible for memorability, as if through top-down attentional mechanisms. With the growth of available eye-tracking data, deep learning approaches can now be used to model attention and predict video saliency maps based on real gaze information [30], which enables the learning of more complex patterns of human attention.

Semantics. Top-down attentional processes are guided by semantic information accessed from long-term memory [78]. However, attention only partially explains the relationship between scene semantics and memorability, suggesting that semantic information relevant to attention and semantic information relevant to memorability are overlapping but different [44]. Semantics have been found time and time again to play an important role in media memorability, with foregrounds, especially humans and human-scale objects, contributing most positively to memorability, and backgrounds, especially exteriors and natural scenes, contributing most negatively to memorability [29, 34]. Objects in the scene appear to be particularly important, as image memorability is greatly influenced by the memorability of its most memorable object [14]. High-level object and scene

semantics were found to be the most predictive features of overall image memorability [29] and can be exploited even without the manual annotation of object labels to predict memorable image regions [35]. In line with findings from studies on human memory [38], conceptual categories influence which regions of an image will be remembered. Some categories of objects are found to be more memorable than others and even affected by the presence of other objects at different rates of decay [14].

Semantic features are also prevalent in video memorability prediction. The first video memorability prediction framework, which used functional magnetic resonance imaging to learn memorability from brain activity, found that the occurrence likelihood of each object in the video had the best prediction capability among the static visual features [23]. Although including other semantically rich features in videos, such as emotions and actions are also effective [71], high-level visual semantics, based on image captioning, was found to be the best predictor of both short-term and long-term memorability [10, 11]. For example, SemanticMemNet [50] combines visual features and semantics through joint learning of caption generation to predict not only video memorability but also its rate of decay over time. Although multimodal models, such as those combining visual and textual representations, perform best [71], even a purely text-based model, based on sentence-level embeddings of short captions, performs quite well in predicting video memorability [37]. Longer and more descriptive texts were found to correspond to videos with high memorability [21], indicating a correlation between the fidelity of semantic information and video memorability. A more recently proposed video memorability prediction model, M3-S [15], attempts to emulate the steps of human memory formation, from encoding to understanding to consolidation, through four modules: raw perception, scene parsing, event understanding, and contextual similarity. The high-level event understanding module, which handles action recognition, was found to be the most contributive to memorability prediction, again supporting the idea of memorability relying mostly on high-level semantics, with low-level features such as color and motion supplementing mostly by helping to differentiate between videos with similar semantics. Results suggest that adding additional high-level features, such as emotion, may improve results even further, as action alone cannot represent complex semantics.

2.2.2 Audio Memorability. General-purpose video content is rarely limited to visual information, but contains auditory information as well. This auditory information can be very distinct, enabling the precise description of a video sequence where it occurs. Examples might include a distinctive piece of dialog, an easily reproducible melody, or a rare but clearly identifiable noise. When composing a query to identify a specific video sequence, it can therefore be useful to consider the auditory modality as well.

Like visual memory, auditory memory can be categorized into sensory, short-term, and long-term memory stores. Since an auditory stimulus typically cannot be fully described by static features alone, the surrounding auditory context and acoustic patterns especially come into play for auditory memory [77]. Auditory signals in sensory memory are transformed into more abstract, higher-level representations in short-term memory, which can then be attended to and even drive changes in perceptual sensitivity to further incoming stimuli [81]. Certain types of auditory stimuli appear to be remembered more than others, with spoken language recognition performing best, followed by sound objects, and lastly, musical excerpts, pointing to the significance of semantics in auditory memory as well [9].

Memorability is also an intrinsic property of sound [53]. High-level conceptual features, especially causal uncertainty, visualizability, emotional valence, and familiarity, were found to be stronger predictors of sound memorability than low-level acoustic and salience features [53], similar to the way that high-level semantic features are the strongest predictors of visual memorability.

Essentially, memorable sounds tend to be familiar, emotional, easy to visualize, and come from apparent sources. Although auditory memory is not as strong as visual memory [4, 9], audio-visual integration of semantically matching tracks has been found to enhance memory performance, even in the presence of temporal asynchrony [48]. Semantically matching environmental sounds, in particular, as opposed to verbalized words of the presented objects, enhance memory performance for the position of the object, likely due to triggering attentional mechanisms toward the location of the perceived sound source [47]. Several video memorability prediction models explore the predictive power of audio features [11, 36, 54, 55, 72–74], but most found little to no effect on overall video memorability. Conditionally including audio features based on estimated audio gestalt memorability [73], however, seems to be a promising approach. High-level auditory features are used to determine the influence of the audio modality on overall memorability, and audio-augmented captions and audio spectrograms are included only if audio is predicted to be useful. Recent findings confirm the dominance of visual long-term memory over auditory memory and provide evidence suggesting that auditory information is associated with visual information as soon as it is available, as opposed to being stored independently from or fully integrated with visual information [49], which may explain the indirect effect audio features have on video memorability.

3 Preprocessing Pipelines

This section describes the design space and proposed pipelines for video preprocessing in the context of evaluating video retrieval systems in order to approximate some relevant aspects of human attention and memory.

3.1 Design Space

There are many ways to preprocess target videos to approximate the effects of human attention and memory. In the context of video retrieval, it is important to consider what information is likely to be attended to and remembered and what information is useful for retrieval tasks. Both the visual and audio tracks of a video can be filtered or transformed in different ways and at various levels of information preservation. For this work, however, we limit ourselves to visual information and omit any auditory information to keep the design space more manageable.

As discussed in the previous section, semantic information is important for encoding events in visual long-term memory. Two possible ways to simulate memory effects may be to identify semantic information and either use it to filter the original video or transform the information into a new representation entirely through the synthesis of new content representing the same semantics differently.

3.1.1 Filtering. The original target video can be filtered at different levels to mimic the effects of memory degradation. We can take advantage of the findings by Cooper et al. [13] to adjust color vibrancy to simulate the low-level degradation of scene representations in memory. Global effects can be applied to the entire video, with resolution and color saturation adjustments applied evenly throughout, or applied in a vignette style to emulate the center bias effect of attention [2]. The applied filters could also be altered across the temporal domain to make forgettable segments appear more blurry and desaturated or even removed entirely through a frame filtering process. For instance, memorability-based key frame selection and extraction have been successfully implemented in static [17] and dynamic [18] video summarization methods. Frame filtering has also been used to improve overall video memorability prediction [12]. Finally, one can adjust filters in the spatial dimensions by extracting meaningful foreground information through semantic segmentation or by emphasizing salient or memorable regions according to fixation or memorability maps, respectively.

Fig. 1. Input video.

3.1.2 Synthesis. Another idea is to abstract out the semantic content of the original videos to translate them into new ones. Synthesizing new representations of the target videos can emulate both the loss of some pieces of information and the formation of new ones, taking into account false memory formation [42]. This could be done via manual methods such as sketching, reenacting, and animating based on the artists' memories of the target video. The advantage of using manual methods is that the results can encapsulate real memory effects, as they are recreations based on the artist's actual memory of the original content. Automatic approaches, on the other hand, have the advantage of being able to streamline the video synthesis pipeline. This could be achieved by taking advantage of recent advances in generative models. Budding image or video synthesis technologies can be used to create new content from the semantics of the original video. Image generation is already impressive, with high-quality text-to-image technologies readily available to the public. Text-to-video generation has evolved rapidly from GAN approaches to transformer-based frameworks and diffusion models in the past few years. However, many of the state-of-the-art models such as Make-A-Video [68], Phenaki [75], NUWA-XL [79], and OpenAI's Sora,[1] the latter three of which are able to produce longer, more complex, higher-quality videos, are not publicly available.

3.2 Proposed Pipelines

We propose three filtering and three synthesis pipelines based on the design space discussed above. We experiment with different levels of retained information to see how they affect the retrieval process. Besides the global filtering pipeline and the text-to-video synthesis pipeline, we use video memorability as a high-level feature to filter or select frames. Memorability estimations in these pipelines are obtained via AMNet [16], an attention-based image memorability estimation network that outputs memorability scores and attention maps, both of which can be used to filter video frames in different ways. Figure 1 shows an unprocessed video frame, which we will use to showcase the effects of each pipeline.

