Seeing Eye to AI:
Comparing Human Gaze and Model Attention in Video Memorability

Prajneya Kumar*[1] Eshika Khandelwal*[2] Makarand Tapaswi†[2] Vishnu Sreekumar†[1]
{[1]Cognitive Science Lab, [2]CVIT}, IIIT Hyderabad
https://katha-ai.github.io/projects/video-memorability/
*† equal contribution

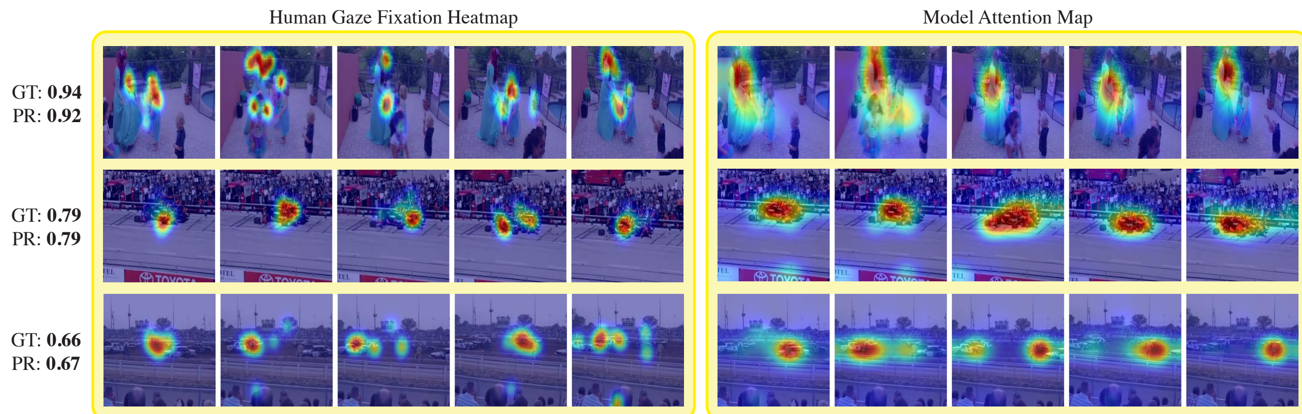

Figure 1. Comparing human gaze fixations (left) and model's attention maps (right) for 3 different videos (one per row). The memorability scores, ground-truth (GT) and model prediction (PR), are provided on the left. The heatmaps depict areas of high visual attention through warmer colors (red-yellow), indicating regions where human observers fixated (left) and model attended (right). The model's attention patterns are aligned with human gaze patterns, especially for more memorable videos. Samples from Memento10k [33].

Abstract

Understanding what makes a video memorable has important applications in advertising and education technology. Towards this goal, we investigate spatio-temporal attention mechanisms underlying video memorability. Different from previous works that fuse multiple features, we adopt a simple CNN+Transformer architecture that enables analysis of spatio-temporal attention *while matching state-of-the-art (SoTA) performance on video memorability prediction. We compare model attention against human gaze fixations collected through a small-scale eye-tracking study where humans perform the video memory task. We uncover the following insights: (i) Quantitative saliency metrics show that our model, trained only to predict a memorability score, exhibits similar spatial attention patterns to human gaze, especially for more memorable videos. (ii) The model assigns greater importance to initial frames in a video, mimicking human attention patterns. (iii) Panoptic segmentation reveals that both (model and humans) assign a greater share of attention to things and less attention to stuff as compared to their occurrence probability.*

1. Introduction

In 2018, Nike's "Dream Crazy" commercial featuring Colin Kaepernick captured nationwide attention in the US[1]. This advertisement was especially memorable because it was aired in the aftermath of Kaepernick's protests against race-based police brutality. While the context made this commercial memorable for US-based audiences, other types of commercials tend to be memorable in general. For example, a famous 2013 E-Trade Super Bowl commercial features a baby seated behind a stack of cash talking about investments and hidden fees[2]. This sort of ad is likely to be memorable regardless of cultural context due to several attention-grabbing features, notably, a baby talking in an adult voice and delivering investment advice. This latter type of memorability, thought to be consistent across individuals and cultures, has been extensively studied in both cognitive science and computer vision using images [4, 21] and words [1, 31]. In this work, we ask: what are the spatial, temporal, and semantic patterns of attention that are associ-

[1]Dream Crazy https://www.youtube.com/watch?v=WW2yKSt2C_A
[2]E-Trade ad https://www.youtube.com/watch?v=EbnWbdR9wSY

ated with video memorability? To answer this question, we train a CNN+Transformer model to predict human memorability of naturalistic videos, use self-attention scores to determine where the model *looks* across space and time, and collect human eye-tracking data to compare the model's attention against human fixations (Fig. 1).

Early work on image memorability reveals the importance of both object and scene categories in predicting memorability [15,21]. Semantic categories are also predictive of memorability across stimuli, including words [1,31] and indeed, prior work shows that context guides eye movements to task-relevant object locations [46]. Thus, we investigate what semantic categories in videos drive memorability. Video captioning approaches have been used in previous semantic analyses of video memorability [13, 33, 39]. However, to our knowledge, we are the first to present a detailed analysis of attention captured by different semantic categories when humans attempt to memorize videos and when a model is trained to predict these memorability scores. We apply panoptic segmentation [11] and adopt the COCO hierarchy [10] to distinguish between *things* (*i.e*. objects with well-defined shapes such as *person*) and *stuff* (*i.e*. amorphous background regions such as *sky*) in the video frames. Next, we compare pixel distributions weighted by model attention and human gaze and find that both the model and humans generally enhance attention to *things* and reduce attention to *stuff*. Furthermore, the model and humans agree on what specific *things* and *stuff* to emphasize or disregard. Overall, these results indicate that the model learns similar attentional strategies as humans *even though it is trained only to predict a memorability score*.