3.2.1 Video Filtering Pipelines. The proposed filtering pipelines apply blur and desaturation in increasing levels of specificity, from global to frame to pixel-level granularity.

Global Filter (F1). A simple vignette effect is applied in the global filter, denoted as F1, so the outer regions are more blurred and desaturated than the video's center. This effect, also commonly employed in the visual arts to designate a memory or flashback, mimics our foveated visual

[1] https://openai.com/sora.

(a) F1 Output (b) F2 Output (c) F3 Output

(d) S1 Output (e) S2 Output (f) S3 Output

Fig. 2. Output of filtering and synthesis pipelines.

perception, where only the center of attention is crisp and in focus, while details at the periphery are ignored.

Frame Memorability Filter (F2). In addition to the global vignette filter, blur and desaturation are applied to the entire frame according to its estimated memorability score, thereby making less memorable details of the video inaccessible to the user. The degree of these effects may differ from frame to frame, but they are applied uniformly within each frame. Inputting the original video into this pipeline, gives us the output in Figure 2(b).

Spatial Memorability Filter (F3). In the spatial filtering pipeline, forgettable regions of the video, as estimated by generated memorability maps, are blurred and desaturated. The values from the estimated memorability maps are first raised to the power of gamma, 0.8, and thresholded so that values below 0.4 are set to 0 to generate smoother areas to which the effect is applied. We then dilate the map with a dilation radius of 4 pixels and apply a Gaussian blur to smoothen the mask gradation. A simple temporal smoothing is then applied with an alpha value of 0.6 to reduce jitter. The amount of preserved information can thus differ within and across frames. This is similar to the pipeline used in [57], except with memorability masks instead of saliency masks. An output of this pipeline is shown in Figure 2(c).

3.2.2 Video Synthesis Pipelines. The proposed synthesis pipelines below are described in order of increasing possibility for deviation from the original video. Once an input video is segmented into its individual shots using the shot boundary detection framework TransNetV2 [70], each shot is input into SwinBERT, an end-to-end transformer-based model for video captioning [41], unless video descriptions have been manually provided. The generated or provided textual descriptions are then fed into a video generation model to produce a short video per shot, which is finally combined to form a sequence of shots that capture the semantics of the original video. Two models are used for video generation: AnimateDiff [22], a text-image-to-video generator, and Text2Video-Zero [32], a zero-shot text-to-video diffusion model. The former enables greater control over the output

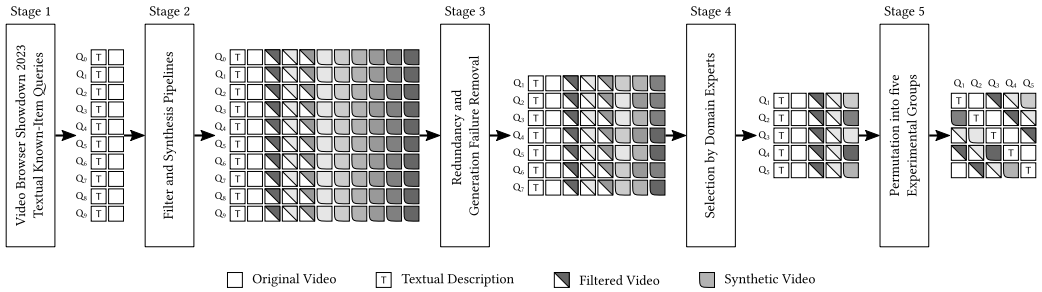


Fig. 3. Overview of the different stages used for query generation and selection.

through an initial input image, while the latter only utilizes textual descriptions to synthesize videos.

Text and Image to Video via Frame Selection (S1). The text and frame-to-video pipeline uses each shot's original frame and description to synthesize video segments. For each shot, the first frame with a memorability score above a threshold, which we define as the average memorability score of the shot, is used as the starting frame. Since the original frame is used in this scenario, the initial visualization should be closer to the original, while actions and resulting frames may deviate. Figure 2(d) shows an example output of this pipeline.

Text and Image to Video via Image Generation (S2). In the text and synthesized image to video pipeline, the starting frame per shot is selected in the same way as above, but this time, novel images are generated, expanding the space of possible deviations from the original content. For each selected frame, ControlNet [80] is used to synthesize a new image based on its textual description and conditioned on its semantic map. The synthesized starting image is then fed into a video generator along with the description of the shot to synthesize a video segment, such as the one shown in Figure 2(e).

Text to Video (S3). The final pipeline uses textual descriptions of shots directly to generate new versions of the shots. This pipeline introduces the most possible noise and false information, as it relies on the text alone to capture original video semantics. Unlike the previous two synthesis pipelines, there are no conditions restricting visual scene composition and details not mentioned in the description, which means that there is a greater likelihood for the output video to stray far from the original scene. One output of this pipeline is shown in Figure 2(f).

4 Pipeline and Query Selection

This section describes the process of selecting the video targets and variations used in the final user experiment. We began with an initial set of videos, using randomly selected tasks from the VBS 2023 archive,[2] and processed each with the pipelines described above. The set was reduced through a qualitative selection process and then further refined based on the results of a preliminary survey gathered from a panel of video retrieval experts. Throughout the rest of the article, the shorthand abbreviation for each pipeline will be used to reduce verbosity. Figure 3 provides an illustration of the query generation and selection process discussed in this section. We will refer back to the figure when discussing each of the steps.

[2]https://github.com/lucaro/VBS-Archive.

4.1 Initial Data Selection

For the initial selection of candidate task videos, we used the targets of the textual KIS tasks of the 2023 VBS. All selected tasks use videos from the V3C video dataset [61]. Videos were selected specifically from the textual KIS task type because the textual descriptions provide a consistent source of human caption input to the synthesis pipelines. Out of the 12 available tasks, we discard two because of short initial textual content descriptions, resulting in 10 candidate videos, see Figure 3 Stage 1. These videos were clipped to the segments defined in the original task. They were each processed in nine different ways: through the three filtering pipelines, the three synthesis pipelines with automatic captioning, and the three synthesis pipelines with human captioning, illustrated in Figure 3 Stage 2.

From the produced videos, qualitative selections were made between the human-captioned generated video and the machine-captioned one for each synthesis pipeline for each video.

Given two generated video clips per synthesis pipeline, we manually eliminated the less semantically accurate or more visually disturbing ones (Figure 3 Stage 3). Some videos with automated captions, for instance, were wildly different in meaning from the original video due to inaccurate caption generation. In some videos with manual captions, the descriptions were not granular enough to detail each shot, resulting in more generic videos. A couple of example clips that were eliminated due to original semantics being "lost in translation" are shown in the Appendix, where Figure A1(b) is synthesized from inputting Figure A1(a) into the S2 pipeline, and Figure A1(d) is synthesized from inputting Figure A1(c) into the S3 pipeline, both using automatically generated captions. Three out of the 10 original queries produced comparatively inferior results for all synthesis pipelines, so they were removed from the candidate pool completely.

This process resulted in 49 video clips, with seven variations (original, three filtered, three synthesized) of seven different videos.

4.2 Preliminary Study with Expert Users

A preliminary study was conducted as a survey distributed to 28 researchers, who were either organizers or long-time participants of the VBS and hence had extensive experience with interactive video retrieval and visual query presentation. We asked them to evaluate the video clips in our selected set based on perceived visual or semantic similarity to help us select the most meaningful set for our experiment, illustrated in Figure 3 Stage 4.

Filtered Videos. For filtered videos, we wanted to ensure that each variation provided perceptibly different information from the original video and each other. We showed participants the original video segment and the F1, F2, and F3 variations. We then asked them to "*Please indicate which videos, if any, you would consider to be nearly visually identical (i.e., you gain equivalent information from the videos).*" Choice options were provided in a matrix format, with subjects providing binary feedback.