Beyond semantics, the time axis in videos begs an important question: how early does the model know about the memorability of a video? Human experiments using extremely fast presentation times reveal that image memorability differences can be observed in brain activity patterns as early as 400 ms [4, 24]. Therefore, it is possible that very early moments in a video are predictive of how memorable it will be. Furthermore, human attention tends to be highest at the beginning of an event and wanes over the course of the event [27]. Thus, video memorability scores may be influenced to a greater extent by the initial frames. Note that memorability scores are computed as a consensus across participants. Therefore, we expect the video frames that most people attend to in similar ways to drive the memorability scores. Despite having no intrinsic temporal bias, can models trained to predict memorability pick up on these human-like temporal attention patterns? To answer this question, we first analyze human-human gaze agreement in our videos and establish that different people are more likely to attend to similar regions in the initial frames. Next, summing over the model's spatial attention scores in a frame, we observe that the model indeed assigns greater importance to earlier frames within videos, thereby discovering a subtle temporal pattern in human behavior.

The video memorability literature [12,16,20] focuses on high prediction performance and lacks analysis of models' (dis)similarities to how humans view and remember videos. We address this gap through the following contributions: (i) We adopt a simple CNN+Transformer model to predict video memorability as it facilitates a study of spatio-temporal attention mechanisms. Even with a single encoder, our model matches state-of-the-art performance. (ii) To compare the model against *what* humans look at and *when*, we collect eye-tracking data of subjects in a video memorability experiment, similar to the original setup [12, 33]. (iii) Through panoptic segmentation and attention-weighted analyses, we show that both the model and humans increase and decrease attention similarly to different *things* and *stuff*. (iv) We show that our model with no intrinsic temporal bias learns to attend to the initial frames of the video with a decreasing pattern over time, consistent with framewise human-human gaze agreement patterns. We will release our code and eye-tracking data to encourage further research.

Note, our work aims to highlight the similarities between *human fixations* when performing memorability experiments, and *model attention* when trained to predict memorability scores. A simple CNN+Transformer architecture enables this, matches SoTA, and has not been used in video memorability before.

2. Related Work

Memorability in cognitive science. While human beings remember a huge amount of visual information, not all visual experiences are equal in our memory [21]. Some images are consistently better remembered across people, suggesting that memorability is observer-independent [3, 4]. This makes algorithms suitable for predicting memorability [25]. Several factors such as scene semantics [21], object category [15], and visual saliency [15] correlate with memorability, yet considerable statistical variance in memorability scores remains unexplained [38]. Although image memorability has been studied extensively in cognitive science, videos have been used primarily in the study of event segmentation and to understand the neural processes underlying learning and memory [5, 6]. Observer-independent memorability of videos has received less attention in cognitive science compared to the work in computer vision.

Memorability in computer vision. The study of visual memorability in computer vision started with a focus on images [21, 25]. Models such as *MemNet* were developed for image memorability prediction on large image datasets [25]. Improvements over the initial models involved incorporating attention mechanisms [17], image captioning modules [41], object and scene semantics [35],

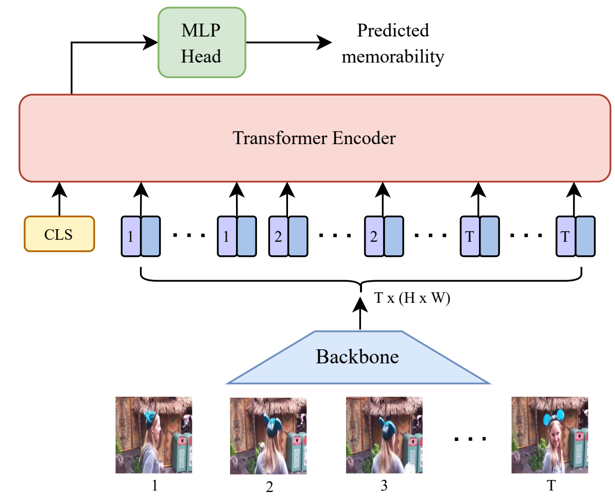

Figure 2. **Model overview.** T video frames are passed through an image backbone encoder to obtain spatio-temporal features $\mathbf{F} \in \mathbb{R}^{T \times H \times W \times D}$. Coupled with position embeddings, and after appending a CLS token, we pass them through a Transformer encoder with self-attention. A memorability score is calculated at the CLS representation with an MLP. *Attention scores between* CLS *and each token are used for downstream analysis.*

and aesthetic attributes [50]. The insights gained from these studies also led to the development of Generative Adversarial Networks (GAN) based models that can modify images to manipulate their memorability [18, 28, 40].

Video memorability has fewer works, typically evaluated on *VideoMem* [12] and *Memento10k* [33]. The semantic embeddings model of *VideoMem* [12] uses an image-captioning pipeline in conjunction with a 2-layer MLP for memorability prediction. *SemanticMemNet* [33] integrates visual cues with semantic information and decay patterns to predict memorability. Recent approaches involve multiple tiered representation structures, *M3S* [16], or use Large Language Models (LLMs) to generate textual descriptions that are then used to predict memorability scores [20]. In contrast, we adopt a simple CNN+Transformer attention-based model that matches SoTA, but also facilitates comparison between model attention and human gaze on semantic and temporal aspects of video memorability.

3. Methods: Model and Human

We present two methods: (i) a CNN+Transformer model that predicts memorability scores using spatio-temporal attention; and (ii) an eye-tracking study to capture human gaze patterns during a memorability experiment.

3.1. Transformer-based Model

We begin by defining some notation. Our dataset consists of multiple videos with associated memorability scores, (V, m) pairs. Each video consists of multiple frames. We sub-sample T frames for memorability prediction and denote a video as $V = \{f_i\}_{i=1}^{T}$.

Our model consists of three parts: (i) a backbone image encoder Φ, (ii) a Transformer encoder that attends over spatio-temporal tokens extracted from T video frames, and (iii) a prediction head that estimates the memorability of a video (see Fig. 2).

1. Image encoder. Our goal is to employ a model that allows us to analyze the spatio-temporal attention over video frames. Thus, we consider CNN backbones such as ResNet-50 [19], trained with contrastive language-image pretraining (CLIP) [36]. We encode each video frame to obtain a space-aware representation (from the conv5 layer):

$$\mathbf{f}_i = \Phi(f_i)\,, \text{ where } \mathbf{f}_i \in \mathbb{R}^{H \times W \times D}, \forall i \in \{1, \dots, T\}\,, \tag{1}$$

where $H \times W$ are height and width of the spatial resolution, and D is the dimensionality of the embeddings.

While previous works use multiple features: frames, flow, and video by [33]; low-, mid-, and high-level representations and a contextual similarity module by [16]; or a host of 10+ models fed to an LLM by [20], our model relies on a single semantic backbone (CLIP). Our simple approach enables the analysis of model's spatio-temporal attention maps through a comparison to human gaze.

2. Video encoder. We use a Transformer encoder [47] to capture attention across spatio-temporal tokens. First, we flatten and encode the image features using a linear layer $\mathbf{W}_d \in \mathbb{R}^{d \times D}$ to reduce dimensionality. Next, to each token, we add two types of position embeddings:

$$\mathbf{f}'_{ij} = \mathbf{W}_d \mathbf{f}_{ij} + \mathbf{E}^t_i + \mathbf{E}^s_j, \forall i \in \{1, \dots, T\}, j \in \{1, \dots, HW\}\,, \tag{2}$$

where $\mathbf{E}^t_i$ is the i^{th} row of the temporal embedding matrix (learnable or Fourier), and $\mathbf{E}^s_j$ is the j^{th} row of the spatial embedding matrix, and $\mathbf{f}_{ij} \in \mathbb{R}^D$ is the feature at frame i and spatial region j.

We prepend a CLS token (with learnable parameters $\mathbf{h}_{\text{CLS}}$) to create a sequence of $1{+}THW$ tokens and post LayerNorm [2] feed this to a Transformer encoder (TE) of L layers with hidden dimension d:

$$[\tilde{\mathbf{h}}_{\text{CLS}}, \tilde{\mathbf{f}}_{11}, \dots, \tilde{\mathbf{f}}_{THW}] = \text{TE}([\mathbf{h}_{\text{CLS}}, \mathbf{f}'_{11}, \dots, \mathbf{f}'_{THW}])\,. \tag{3}$$

3. Predicting memorability. We pass the CLS token's contextualized representation to an MLP and predict the memorability score: $\hat{m} = \text{MLP}(\tilde{\mathbf{h}}_{\text{CLS}})$.

Extracting attention scores. We extract the self-attention matrix from the multi-head attention module of the last layer of the TE. We mean pool over the heads and pick the row corresponding to the CLS token. Ignoring the self token, this attention vector $\alpha \in \mathbb{R}^{THW}, \sum \alpha = 1$, is used for further spatio-temporal analysis. We obtain an attention map of the size of the image by applying upscaling (pyramid expand) on the $H \times W$ attention scores of each frame.

Training and inference. Similar to previous work [16, 33] we use the MSE loss $\mathscr{L} = \|m - \hat{m}\|^2$ to train our model. We also considered the Spearman loss [16], but did not see significant performance gains. For most experiments, we freeze the backbone and rely on the strong semantic features extracted by CLIP pretraining.

3.2. Eyetracking Study: Capturing Gaze Patterns

We collect eye-tracking data while participants view videos in a memory experiment. The setup (schematic in supplement Fig. 9) follows the original video memorability experiments [12, 33], as we want the gaze patterns to accurately reflect the cognitive and visual processes involved in viewing and remembering videos. Further details regarding the setup are provided in supplement Appendix A.1.

Data collection. Our study has 20 participants (9 females, 11 males, Age 22.15 ± 0.52 (mean $\pm$SEM)). Memento10K: 6 females, 4 males, Age 22.9 ± 0.94. VideoMem: 3 females, 7 males, Age 21.4 ± 0.37. We choose 140 unique videos each from both video datasets: *Memento10K* [33] and *Videomem* [12]. We use the SR Research EyeLink 1000 Plus [42] to capture binocular gaze data, sampling pupil position at 500 Hz. A 9-point target grid is used to calibrate the position of the eye. Saccades and fixations are defined using the algorithm supplied by SR Research.

We perform clustering to select videos spanning diverse visual content and memorability attributes (see supplement Appendix A.2 for details). Participants watch multiple videos and are instructed to press the `SPACEBAR` upon identifying a repeated video. Each participant watched a total of 200 videos: 140 unique videos, 20 target repeats occurring at an interval between $9-200$, and 40 vigilance repeats interspersed every $2-3$ videos. All videos are displayed in their original aspect ratios at the center of a white display screen with resolution 1024×768 pixels.

Data processing. The fixation coordinates for both eyes are obtained using the EyeLink Data Viewer software package (SR Research Ltd., version 4.3.210). These coordinates are then used to construct a binary matrix for each participant, corresponding in size to the original video dimensions. To account for the visual angle of approximately 1 degree, a Gaussian blur is applied to these matrices (see supplement Appendix A.3 for details). To create the human fixation density maps, we average the matrices corresponding to the same frame of the same video across participants. To ensure compatibility with model's attention maps, the fixation maps are resized to a resolution of 224×224 pixels.

4. Experiments

Video memorability datasets. We perform experiments on two datasets: (i) **VideoMem** [12] consists of 10K, 7 second video clips, each associated with a memorability score. (ii) **Memento10K** [33], introduced as a dynamic video memorability dataset, contains human annotations at different viewing delays. This dataset consists of 10K clips, but they are shorter in duration (3 seconds).

Data splits. VideoMem has 7000 videos in the training set and 1000 in the validation set (MediaEval workshop [43]). Past works report results on the validation set as the test labels are not publicly available. Memento10k is split into 7000 videos for train and 1500 each for validation and test. We provided our model's outputs to the competition organizers and report results on the test set.

Memorability metrics. The memorability score associated with each video in the datasets captures the proportion of people in the original experiments who correctly recognized the video. We evaluate model's predictions relative to ground-truth (GT) memorability scores, using the Spearman rank correlation (RC $\uparrow$). Following previous works, we also report the mean squared error (MSE $\downarrow$) to measure the gap between GT and predictions.