Based on the results from this survey, the global vignette filter (F1) appeared to have the most similarities with other variations, especially to the original video and to the memorability-masked video (F3), as shown in Table 1. To reduce redundancy, we removed this variation in our final set.

Synthesized Videos. For the synthesized variations, the goal was to select variations that (1) are similar enough to the original video to make retrieval possible and (2) communicate different information from each other. To address these, we had two types of questions for each video: (1) "*Please indicate which video clips, if any, you would consider to be semantically similar to this video:*" followed by the original video clip, and (2) "*Please indicate which of these video clips, if any, you would consider to be visually or semantically identical (i.e., there is no meaningful difference*

Table 1. Aggregated (Binary) Similarity Votes between Original and Filtered Videos by 28 Raters

Pipeline	Original	F3	F2
F1	**137**	<u>114</u>	70
F2	48	104	
F3	74		

The largest value is in bold, and the second largest is underlined.

between them)." Survey participants could select multiple options for both question types among the S1, S2, and S3 video clips. From the answers to these questions, no difference was found in the level of semantic similarity to the original video between S1, S2, and S3, each receiving 29.5%, 26.9%, and 33.6% of the votes, respectively. Rather, the videos that most people found to be more semantically close to the original depended heavily upon the video clip.

Between the synthesized variations, S1 (44% of votes) and S2 (39.2% of votes) were the most similar across the seven video clips. This is unsurprising, considering S1 and S2 are generated from the same video synthesis model.

Since we want the synthesized variation to be similar enough to the original to provide meaningful information for retrieval, and there is no single synthesized variation that is consistently meaningful for all video clips, we selected only the variation that most people found to be most like the original for each video clip. Five of the remaining seven videos were selected based on the number of votes for their most similar synthesized clip.

As a final result, this process resulted in two filtered versions (F2 and F3) and one synthesized version based on the best-rated pipeline per video.

5 Experimental Setup

In our main experiment, we want to determine the effects of the different ways of presenting the search target on the retrieval performance. The experimental setup consists of a wrapper around the DRES [59], which provides KIS tasks and receives submissions, and the interactive video retrieval system, vitrivr [60], through which participants can search and send their submissions. This section outlines the components of the platform and how they interact.

5.1 Crowdsourcing Platform

Prolific[3] is an online platform designed to enable large-scale data collection for academic research or machine learning model training and evaluation. We chose this platform over many others for its ease of use and thorough participant verification and bias mitigation processes. When a participant starts the experiment, the participant's ID within Prolific is passed into our demographic survey on Qualtrics,[4] and then into our custom platform, where it is used to match a participant to their DRES credentials. Once all tasks are complete, the participant is redirected back to Prolific with a completion code.

5.2 Experimental Platform

After the demographic survey, there are three different stages that the participant can see on the frontend of the platform. First is the starting screen, where instructions are provided. Clicking the

[3]https://www.prolific.co.
[4]https://www.qualtrics.com.

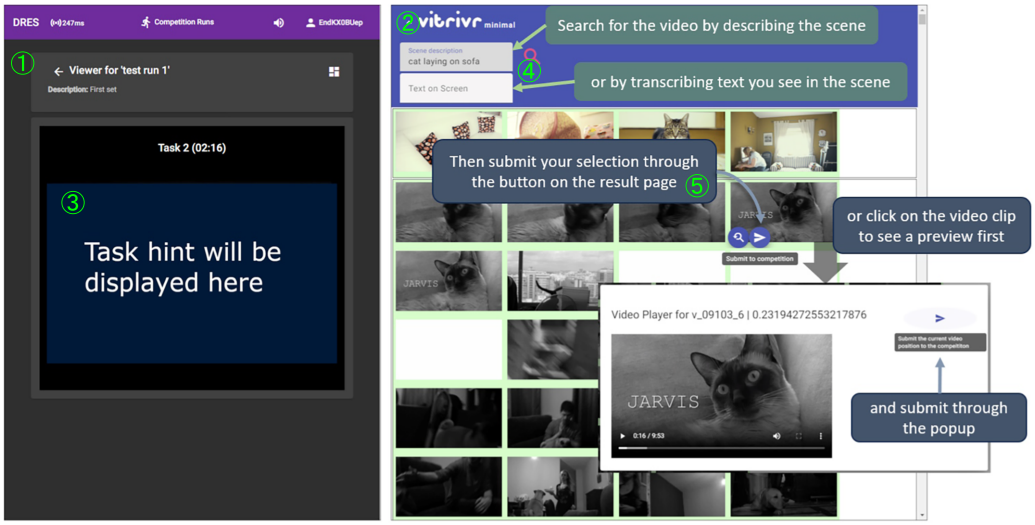


Fig. 4. Platform screenshot (cf. Section 5.2 for a discussion).

"Start" button on this screen triggers the backend to assign DRES credentials and a vitrivr instance to the participant. Credentials are assigned randomly from a list of preregistered accounts. If a participant has already been assigned to an account, their preexisting credentials will be reused such that they can pick up from where they left off. These credentials are used to automatically log in to the DRES system and obtain the session token.

Once the DRES credentials and vitrivr instance are assigned, the frontend renders two iframes, one with a viewer of the task from DRES (Figure 4 ①) and another containing vitrivr's minimal frontend (Figure 4 ②). The participant must search and submit their answers through the vitrivr iframe based on what they are shown from DRES until a correct submission is made or until the time is up, whichever comes first. Figure 4 shows a screenshot of the main screen, annotated with instruction bubbles for the participant. This is the image we display on the starting screen to familiarize participants with the task layout.

The hints for the different tasks are displayed in the box on the left side (Figure 4 ③). Participants can then use the text input boxes on the top (Figure 4 ④) to formulate their queries. The retrieved results will subsequently be displayed below. Once a participant has found what they think to be the correct target, they can submit it using the submit button (Figure 4 ⑤).

When all tasks are complete, the participant is shown the final screen, where they can click the "Finish" button to be directed back to Prolific and mark their completion.

5.2.1 DRES. As discussed in Section 2, DRES facilitates the evaluation of interactive multimedia retrieval systems, including remote and asynchronous settings [62]. This flexibility enables the crowdsourced exploration of different task presentations, which perfectly applies to our experimental setup. From the administrative perspective, DRES allows us to configure an evaluation with a specific database, targets, task presentation, duration, and participants. From the participant perspective, DRES displays the target, or a representation of the target, and receives participant submissions. DRES provides an API that we use to make the participant's experience as smooth and seamless as possible.

5.2.2 vitrivr. The interactive video retrieval system we use in our platform is vitrivr [60]. This system offers a great variety of additive query modes. Still, we stick with their minimal frontend,

Table 2. Experimental Conditions

	Task 1 (01140)	Task 2 (02024)	Task 3 (05722)	Task 4 (13872)	Task 5 (14700)
Condition 1	Original	F2	F3	S	Textual
Condition 2	F2	F3	S	Textual	Original
Condition 3	F3	S	Textual	Original	F2
Condition 4	S	Textual	Original	F2	F3
Condition 5	Textual	Original	F2	F3	S

vitrivr-ng-min [64], for the sake of simplicity and to avoid overwhelming potential novice users. This frontend supports textual queries, namely text displayed on the screen and scene description. Feature extraction must be conducted on the same video collection that was input to DRES before any video clip can be retrieved. In addition to video search and retrieval, the vitrivr frontend supports making submissions to DRES directly through a button on the retrieved video segment. The DRES session token obtained upon authenticating the participant is passed from our platform to vitrivr as soon as the vitrivr iframe is loaded to enable task submissions.

5.3 Experimental Setup

Using the final video set derived from the preliminary study, we set up five evaluations on DRES, one per experimental condition. Each condition consists of five tasks, where each task is a unique video and variation combination, as summarized in Table 2 and illustrated in Figure 3 Stage 5.

Textual KIS tasks were included in the final evaluations to allow for comparison. We used only the first textual hint from VBS to keep it consistent with the description we used as manual input to the relevant synthesis pipelines. Table A1 in the Appendix contains the descriptions of the videos in the final set.