Implementation details. We break each video into T uniform segments and pick one frame at random from each segment during training - this acts as data augmentation [49]. For inference, we take the middle frame of the segment. $T{=}5$ works well for Memento10k (1.66fps) and $T{=}7$ for VideoMem (1fps). When not specified otherwise, we train our model with the Adam optimizer [26], learning rate 10^{-5}, and a step scheduler (for VideoMem only) with step size 10 epochs and multiplier 0.5.

4.1. Video Memorability Prediction

We begin with model ablation studies for Memento10k. VideoMem has some challenges with respect to data leakage (Sec. 4.2) and results are presented in Appendix C.1.

Ablation of vision models. Tab. 1 rows 1-6 show the results of various hyperparameters of the vision model evaluated on the validation set. Row 1 (R1) achieves best performance and is the *default configuration* for further experiments. Using spatio-temporal (ST, R1) image embeddings and not performing global average pooling (R2) shows a small improvement in RC. Similarly, using Fourier embeddings (R1) is better than learnable ones (R3), perhaps due to the small dataset size. Surprisingly, using spatial embeddings to identify the $H \times W$ tokens reduces performance (R1 *vs.* R4 or R5), perhaps due to the pyramidal nature of the CNN representations. Finally, using random sampling during training (R1) instead of picking the middle frame of the segment (R6) results in a small increase. In general, the gap between all rows is small, indicating that results are not impacted strongly by hyperparameter changes. However, spatio-temporal (ST) CLIP embeddings are required to obtain spatio-temporal model attention maps.

Use of captions. [33] introduced captions (descriptions) for

	Embedding					Memento10k (val)	
	CLIP	Time	Space	Sampling	Caption	RC ↑	MSE ↓
1	ST	F	-	Random	-	**0.706**	0.0061
2	T	F	-	Random	-	0.687	0.0062
3	ST	L	-	Random	-	0.696	0.0059
4	ST	F	1D	Random	-	*0.703*	*0.0057*
5	ST	F	2D	Random	-	0.701	**0.0056**
6	ST	F	-	Middle	-	*0.703*	0.0066
7	ST	F	-	Random	Orig.	**0.745**	**0.0050**
8	ST	F	-	Random	Pred.	*0.710*	*0.0056*

Table 1. **Model ablations.** Column 1 (C1) compares the impact of using spatio-temporal (ST) features versus temporal (T) features with global average pooling. C2 and C3 specify the types of temporal (L: learnable, F: Fourier) and spatial position embeddings used. C4 is the frame sampling method used during training. C5 indicates whether the video caption (Orig: original caption, Pred: predicted caption) is used in modeling. *Row 1 (R1) is chosen as the default configuration for further experiments* and represents the best vision-only model. R2-6 evaluate vision model choices: features, position-encodings, and frame sampling methods. R7 presents results with original captions (Orig.) as a part of the model and R8 aims to predict the captions on the fly. The best results in each section are in **bold**, with second-best in *italics*.

the short videos in Memento10k as a way to emphasize semantic categories for predicting memorability. We modify our model by extending the sequence length of our Transformer encoder to include additional description tokens. Visual and text tokens are differentiated through a type embedding (additional details in the supplement, Appendix D).

In Tab. 1 (bottom) using the original captions (OC) strongly benefits Memento10k as Spearman RC goes up from 0.706 (R1) to 0.745 (R7). However, when the visual tokens predict both the memorability score and the caption (similar to CLIPCap [32]) the memorability score shows modest improvement (to 0.710, R8).

SoTA comparison. Comparison to state-of-the-art works on Memento10k with different setups (val or test split, with / without captions) is presented in Tab. 2. Note, our goal is to understand the attentional factors driving video memorability through a model that provides spatio-temporal attention. Nevertheless, our model with a single feature encoder (CLIP) achieves results comparable to SoTA (Memento10k: 0.706 val, 0.662 test). With captions, we obtain 0.713 (test). To interpret model performance reported as RC scores, we note that a model that performs well is expected to approach a human-human consistency RC of 0.73 for *Memento10K* [33].

Furthermore, our model is trained only on the Memento10k training set, while all baselines train on a combination of image and video memorability datasets. For example, pretraining on LaMem [41] and fine-tuning on Memento10k improves performance from 0.706 to 0.715. For

		Memento10k			
		Test		Val	
Methods	Caption	RC	MSE	RC	MSE
SemanticMemNet ECCV20	No	0.659	-	-	-
M3-S CVPR23	No	-	-	0.670	0.0062
Ours (R1 Tab. 1)	No	**0.662**	0.0065	**0.706**	0.0061
SemanticMemNet ECCV20	Yes	0.663	-	-	-
Sharingan arXiv	Yes	-	-	0.72	-
Ours (R7 Tab. 1)	Yes	**0.713**	0.0050	**0.745**	0.0050

Table 2. Comparison against SoTA for video memorability. Baselines considered are SemanticMemNet [33], M3-S [16], and Sharingan [20]. Split-half human-human consistency RC for Memento10k is 0.73. See supplement Tab. 6 for VideoMem.

completeness, we present cross-domain transfer results of pretraining and fine-tuning our model on image or video memorability datasets and evaluation on all in the supplement, Appendix B.

All further analyses and experiments are conducted using the vision-only model, without incorporating captions.

4.2. Why is VideoMem challenging?

The RC scores on VideoMem [12] are significantly lower than on Memento10k, even with additional information like captions providing no improvement. Detailed results can be found in supplement Appendix C.1. In fact, most methods achieve RC greater than the human-human RC at 0.481, indicating that models have probably overfit to the dataset, especially as a held-out test set is not available. As evidence, the code repository of a recent work, M3-S [16][3] shows that achieving a Spearman RC of 0.5158 is possible after using highly specific random seeds and hyperparameters.

Similar videos across splits. We propose a nearest-neighbors (NN) analysis of representations and observe that improving results on VideoMem is challenging due to problems in *split creation*. We visualize the NN in the training set for each validation video based on $\hat{\mathbf{h}}_{\mathsf{CLS}}$, the representation before the MLP regressor. On VideoMem, from a random sample of 30 validation videos, 14 clips have visually identical NN in the training set. In contrast, on Memento10k, we are only able to find 1 clip among 30.

Fig. 3 displays a few videos illustrating this problem. On Memento10k (left), we see that NNs show semantic awareness and matching (I food, II speaker, IV sports field). On the other hand, on VideoMem, the NN are (probably) from the same long video. See right: I surfer, II astronaut, III news anchor, IV farmer. Given the identical visual stimuli, the model can do no better than predict the average memorability score of the NNs on the training set (which it does). *E.g.* in row II with the astronaut, PR=0.81 is equal to the average memorability, but is away from GT=0.86. In

[3] https://github.com/theodumont/modular-memorability

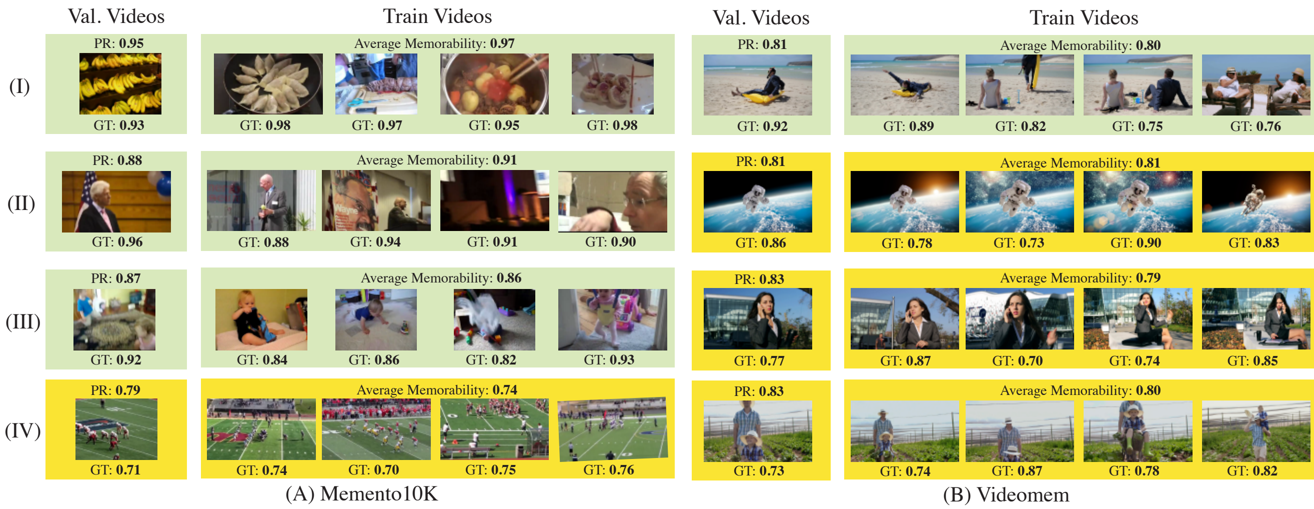

Figure 3. Nearest neighbor (NN) analysis for videos from Memento10K (left) and VideoMem (right). We illustrate four validation set videos and for each, four NN from the training set. We provide the GT memorability score (below), the predicted score on the val set (above), and the average of 4 NN scores from the training set. In B (right), multiple video clips with high visual similarity between train and validation sets are highlighted with a *yellow* background. Conversely, the green rows highlight clips that have similar content, but are likely from different source videos. We discuss how data leakage and variance in GT scores may adversely affect evaluation in Sec. 4.2.

row IV farmer, PR=0.83 is close to the average 0.80, but away from GT=0.73. While using multiple feature backbones may help, this is not a satisfactory solution to a fundamental issue of data leakage across splits. To address this, we attempted to recreate the splits. However, as the original source video ids are unavailable, it is not easy to detect which video clips belong to the source video.

Implications for data collection. We encourage researchers to analyze new datasets before they are released. Information about the video source and split creation process are crucial aspects for any dataset. Additionally, memorability scores are a measure of consensus among viewers and are therefore closely tied to the number of viewers per video. While LaMem averages 80 scores per image, Memento10K has over 90 annotations per video, Videomem averages 38 annotations per video, much smaller than the others. This variance in GT scores is also observed in Fig. 3 (B-II), videos of the same astronaut have GT scores varying from 0.73 to 0.90, making learning difficult.

4.3. Comparing Model Attention and Human Gaze

Setup. To compare the human gaze fixation density maps and model-generated attention maps, we first min-max normalize them to $[0, 1]$. Next, we compute multiple popular metrics [4] in saliency evaluation [9]: AUC-Judd [23], Normalized Scanpath Saliency (NSS) [7], Linear Correlation Coefficient (CC) [34], and Kullback-Leibler Divergence (KLD) [37, 44].

[4]We compute all metrics following the methods used by https://github.com/imatge-upc/saliency-2019-SalBCE/blob/master/src/evaluation/metrics_functions.py

	Memento10k			VideoMem		
Metrics	M-H	H-H	H-H Shuff.	M-H	H-H	H-H Shuff.
AUC-J ↑	0.89 ±0.007	0.90 ±0.001	0.70 ±0.002	0.89 ±0.007	0.80 ±0.002	0.55 ±0.001
AUC-P ↑	82.91 ±1.65	-	-	88.88 ±1.29	-	-
NSS ↑	1.95 ±0.074	3.07 ±0.024	0.84 ±0.022	2.00 ±0.068	3.12 ±0.023	0.23 ±0.012
CC ↑	0.46 ±0.014	0.49 ±0.003	0.16 ±0.003	0.27 ±0.007	0.27 ±0.018	0.03 ±0.001
KLD ↓	1.48 ±0.035	2.17 ±0.023	4.61 ±0.022	2.65 ±0.020	4.02 ±0.018	6.49 ±0.013

Table 3. Comparing gaze fixation maps against model's attention map via different metrics, along with human-human split-half reliability scores over 10 iterations. ↑ (↓) indicates higher (lower) is better. M-H: Model-human; H-H: Human-human; and H-H Shuff.: Human-Human_shuffled (random performance).