We aimed to have 30–40 participants per condition and registered 50 members each to leave room to make up for incomplete or returned participation after DRES credentials assignment. The platform design and study instructions were iterated from the feedback we received through test runs. Having 500 distractor videos in the database and a duration of 3 minutes per task were set to balance the smaller pool of options with a shorter time limit. These parameters were also confirmed to be sufficient in our test runs. The participants we recruited for our study had a balanced distribution among males and females and were screened for English proficiency and high approval rates (80–100).

6 Results

Excluding those who withdrew from the evaluation, did not give consent, or did not make any submission attempts, we gathered a total of 200 valid participants for this study. For evaluations 1 through 5, corresponding to Table 2, we had 35, 43, 43, 36, and 43 participants, respectively. Most participants indicated that they were familiar with using video retrieval systems. The participants' technology interaction affinity scores, measured by the ATI scale [19], were also on the higher end, with more than 90% of the participants having a score above 3.5 out of 6. No correlation between ATI scores and video retrieval performance could be identified.

Answer Correctness. Table 3 shows the submissions that are correctly within the target segment, miss the target segment by at most 30 seconds, between 30 seconds and 1 minute, and more than 1 minute, as well as the submissions that missed the target video entirely. It also shows the total number of submissions per task type as well as the total instances per task type. A detailed breakdown per task variant is shown in Table A2 in the Appendix. The fractional numbers show

Table 3. Total Number of Submissions per Task Type and Submission Result

	Correct	Within 30 Seconds	Within 1 Minute	Within Video	Wrong	Total	Tasks
Original	117 (26.8%)	103 (23.6%)	25 (5.7%)	80 (18.3%)	111 (25.5%)	436 (21.8%)	197
F2	91 (26.4%)	46 (13.3%)	13 (3.8%)	40 (11.6%)	155 (44.9%)	345 (17.3%)	198
F3	104 (22.7%)	96 (21.0%)	41 (9.0%)	108 (23.6%)	109 (23.8%)	458 (23.0%)	199
S	6 (2.1%)	37 (13.0%)	25 (8.7%)	54 (18.8%)	165 (57.5%)	287 (14.4%)	198
Textual	33 (7.0%)	109 (23.1%)	50 (10.6%)	81 (17.2%)	199 (42.2%)	472 (23.6%)	196
Total	351 (17.6%)	391 (19.7%)	154 (7.7%)	363 (18.2%)	739 (37%)	1998	988

We distinguish between submissions that are within the target segment, within 30 seconds or a minute of the target segment, within the target video, or outside the target video. The Tasks column shows the number of individual task instances. Since some participants did not complete all five tasks, the total number of task instances per type is below 200. Percentages are shown per task type (column-wise) and per total across task types and time intervals.

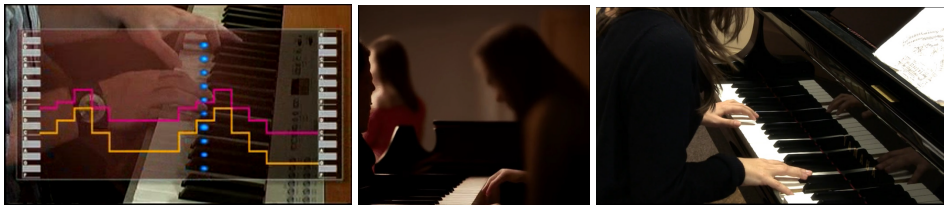
the ratio between correctness types per task type as well as the proportion of submissions per task type with respect to the total number of submissions. Since a few participants did not complete the entire experiment, we were only able to obtain slightly below the targeted 200 instances per task type. It is important to note that while a participant could make an arbitrary number of incorrect submissions per task, the task would end after a correct submission.

We can see that the unmodified video target was found most often, followed by the two filtering pipelines. The lowest number of correct submissions, in both absolute and relative terms, as well as the lowest number of total submissions, can be seen for the tasks using a synthetic video target. The tasks with textual targets, while also having a comparatively small number of correct submissions, have a much larger number of near-misses when compared to the synthetic targets. The general success rate for solving any task, i.e., the number of correct submissions divided by the total number of tasks per type, ranges from 59.4% for unmodified targets to 3% for synthetic targets.

Time for Task Completion. When considering only the correct submissions, the distribution of the times taken from the start of a task until the participants could find the correct segment can also provide some insights into the difficulties of the task types. The mean time in seconds within one SD for the different task types is as follows: 97.6 ± 38.7seconds for the original targets, 105.3 ± 43.3seconds for F2, 96.3 ± 38.9seconds for F3, 131.1 ± 26.1seconds for S, and 118.3 ± 34.4seconds for the textual tasks. We can see no substantial difference in the times for the Original and F3 tasks, with F2 tasks taking only slightly longer to solve. Tasks with synthetic video targets needed by far the longest time to be solved, with, on average, over 2 out of the available 3 minutes per task.

Query Terms. Looking at the search terms corresponding to wrong video submissions, some common patterns include terms that are not related to the video hint itself (e.g., "Second set," "Competition Runs"), descriptions that are too general (e.g., "music," "wedding," "race"), and terms that focused on aspects of the provided hint that are not relevant to the actual target (e.g., "kayak blurry," "AI generated climber").

Most Common Failure. The most commonly submitted incorrect video among all task types and videos is Video 14607 given a synthesized variation of the target Video 02024. This is, in fact, the only synthesized video task that received no correct submissions. Taking a closer look at this, we can see that Video 14607 appears to be visually closer to the synthesized task hint than the synthesized video is to the target video. Figure 5(a) shows a frame of the original target video. The synthesized task hint is displayed in Figure 5(b), and a frame of the commonly submitted Video 14607 is shown in Figure 5(c). This says more about the quality of the synthesized video than the ability of participants to extract relevant semantic information from the hint provided. Limitations



(a) Target (b) Hint (c) Common Submission

Fig. 5. Most common wrong video submission for video 02024, given a synthesized video hint.

in different stages in the synthesis process may make identifying and representing more complex qualities, such as transparent graphical overlays, difficult.

7 Discussion

The findings emerging from the user experiment offer valuable insights into the effect of different target representations on the overall retrieval performance. In this section, we examine the implications of these findings, identify the limitations inherent in this study, and explore potential avenues for future work in the domain of interactive media retrieval evaluations.

7.1 Interpretation

The filtered video variations in this study exhibit a level of comparability to the original Visual KIS task. In contrast, the synthesized variations currently demonstrate a greater degree of comparability to the textual KIS task. Results from the user experiment provide us with an initial insight into the range and effects of different task presentation methods and what this could mean in relation to our cognitive processes.

From the results, we can see that filtering a target video to blur and desaturate less memorable frames (F2) or regions (F3) can produce similar video retrieval results to showing the target as a hint itself. In general, F2 variations tend to appear blurrier than F3 variations due to the entire frame being blurred at the same level, while F3 variations appear to spotlight more memorable regions. This could potentially be related to why a greater proportion of wrong submissions in the F2 task type are not even within the correct video, as more participants could be searching with terms related to the blurriness of the F2 video.