We split participants into two random groups and for a given video, compute agreement between the two groups using the saliency metrics. These human-human (H-H) agreement scores are averaged over 10 random split iterations and then across videos. H-H scores act as a ceiling against which our model-human (M-H) agreement scores are compared. To obtain chance-level performance, we compute H-H agreement scores but now with shuffled videos (H-H Shuff.).

Results. While Fig. 1 shows qualitative results of human gaze and model attention, Tab. 3 indicates that there is a high degree of M-H similarity across both datasets. We observe that metrics (AUC-J, CC) often approach the H-H scores, and importantly, significantly improve over random chance (H-H Shuff.). In Fig. 5, we plot AUC-Judd and NSS

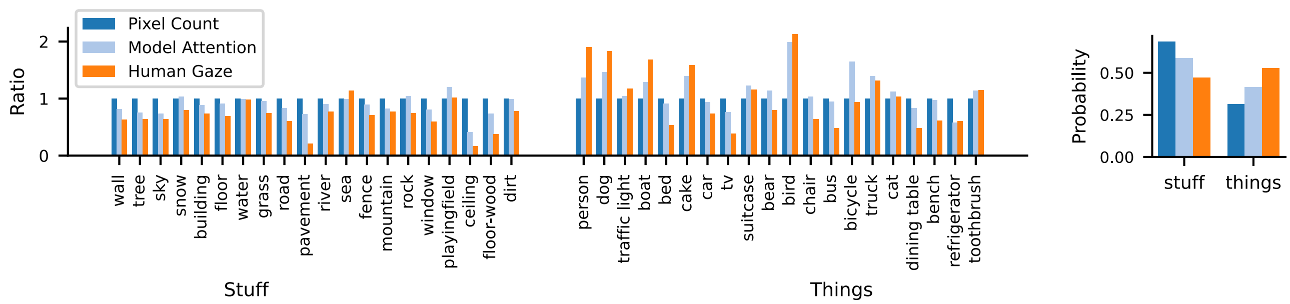

Figure 4. Analysis of panoptic segmentation for the most common 40 classes (20 stuff, 20 things). Left shows normalized pixel counts (blue), model attention-weighted counts (light blue), and human gaze-weighted counts (orange). Both, model and humans, show lower affinity for stuff classes and higher for thing classes, indicating their importance in memorability. Right Pixel counts are accumulated across stuff and thing classes, highlighting the above trend clearly. Best viewed on screen with zoom.

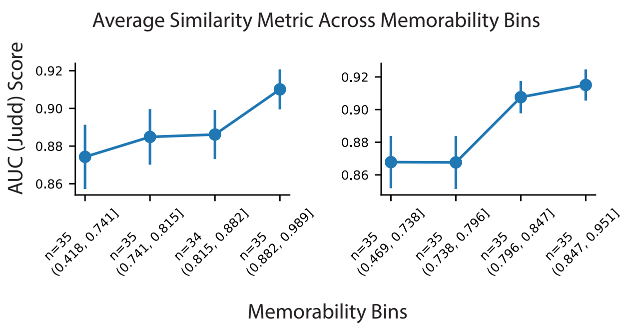

Figure 5. Gaze *vs.* attention similarity metrics with AUC-Judd scores on the Y-axis and Ground Truth on the X-Axis. (See supplement Appendix C.3, Fig. 11 for other metrics and their trends.) Left: Memento10k, Right: VideoMem. Error bars depict SEMs.

against GT memorability bins and observe that the similarity between model attention and human gaze maps increases with GT memorability scores in both datasets. This suggests that highly memorable videos have clear regions of focus for both humans and the model. Please refer to the supplement Appendix C.3 for other metrics.

Furthermore, we replicate these results on image datasets by using a model pretrained on LaMem [25] and fine-tuned on FIGRIM [8] (supplement Appendix C.4).

Center bias. Among metrics, we also considered the shuffled AUC (sAUC) [48], but it tends to unjustly penalize valid central predictions [22]. Therefore, we introduce a metric to measure relative similarity, *AUC-Percentile*. For a given video, we compare the true AUC-Judd between model attention and human gaze against a distribution of AUC-Judd values calculated by comparing model attention from that video and human gaze from other randomly selected videos. The percentile of the true AUC-Judd score within the distribution of random AUC-Judd scores estimates the probability that the true score is video-specific and is not obtained by chance or due to center bias. For instance, a model driven purely by center-bias (using a 2D Gaussian, σ=10% of the scene height [30]) yields an average AUC-Percentile score of 76.17 ± 2.62 on Memento10K and 68.47 ± 2.82 for VideoMem. Results in Tab. 3 show that our model's AUC-P scores at 82.91 ± 1.65 and 88.88 ± 1.29 exceed these center-bias-driven AUC-P scores.

Another approach to rule out the possibility that the high M-H similarity is due to center bias involves a direct comparison between the performance of the previously explained Gaussian-based center bias model [30] and our proposed gaze prediction model. We use the Gaussian to simulate central fixation and calculate median AUC-J score across frames per video. Compared to the Gaussian, our model is better aligned with human fixations across videos on both datasets, Memento10K ($p = 0.003$) and VideoMem ($p = 5.80 \times 10^{-12}$).

4.4. Panoptic Segmentation

We extract panoptic segmentation labels from Mask-Former [11], a SoTA model for segmentation, on the T selected video frames (see supplement Fig. 16 for examples). We use the COCO-stuff hierarchy [10] to classify labels as *stuff* or *things*. We create three sets of counts: (i) *Pixel Count* sums the number of pixels attributed to each label across frames and videos (normalized by the total number of pixels in the frame). (ii) *Model Attention* weighted counts multiply the attention map with segmentation masks of each category, summing across frames and videos. (iii) *Human Gaze* weighted counts are similar and multiply gaze fixation densities with segmentation masks.

Stuff *vs.* things classes. We consider the most prevalent *stuff* and *things* labels (20 each) across the 140 videos of the eye-tracking dataset and observe that attention increases/decreases relative to normalized pixel counts in similar ways for models and humans (Fig. 4 left). Specifically, we observe a tendency for decreased attention to *stuff* and increased attention to *things*, which is clear in the cumulative distributions (Fig. 4 right).

Simple *vs.* complex videos. Panoptic segmentation also allows us to answer a crucial question about the impact of video complexity on model-human alignment. We split our videos into simple and complex based on the number of objects averaged over frames (median split). Comparing model-human and human-human alignment in these videos, we find no significant differences in most metrics (see supplement Appendix C.3) suggesting that our results are not influenced by the complexity of videos.

4.5. Temporal Attention

We first analyze whether humans look at similar regions across frames of a video and find that they are more consistent in the initial frames of the video as compared to later frames, see Fig. 6 (blue). However, it is possible that this result is driven by center bias if most videos have salient central regions at the start. To rule this out, we identify a subset of videos that have off-center salient regions in the initial frames.[5] Fig. 6 (green) shows us that there is stronger consensus across participants for the off-centered videos, and this too goes down as the video progresses.

Next, to ascertain whether our model displays similar temporal patterns of attention, we compute attention scores as $\alpha \in \mathbb{R}^{T \times HW}$ and sum over the spatial dimensions to obtain temporal attention, $\alpha_T \in \mathbb{R}^T$. As visualized in Fig. 7 left, our model preferentially attends to the initial frames of the video sequence, without any architectural bias towards this. We further rule out two possibilities: (i) reversing the frames (and preserving the same temporal position embeddings), we observe that the model still gives more attention to early frames (now appearing at the end, Fig. 7 middle); (ii) computing optical flow magnitude [45] per frame, averaged across all pixels, we find that motion is strongest around the middle (Fig. 7 right) and cannot be the reason for increased attention to early frames.

Therefore, we conclude that our model, only trained to predict memorability scores, has learned to attend to the visual information that most participants look at earlier on in the videos.

5. Conclusion

We adopted a simple CNN+Transformer model that not only matches SoTA in predicting video memorability scores, but also enables exploring the underlying spatio-temporal attention mechanisms. Furthermore, we collected human gaze data to compare against model attention and observed that the model and humans look at similar regions. We also discovered novel semantic attention patterns relevant for video memorability. On the temporal dimension, the model exhibited strong preference for early frames of the videos, mimicking temporal patterns in human attention. We also analyzed a widely used video memorability

[5] We adopt DeepGaze [29] and compute saliency maps for T video frames. Next, we compute a distance between the predicted saliency map and a center bias, modeled as a Gaussian, and sort the videos in decreasing distance. For this analysis, we consider 25[th] percentile most off-centered videos for Memento10k and VideoMem separately.

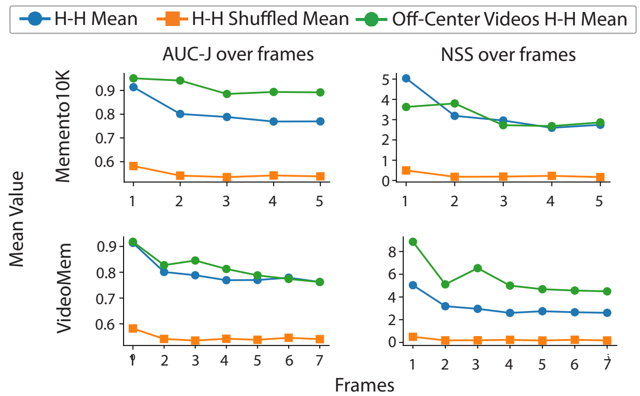

Figure 6. Framewise split-half AUC-J and NSS scores for Memento10K (left) and VideoMem (right). The x-axis shows sub-sampled frames at T=5 for Memento10K and T=7 for VideoMem. The blue line (H-H) indicates the framewise alignment between gaze patterns, averaged over all 140 videos. The green line captures framewise alignment averaged over 35/140 videos that have most off-center saliency in the initial frames. The orange line represents H-H shuffled, mean alignment when gaze patterns are compared across random videos.

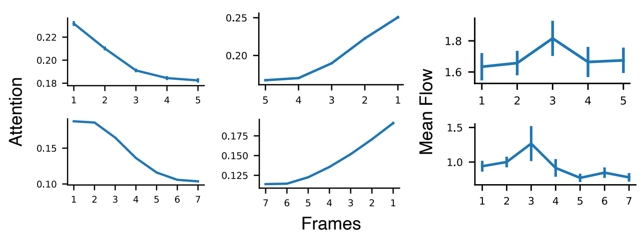

Figure 7. Left: Distribution of temporal attention across video frames in normal order, showing peak at the early frames. Middle: Distribution of temporal attention across video frames in *reversed* order as a control to rule out position bias. Right: Mean optical flow magnitude across frames to rule out motion as a bias for the stronger temporal attention at the beginning. The x-axis indicates the number of sub-sampled frames; T=5 for Memento10K (top) and T=7 for VideoMem (bottom).

dataset, identifying several critical issues that researchers must consider when constructing new datasets.

Limitations. The current datasets have 10k videos each. A model trained on them may not generalize well to any video from the internet, especially in specific domains where the visual stimuli are typically similar across all clips, *e.g.* identifying memorable parts from a lecture video. Additionally, the model processes extracted frames rather than full videos, which may result in the loss of important details for memorability and could affect comparison with human data, where viewers see the entire video.

Acknowledgments. We thank Sriya Ravula and Dr. Priyanka Srivastava for help with eye-tracking equipment and its setup. The study was supported by the IIIT-H Faculty Seed Fund (VS) and an Adobe Research Gift (MT).