Although the proportions of incorrect submission types differ between F2 and F3 task types, it is noteworthy that the F2 task received a good proportion of correct submissions, comparable to the unfiltered task hint, and slightly surpassing F3. Upon a closer examination of individual task results, participants generally outperformed on F3 tasks compared to F2 tasks. However, there are a couple of exceptions, namely video 14700, where F2 garnered more correct submissions but fewer near-correct submissions than F3, and video 02024, where F2 received a significantly higher proportion of correct submissions (48.8%) compared to F3 (14.4%). In these instances, perhaps the overall blur of F2 filtering reduced the prominence of distracting features within the video, or the effects of F3 filtering were more pronounced in specific areas of the video. The latter would suggest that regions estimated to be forgettable were, in fact, containing elements people used in their search queries. Generally, *the higher prevalence of incorrect submissions in filtered tasks compared to the original unfiltered task type indicates some information loss due to filtering that affects video search.*

The synthesized task type resembles the textual task because three of the synthesized video hints directly mirror the textual hints of their respective videos, and the participants face greater

difficulty in pinpointing the correct video segments in both scenarios. In both cases, the hint is more abstract, relying heavily on the participant's interpretation of the information provided. Although the synthesized and textual task hints have similar submission accuracies in general, submissions for synthesized task types are slightly worse overall, suggesting that the *artificially generated visualizations may either be distracting, causing participants to focus more on its visual aspects than the semantic ones, or portraying new or inaccurate semantic information, causing participants to misinterpret the idea of the target.*

Two cases comparing synthesized and textual tasks stand out in particular. One is video 02024, where the textual task had some correct submissions and many incorrect submissions that were still within the target video. In contrast, the synthesized version had no correct submissions, and all incorrect submissions were from outside the target video. This disparity likely stems from a significant missing element in the synthesized video—the absence of the colorful graphical overlay. Complex videos, such as those with transparent layers, may lose some meaning when processed through the synthesis pipelines due to the limited capabilities of the models themselves. Another interesting case is video 05722, where the textual task yielded no correct submissions or even submissions within the target video. The synthesized video, on the other hand, had a small portion of submissions that were correct within a minute of the target segment or at least within the same video. This could be a case where the synthesized hint provides something meaningful and useful beyond text alone.

While *both textual and synthesized task representations extract and portray the semantics of a video, the synthesized variations take it a step further by reinterpreting the extracted semantics.* The synthesis pipelines' premise was that high-level semantics, the aspect most crucial to memorability, are effectively captured by textual descriptions. Generating a visualization based on the text yields a common interpretation of the semantics for all participants. In contrast, textual KIS tasks rely more on individual interpretation and can be influenced by personal experience, context, and language proficiency. *Through video synthesis mechanisms, we can provide a common visualization and fill in visual details in potentially inaccurate ways, bridging gaps with "fake memories"* that are consistently simulated across participants. One challenge, however, is determining the degree of deviation from the original source that replicates real memory effects. Various methods of synthesizing videos were explored, including using manual or automatic captions, frame-based video generation, frame-to-image synthesis with semantic mapping, and text-to-video synthesis. No one method consistently outperformed the others, varying instead from video to video. This variability could be attributed to this specific set of videos or may stem from technological limitations. Nonetheless, a more comprehensive study on video synthesis would be needed to determine the most appropriate method for this use case.

In summary, we found that different ways of presenting the target of a retrieval task have a clear effect on overall retrieval performance. Applying filtering pipelines that aim to simulate effects similar to those caused by attention and memory shows a slight decrease in retrieval performance, indicating that such effects should be considered in interactive retrieval evaluations. Using filtered versions of the original video did, however, still lead to a higher rate of solved tasks compared to the use of a textual description. This suggests that finding the target based on even partial visual information is easier than having to imagine the target based on a textual summary alone. Using a synthetic target appears to make the task even more difficult, even though it alleviates the differences caused by different users imagining the target in different ways. This type of target representation can be interpreted as a way of causing a consistent way of misremembering a target, which deserves further study.

This has implications for the task design of evaluations as it questions the external validity of some of the designs. Someone aiming to find a video from memory appears to be a vastly different

task than being told to find some video containing some scene based on a textual description or even the task of finding something similar to some given artifact. Hence, engaging in use-case studies of the actual usage of multi-media retrieval methods would be valuable in informing the next generation of benchmark design.

7.2 Limitations

The constraints in this work mainly stem from technology or scope. As we relied on open source materials for constructing different experimental pipelines, we were limited by the technologies that are currently openly available. The second type of limitation is more intentional. As there are many possible paths we can take in an exploratory study, we needed to select which areas to apply our focus, inevitably leaving out other aspects for future work.

As discussed previously, this study explored artificially generated video clips, among other methods, to convey hints in KIS tasks. At this stage, however, it is crucial to acknowledge that the openly available video synthesis technologies exhibit certain limitations. The generated clips often suffered from distortions and instability, with objects seemingly morphing rather than moving smoothly over time. Such anomalies can potentially confuse viewers or divert their attention toward the unintended visual effects rather than the intended semantic content. Given the rapid advancements in the field of video synthesis and the growing body of research focused on improving these techniques, it may be worthwhile to revisit this approach in the future, exploring different, more advanced models that can mitigate these issues.

Another technical constraint arises from the memorability prediction method. The availability of pre-trained video memorability models, which also enable the extraction of spatial mappings, remains limited. Predicting video memorability with precision continues to be a considerable challenge. While this study incorporated an image memorability estimation model into the pipelines, there is potential merit in investigating the use of a video-specific model to enhance the accuracy and stability of predicting memorable regions within videos.

Constraints on what conclusions can be drawn also arise simply from the limited scope of this study. The pipelines used to preprocess KIS task hints are derived from heuristics and models based on research on human perception and memory but are not independently validated to be accurate representations of such effects. While we can see that each pipeline produces a unique effect on video retrieval success, it remains uncertain which, if any, most faithfully emulates real memory effects. Additionally, we focused on visual and automated approaches, but numerous other methods can be explored.

7.3 Future Work

This work can be expanded upon in several different ways. For instance, one could go in the direction of validating whether or not any of the presented pipelines effectively capture real perception and memory effects. It could also be worth continuing to explore other task presentation methods and evaluation protocols.

A logical progression in this research would involve an experimental examination of the influences of perception and memory on video retrieval. In such an experiment, participants would be presented with the actual target segment and then asked to retrieve it after a certain amount of time has elapsed. The resulting findings could then be compared to the outcomes of this study, shedding light on which pipeline most faithfully captures the effects of perception and memory. There are many well-known studies on human memory and even video memorability, but they are conducted by re-showing videos or using other recall prompts [10, 11, 66]. It would be interesting to conduct analogous investigations within the context of interactive retrieval, leveraging an evaluation server and content-based media retrieval system such as DRES and vitrivr.

Another avenue for future research is to explore alternative methods of task presentation. The video filtering pipelines used in this article operated on a comparatively low level. A range of other automatic video manipulation mechanisms exist, e.g., video style transfer [27], which might be worth investigating in the future. While this article concentrated on automated approaches to representation synthesis, numerous manual methods exist that could be considered, such as engaging artists to sketch, reenact, or animate a target video from memory. These representations, grounded in an individual's human interpretation and memory of a video, might yield more informative hints for video retrieval and offer a closer approximation to real memory effects.

Diverging from the visual domain, another promising area for exploration could revolve around audio, which was intentionally excluded from this study to maintain a focused scope. Although audio is not as predominant in memory as visuals, its impact on the video retrieval process could be investigated. Possibilities include applying a basic band-pass filter, experimenting with noise addition, and delving into the memorability of sound. Exploration of the latter could involve identifying and diminishing less memorable audio features while amplifying the more memorable ones, such as speech [9]. In general, one might take different high-level aspects of sound memorability into consideration, such as sound source clarity, emotional valence, and familiarity [53]. Examining visual-audio interactions [49], such as visualized sound cues and sources, adds another layer of complexity to this investigation.

8 Conclusions

This article presented an exploratory investigation into the integration of perceptual and memory effects within interactive video retrieval tasks. Concepts from the literature were applied through the design and implementation of video processing pipelines, focusing on visuals and automated approaches. Six different pipelines emerged from this process: three of which filter the original target segment in various ways, and three synthesize new videos from the original input using generative models. A focus was placed on memory effects, as they also implicitly encapsulate influences of perception and attention.

We then created a custom experimental platform and used it to conduct a crowdsourced KIS evaluation that employed specific video variations, which were generated through the pipelines and selected through a preliminary evaluation. These various task representations have varying effects on the video retrieval process. Still, in general, people can find the target video despite obscuring, removing, or even altering large parts of the video shown as a hint. We found that the way in which task targets are presented impacts overall retrieval performance. While filters aiming to simulate attention and memory effects appear to make the task slightly more difficult, the differences are small compared to using textual or synthetic visual task targets. The findings might have consequences for the design of future benchmarks, as they show the importance of considering additional aspects of real-world scenarios after which the benchmark is modeled.

To the best of our knowledge, the experiment presented in this article was, by a substantial margin, the largest interactive video retrieval evaluation in terms of the number of participants conducted so far.

The extent to which these representations accurately capture human memory effects remains a subject for future investigation. Nevertheless, the results showcased in this study are highly promising. They underscore the significant influence of target representation on retrieval success and demonstrate the malleability of this influence through different approaches. It is highly plausible that real perception and memory effects would introduce nuances in retrieval success rates that are not captured by the current practices of evaluating KIS tasks in media retrieval evaluations. It would be interesting to see a more realistic scenario being emulated in large-scale evaluations and how this might influence evaluation outcomes.

Acknowledgment

The authors would like to thank all the participants in the preliminary study.

References

[1] Richard A. Abrams and Shawn E. Christ. 2003. Motion onset captures attention. *Psychological Science* 14, 5 (Sep. 2003), 427–432.

[2] George A. Alvarez and Brian J. Scholl. 2005. How does attention select and track spatially extended objects? New effects of attentional concentration and amplification. *Journal of Experimental Psychology: General* 134, 4 (2005), 461–476.

[3] Alan D. Baddeley. 1999. *Essentials of Human Memory*. Psychology Press.

[4] James Bigelow and Amy Poremba. 2014. Achilles' Ear? Inferior human short-term and recognition memory in the auditory modality. *PLOS One* 9, 2 (Feb. 2014), e89914.

[5] Timothy F. Brady, Talia Konkle, George A. Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences* 105, 38 (Sep. 2008), 14325–14329.

[6] Bora Celikkale, Aykut Erdem, and Erkut Erdem. 2013. Visual attention-driven spatial pooling for image memorability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR Workshops '13)*. IEEE Computer Society, 976–983.

[7] Marvin M. Chun and Nicholas B. Turk-Browne. 2007. Interactions between attention and memory. *Current Opinion in Neurobiology* 17, 2 (Apr. 2007), 177–184.

[8] Paul D. Clough and Mark Sanderson. 2003. The CLEF 2003 cross language image retrieval track. In *Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum (CLEF '03), Lecture Notes in Computer Science, Vol. 3237*, Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck (Eds.) Springer, 581–593.

[9] Michael A. Cohen, Todd S. Horowitz, and Jeremy M. Wolfe. 2009. Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences* 106, 14 (Apr. 2009), 6008–6010.

[10] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, and Martin Engilberge. 2019. VideoMem: Constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV '19)*. IEEE, 2531–2540.

[11] Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, and Claire-Hélène Demarty. 2018. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18)*. ACM, 178–186.

[12] Mihai Gabriel Constantin and Bogdan Ionescu. 2022. AIMultimediaLab at MediaEval 2022: Predicting media memorability using video vision transformers and augmented memorable moments. In *Working Notes Proceedings of the MediaEval 2022 Workshop*, *CEUR Workshop Proceedings*, Vol. 3583. Bergen, Norway and Online, 12-13 January 2023. Retrieved from https://ceur-ws.org/Vol-3583/paper6.pdf

[13] Rose A. Cooper, Elizabeth A. Kensinger, and Maureen Ritchey. 2019. Memories fade: The relationship between memory vividness and remembered visual salience. *Psychological Science* 30, 5 (Mar. 2019), 657–668.

[14] Rachit Dubey, Joshua C. Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. 2015. What makes an object memorable? In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV '15)*. IEEE Computer Society, 1089–1097.

[15] Théo Dumont, Juan Segundo Hevia, and Camilo Luciano Fosco. 2023. Modular memorability: Tiered representations for video memorability prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '23)*. IEEE, 10751–10760.

[16] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. AMNet: Memorability estimation with attention. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '18)*. Computer Vision Foundation/IEEE Computer Society, 6363–6372. Retrieved from http://openaccess.thecvf.com/content_cvpr_2018/html/Fajtl_AMNet_Memorability_Estimation_CVPR_2018_paper.html

[17] Mengjuan Fei, Wei Jiang, and Weijie Mao. 2017. Memorable and rich video summarization. *Journal of Visual Communication and Image Representation* 42 (2017), 207–217.

[18] Mengjuan Fei, Wei Jiang, and Weijie Mao. 2018. Creating memorable video summaries that satisfy the user's intention for taking the videos. *Neurocomputing* 275 (2018), 1911–1920.

[19] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.

[20] Jason M. Gold, Richard F. Murray, Allison B. Sekuler, Patrick J. Bennett, and Robert Sekuler. 2005. Visual memory decay is deterministic. *Psychological Science* 16, 10 (Oct. 2005), 769–774.

[21] Camille Guinaudeau and Andreu Girbau Xalabarder. 2022. Textual analysis for video memorability prediction. In *Working Notes Proceedings of the MediaEval 2022 Workshop, CEUR Workshop Proceedings*, Vol. 3583 (2022). Retrieved from https://ceur-ws.org/Vol-3583/paper16.pdf

[22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2023. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv:2307.04725. DOI: https://doi.org/10.48550/arXiv.2307.04725

[23] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. 2015. Learning computational models of video memorability from fMRI brain imaging. *IEEE Transactions on Cybernetics* 45, 8 (2015), 1692–1703.

[24] Donna Harman. 1992. Overview of the first text retrieval conference (TREC-1). In *Proceedings of the 1st Text REtrieval Conference (TREC '92)*. National Institute of Standards and Technology (NIST), 1–20. Retrieved from http://trec.nist.gov/pubs/trec1/papers/01.txt

[25] Christopher G. Healey and James T. Enns. 2012. Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics* 18, 7 (2012), 1170–1188.

[26] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoc, Andreas Leibetseder, Frantisek Mejzlík, Ladislav Peska, Luca Rossetto, et al. 2022. *Interactive video retrieval evaluation at a distance: Comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. International Journal of Multimedia Information Retrieval* 11, 1 (2022), 1–18.

[27] Nisha Huang, Yuxin Zhang, and Weiming Dong. 2024. Style-a-video: Agile diffusion for arbitrary text-based video style transfer. *IEEE Signal Processing Letters* 31 (2024), 1494–1498.

[28] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1469–1482.

[29] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*. IEEE Computer Society, 145–152.

[30] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. 2018. DeepVS: A deep learning based video saliency prediction approach. In *Proceedings of the 15th European Conference on Computer Vision (ECCV '18), Lecture Notes in Computer Science, Vol. 11218*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.) Springer, 625–642.

[31] Akankshya Kar, Prashasthi Mavin, Yogesh Ghaturle, and VaniM. 2017. What makes a video memorable?. In *Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA '17)*. IEEE, 373–381.

[32] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators. arXiv:2303.13439. Retrieved from https://arxiv.org/abs/2303.13439

[33] Meera Thapar Khanna, Chetan Ralekar, Anurika Goel, Santanu Chaudhury, and Brejesh Lall. 2019. Memorability-based image compression. *IET Image Processing* 13, 9 (2019), 1490–1501.

[34] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV '15)*. IEEE Computer Society, 2390–2398.

[35] Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2012. Memorability of Image Regions. Advances in neural information processing systems 25, 305–313. Retrieved from https://proceedings.neurips.cc/paper/2012/hash/e9dae45ec08b498f7e1af247757c9b35-Abstract.html

[36] Ricardo Kleinlein, Cristina Luna Jiménez, Zoraida Callejas, and Fernando Fernández Martínez. 2020. Predicting media memorability from a multimodal late fusion of self-attention and LSTM models. In *Working Notes Proceedings of the MediaEval 2020 Workshop*, CEUR Workshop Proceedings, Vol. 2882, Steven Hicks, Debesh Jha, Konstantin Pogorelov, Alba Garcia Seco de Herrera, Dmitry Bogdanov, Pierre-Etienne Martin, Stelios Andreadis, Minh-Son Dao, Zhuoran Liu, Jose Vargas Quiros, Benjamin Kille, and Martha A. Larson (Eds.) CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-2882/paper61.pdf

[37] Ricardo Kleinlein, Cristina Luna-Jiménez, David Arias-Cuadrado, Javier Ferreiros, and Fernando Fernández-Martínez. 2021. Topic-oriented text features can match visual deep models of video memorability. *Applied Sciences* 11, 16 (Aug. 2021), 7406.

[38] Talia Konkle, Timothy F. Brady, George A. Alvarez, and Aude Oliva. 2010. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General* 139, 3 (Aug. 2010), 558–578.

[39] Talia Konkle, Timothy F. Brady, George A. Alvarez, and Aude Oliva. 2010. Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science* 21, 11 (Oct. 2010), 1551–1556.

[40] Jin Ha Lee, Allen Renear, and Linda C. Smith. 2006. Known-item search: Variations on a concept. *Information Realities: Shaping the Digital Future for All - Proceedings of the 69th ASIS & T Annual Meeting (ASIST 2006)*. Austin, TX, USA, November 3-8, 2006 , 1–17.

[41] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. SwinBERT: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '22)*. IEEE, 17928–17937.

[42] Elizabeth F. Loftus and Jacqueline E. Pickrell. 1995. The formation of false memories, *Psychiatric Annals* 25, 12 720–725.

[43] Jakub Lokoc, Werner Bailer, Kai Uwe Barthel, Cathal Gurrin, Silvan Heller, Björn Þór Jónsson, Ladislav Peska, Luca Rossetto, Klaus Schoeffmann, Lucia Vadicamo, et al. 2022. A task category space for user-centric comparative multimedia search evaluations. In *Proceedings of the 28th International Conference on MultiMedia Modeling (MMM '22), Lecture Notes in Computer Science, Vol. 13141*, Björn Thor Jonsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Huynh Thi Thanh Binh, Benoit Huet (Eds.) Springer, 193–204.

[44] Muxuan Lyu, Kyoung Whan Choe, Omid Kardan, Hiroki P. Kotabe, John M. Henderson, and Marc G. Berman. 2020. Overt attentional correlates of memorability of scene images and their relationships to scene semantics. *Journal of Vision* 20, 9 (Sep. 2020), 2.

[45] Matei Mancas and Olivier Le Meur. 2013. Memorability of natural scenes: The role of attention. In *Proceedings of the IEEE International Conference on Image Processing (ICIP '13)*. IEEE, 196–200.

[46] Jean M. Mandler and Gary H. Ritchey. 1977. Long-term memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory* 3, 4 (Jul. 1977), 386–396.

[47] Viorica Marian, Sayuri Hayakawa, and Scott R. Schroeder. 2021. Cross-modal interaction between auditory and visual input impacts memory retrieval. *Frontiers in Neuroscience* 15 (Jul. 2021), 661477.

[48] Hauke S. Meyerhoff and Markus Huff. 2015. Semantic congruency but not temporal synchrony enhances long-term memory performance for audio-visual scenes. *Memory & Cognition* 44, 3 (Nov. 2015), 390–402.

[49] Hauke S. Meyerhoff, Oliver Jaggy, Frank Papenmeier, and Markus Huff. 2022. Long-term memory representations for audio-visual scenes. *Memory & Cognition* 51, 2 (Sep. 2022), 349–370.

[50] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry A. McNamara, and Aude Oliva. 2020. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *Proceedings of the 16th European Conference on Computer Vision (ECCV '20)*. Springer, 223–240.

[51] Henrik Olsson and Leo Poom. 2005. Visual memory needs categories. *Proceedings of the National Academy of Sciences* 102, 24 (Jun. 2005), 8776–8780.

[52] W. A. Phillips and D. F. M. Christie. 1977. Components of visual memory. *Quarterly Journal of Experimental Psychology* 29, 1 (Feb. 1977), 117–133.

[53] David B. Ramsay, Ishwarya Ananthabhotla, and Joseph A. Paradiso. 2018. The intrinsic memorability of everyday sounds. arXiv:1811.07082. Retrieved from http://arxiv.org/abs/1811.07082

[54] Alison Reboud, Ismail Harrando, Jorma Laaksonen, and Raphaël Troncy. 2020. Predicting media memorability with audio, video, and text representations. *In Working Notes Proceedings of the MediaEval 2020 Workshop, CEUR Workshop Proceedings, Vol. 2882*, Steven Hicks, Debesh Jha, Konstantin Pogorelov, Alba Garcia Seco de Herrera, Dmitry Bogdanov, Pierre-Etienne Martin, Stelios Andreadis, Minh-Son Dao, Zhuoran Liu, Jose Vargas Quiros, Benjamin Kille, and Martha A. Larson (Eds.) CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-2882/paper57.pdf

[55] Alison Reboud, Ismail Harrando, Jorma Laaksonen, and Raphaël Troncy. 2021. Exploring multimodality, perplexity and explainability for memorability prediction. *In Working Notes Proceedings of the MediaEval 2021 Workshop*, *CEUR Workshop Proceedings, Vol. 3181*, Steven Hicks, Konstantin Pogorelov, Andreas Lommatzsch, Alba Garcia Seco de Herrera, Pierre-Etienne Martin, Syed Zohaib Hassan, Alastair Porter, Asem Kasem, Stelios Andreadis, Mathias Lux, Marc Gallofre Ocana, Alex Liu, and Martha A. Larson (Eds.), CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-3181/paper53.pdf

[56] Ronald A. Rensink, J. Kevin O'Regan, and James J. Clark. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* 8, 5 (Sep. 1997), 368–373.

[57] Luca Rossetto, Werner Bailer, and Abraham Bernstein. 2021. Considering human perception and memory in interactive multimedia retrieval evaluations. In *27th International Conference on MultiMedia Modeling (MMM '21), Lecture Notes in Computer Science, Vol. 12572*, Jakub Lokoc, Tomas Skopal, Klaus Schoeffmann, Vasileios Mezaris, Xirong Li, Stefanos Vrochidis, and Ioannis Patras (Eds.) Springer, 605–616.

[58] Luca Rossetto, Ralph Gasser, Jakub Lokoc, Werner Bailer, Klaus Schoeffmann, Bernd Münzer, Tomás Soucek, Phuong Anh Nguyen, Paolo Bolettieri, Andreas Leibetseder, et al. 2021. Interactive video retrieval in the age of deep learning - detailed evaluation of VBS 2019. *IEEE Transactions on Multimedia* 23 (2021), 243–256.

[59] Luca Rossetto, Ralph Gasser, Loris Sauter, Abraham Bernstein, and Heiko Schuldt. 2021. A system for interactive multimedia retrieval evaluations. In *27th International Conference on MultiMedia Modeling (MMM '21), Lecture Notes in Computer Science, Vol. 12573*, Jakub Lokoc, Tomas Skopal, Klaus Schoeffmann, Vasileios Mezaris, Xirong Li, Stefanos Vrochidis, and Ioannis Patras (Eds.), Springer, 385–390.

[60] Luca Rossetto, Ivan Giangreco, Claudiu Tanase, and Heiko Schuldt. 2016. vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. In *Proceedings of the 2016 ACM Conference on Multimedia Conference (MM '16)*. ACM, 1183–1186.

[61] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A. Butt. 2019. V3C - A research video collection. In *Proceedings of the 25th International Conference on MultiMedia Modeling (MMM '19), Lecture Notes in Computer Science, Vol. 11295*, Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis (Eds.) Springer, 349–360.

[62] Loris Sauter, Ralph Gasser, Abraham Bernstein, Heiko Schuldt, and Luca Rossetto. 2022. An asynchronous scheme for the distributed evaluation of interactive multimedia retrieval. In *Proceedings of the 2nd International Workshop on Interactive Multimedia Retrieval (MM '22)*. ACM.

[63] Loris Sauter, Ralph Gasser, Heiko Schuldt, Abraham Bernstein, and Luca Rossetto. 2024. Performance evaluation in multimedia retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* (Oct. 2024). DOI: https://doi.org/10.1145/3678881

[64] Loris Sauter, Heiko Schuldt, Raphael Waltenspül, and Luca Rossetto. 2023. Novice-friendly text-based video search with vitrivr. In *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing (CBMI '23)*. ACM, 163–167.

[65] Klaus Schoeffmann. 2019. Video browser showdown 2012-2019: A review. In *Proceedings of the 2019 International Conference on Content-Based Multimedia Indexing (CBMI '19)*. IEEE, 1–4.

[66] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and recall: Learning what makes videos memorable. In *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCV Workshops '17)*. IEEE Computer Society, 2730–2739.

[67] Daniel J. Simons and Christopher F. Chabris. 1999. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception* 28, 9 (Sep. 1999), 1059–1074.

[68] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2023. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, (CLR 2023)*. Kigali, Rwanda, May 1-5, 2023. Retrieved from https://openreview.net/pdf?id=nJfylDvgzlq

[69] Alan F. Smeaton, Paul Over, and R. Taban. 2001. The TREC-2001 video track report. In *Proceedings of the 10th Text REtrieval Conference (TREC '01), NIST Special Publication, Vol. 500-250*, Ellen M. Voorhees and Donna K. Harman (Eds.), NIST. Retrieved from http://trec.nist.gov/pubs/trec10/papers/TREC10Video_Proc_Report.pdf

[70] Tomás Soucek and Jakub Lokoc. 2020. TransNet V2: An effective deep network architecture for fast shot transition detection. arXiv:2008.04838. Retrieved from https://arxiv.org/abs/2008.04838

[71] Lorin Sweeney, Mihai Gabriel Constantin, Claire-Hélène Demarty, Camilo Fosco, Alba Garcia Seco de Herrera, Sebastian Halder, Graham Healy, Bogdan Ionescu, Ana Matran-Fernandez, et al. 2022. Overview of the MediaEval 2022 predicting video memorability task. in *Working Notes Proceedings of the MediaEval 2022 Workshop*, CEUR Workshop Proceedings, Vol 3583. Bergen, Norway and Online, 12-13 January 2023. Retrieved from https://ceur-ws.org/Vol-3583/paper17.pdf

[72] Lorin Sweeney, Graham Healy, and Alan F. Smeaton. 2020. Leveraging audio gestalt to predict media memorability. In *Working Notes Proceedings of the MediaEval 2020 Workshop*, CEUR Workshop Proceedings Vol. 2882. Online, 14-15 December 2020, . Retrieved from https://ceur-ws.org/Vol-2882/paper43.pdf

[73] Lorin Sweeney, Graham Healy, and Alan F. Smeaton. 2021. The influence of audio on video memorability with an audio gestalt regulated video memorability system. In *Proceedings of the 18th International Conference on Content-Based Multimedia Indexing (CBMI '21)*. IEEE, 1–6.

[74] Lorin Sweeney, Graham Healy, and Alan F. Smeaton. 2021. Predicting media memorability: Comparing visual, textual, and auditory features. In *Working Notes Proceedings of the MediaEval 2021 Workshop*, CEUR Workshop Proceedings Vol. 3181. Online, 13-15 Dec 2021. . Retrieved from https://ceur-ws.org/Vol-3181/paper35.pdf

[75] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2023. Phenaki: Variable length video generation from open domain textual descriptions. *The Eleventh International Conference on Learning Representations, ICLR 2023*. Kigali, Rwanda, May 1-5, 2023. Retrieved from https://openreview.net/pdf?id=vOEXS39nOF

[76] Wulin Wang, Jiande Sun, Jing Li, Qiang Wu, and Ju Liu. 2015. Investigation on the Influence of visual attention on image memorability. In *8th International Conference on Image and Graphics (ICIG '15), Lecture Notes in Computer Science, Vol. 9219*, Yu-Jin Zhang (Eds.) Springer, 573–582.

[77] István Winkler and Nelson Cowan. 2005. From sensory to long-term memory: Evidence from auditory memory reactivation studies. *Experimental Psychology* 52, 1 (Jan. 2005), 3–20.

[78] Chia-Chien Wu, Farahnaz Ahmed Wick, and Marc Pomplun. 2014. Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology* 5 (2014) 54.

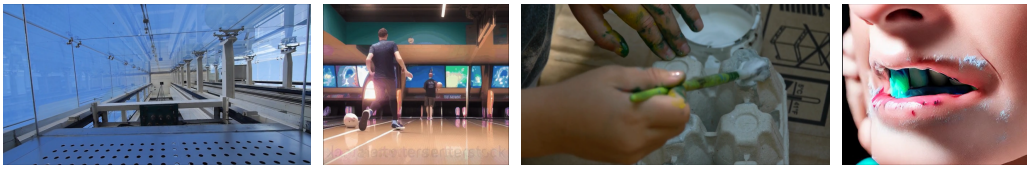
[79] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. 2023. NUWA-XL: Diffusion over diffusion for extremely long video generation, In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, V*ol 1: Long Papers), ACL 2023. Toronto, Canada, 9-14 Jul 2023 1309–1320.

[80] Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. arXiv:2302.05543. Retrieved from https://arxiv.org/abs/2302.05543

[81] Jacqueline F. Zimmermann, Morris Moscovitch, and Claude Alain. 2016. Attending to auditory memory. *Brain Research* 1640 (Jun. 2016), 208–221.

Appendix

A Additional Figures and Tables



(a) Frame from video 06791 (b) S2 Variation (c) Frame from video 13062 (d) S3 Variation

Fig. A1. Examples of eliminated synthesis outputs.

Table A1. Textual Hints

ID	Description
01140	Start of an indoor bike race with six riders. A motorbike with a camera crosses the start line just after the starting shot.
02024	Singing instruction video, showing two singers and a keyboarder, with an overlaid graphical visualization.
05722	Shot of a wedding party panning from left to right, the party is grouped around bride and groom, then a shot of bride and groom walking and guests following them.
13872	Kids in kayaks on a river, throwing paddles through three colored hoops placed over the water.
14700	View down the surface of a boulder, with a forest in the background. A bearded man in a cyan shirt climbing up the boulder.

Table A2. Submission Accuracies per Task

Video	Pipeline	Correct	Within 30 Seconds	Within 1 Minute	Within Video	Wrong
01140	Original	83.3%	0%	4.2%	8.3%	4.2%
	F2	32.4%	1.5%	0%	0%	66.2%
	F3	39.2%	7.8%	0%	0%	52.9%
	S	6.5%	0%	0%	16.1%	77.4%
	Text	6.5%	15.2%	0%	2.2%	76.1%
02024	Original	28.4%	7.8%	4.9%	46.1%	12.7%
	F2	48.8%	7.3%	0%	31.7%	12.2%
	F3	14.4%	0.8%	2.4%	62.4%	20.0%
	S	0%	0%	0%	0%	100%
	Text	20.0%	1.7%	0%	53.3%	25.0%
05722	Original	19.6%	7.2%	0%	12.4%	60.8%
	F2	22.9%	13.5%	8.3%	15.6%	39.6%
	F3	55.0%	10.0%	5.0%	17.5%	12.5%
	S	4.5%	0%	2.3%	4.5%	88.6%
	Text	0%	0%	0%	0%	100%
13872	Original	31.0%	56.0%	6.0%	2.4%	4.8%
	F2	19.8%	22.1%	4.7%	3.5%	50.0%
	F3	36.8%	35.5%	5.3%	7.9%	14.5%
	S	4.2%	41.7%	25.0%	4.2%	25.0%
	Text	8.7%	35.6%	24.0%	4.8%	26.9%
14700	Original	17.8%	31.8%	10.9%	13.2%	26.4%
	F2	18.5%	18.5%	1.9%	16.7%	44.4%
	F3	9.6%	36.1%	19.3%	10.2%	24.7%
	S	0.9%	25.0%	16.7%	42.6%	14.8%
	Text	5.6%	4.0.%	15.6%	26.9%	11.9%

Received 18 March 2024; revised 10 March 2025; accepted 26 March 2025