References

[1] Ada Aka, Sudeep Bhatia, and John McCoy. Semantic determinants of memorability. *Cognition*, 239:105497, Oct. 2023. 1, 2

[2] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv: 1607.06450*, 2016. 3

[3] Wilma A. Bainbridge. The memorability of people: Intrinsic memorability across transformations of a person's face. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 43(5):706–716, May 2017. 2

[4] Wilma A. Bainbridge. Chapter One - Memorability: How what we see influences what we remember. In Kara D. Federmeier and Diane M. Beck, editors, *Psychology of Learning and Motivation*, volume 70 of *Knowledge and Vision*, pages 1–27. Academic Press, Jan. 2019. 1, 2

[5] Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W. Pillow, Uri Hasson, and Kenneth A. Norman. Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, 95(3):709–721.e5, Aug. 2017. 2

[6] Chris M. Bird, James L. Keidel, Leslie P. Ing, Aidan J. Horner, and Neil Burgess. Consolidation of Complex Events via Reinstatement in Posterior Cingulate Cortex. *Journal of Neuroscience*, 35(43):14426–14434, Oct. 2015. 2

[7] Ali Borji, Dicky N Sihite, and Laurent Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2012. 6

[8] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015. 7, 11, 13, 14

[9] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. 6

[10] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 2, 7

[11] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 7, 17

[12] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, and Martin Engilberge. VideoMem: Constructing, Analyzing, Predicting Short-term and Long-term Video Memorability. In *ICCV*, 2019. 2, 3, 4, 5, 11, 14, 15

[13] Romain Cohendet, Karthik Yadati, Quin Ngoc, and Claire-Hélène Demarty. Annotating, understanding, and predicting long-term video memorability. In *ICMR '18: 2018 International Conference on Multimedia Retrieval*, pages 178–186, 2018. 2

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of Association of Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 15

[15] Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. What Makes an Object Memorable? In *ICCV*, 12 2015. 2

[16] T. Dumont, J. Hevia, and C. L. Fosco. Modular memorability: Tiered representations for video memorability prediction. In *CVPR*, 2023. 2, 3, 4, 5, 15

[17] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. AMNet: Memorability Estimation with Attention. In *CVPR*, pages 6363–6372, June 2018. 2

[18] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. GANalyze: Toward Visual Definitions of Cognitive Image Properties. In *ICCV*, 2019. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 3

[20] Harini S I, Somesh Singh, Yaman K Singla, Aanisha Bhattacharyya, Veeky Baths, Changyou Chen, Rajiv Ratn Shah, and Balaji Krishnamurthy. Long-Term Memorability On Advertisements. *arXiv:2309.00378v1*, 2023. 2, 3, 5, 15

[21] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR*, 2011. 1, 2

[22] Sen Jia and Neil Bruce. Revisiting Saliency Metrics: Farthest-Neighbor Area Under Curve. In *CVPR*, 2020. 7

[23] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113. IEEE, 2009. 6

[24] Seyed-Mahdi Khaligh-Razavi, Wilma A. Bainbridge, Dimitrios Pantazis, and Aude Oliva. From what we perceive to what we remember: Characterizing representational dynamics of visual memorability. *bioRxiv doi: 10.1101/049700*, 2016. 2

[25] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and Predicting Image Memorability at a Large Scale. In *ICCV*, 2015. 2, 7, 13, 14

[26] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 4

[27] Jessica E. Kosie and Dare Baldwin. Attentional profiles linked to event segmentation are robust to missing information. *Cognitive Research: Principles and Implications*, 4(1):8, Mar. 2019. 2

[28] Cameron Kyle-Davidson, Adrian G Bors, and Karla K Evans. Generating memorable images based on human visual memory schemas. *arXiv preprint arXiv:2005.02969*, 2020. 3

[29] Matthias Kümmerer, Thomas Wallis, Leon Gatys, and Matthias Bethge. Understanding Low- and High-Level Contributions to Fixation Prediction. In *ICCV*, 2017. 8

[30] Muxuan Lyu, Kyoung Whan Choe, Omid Kardan, Hiroki Kotabe, John Henderson, and Marc Berman. Overt attentional correlates of memorability of scene images and their relationships to scene semantics. *Journal of Vision*, 20, 2020. 7

[31] Christopher R. Madan. Exploring word memorability: How well do different word properties explain item free-recall probability? *Psychonomic Bulletin & Review*, 28(2):583–595, Apr. 2021. 1, 2

[32] Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: CLIP Prefix for Image Captioning. *arXiv:2111.09734*, 2021. 5, 16

[33] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. Multimodal Memorability: Modeling Effects of Semantics and Decay on Video Memorability. In *ECCV*, 2020. 1, 2, 3, 4, 5, 11, 13, 14, 15

[34] Nabil Ouerhani, Heinz Hügli, René Müri, and Roman Wartburg. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, 2004. 6

[35] Shay Perera, Ayellet Tal, and Lihi Zelnik-Manor. Is Image Memorability Prediction Solved? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 800–808, June 2019. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv: 2103.00020*, 2021. 3, 12

[37] Umesh Rajashekar, Lawrence Cormack, and Alan Bovik. Point of gaze analysis reveals visual search strategies. *Proceedings of SPIE - The International Society for Optical Engineering*, 2004. 6

[38] Nicole C. Rust and Vahid Mehrpour. Understanding image memorability. *Trends in cognitive sciences*, 24(7):557–568, July 2020. 2

[39] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. Show and recall: Learning what makes videos memorable. In *ICCV Workshops*, pages 2730–2739, 2017. 2

[40] Oleksii Sidorov. Changing the image memorability: From basic photo editing to gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

[41] Hammad Squalli-Houssaini, Ngoc Duong, Marquant Gwenaelle, and Claire-Hélène Demarty. Deep learning for predicting image memorability. In *ICASSP*, 2018. 2, 5, 13

[42] SR Research. EyeLink 1000 Plus - SR Research. https://www.sr-research.com/eyelink-1000-plus/, 2023. 4

[43] Lorin Sweeney, Mihai Gabriel Constantin, Claire-Hélène Demarty, Camilo Fosco, Alba G Seco de Herrera, Sebastian Halder, Graham Healy, Bogdan Ionescu, Ana Matran-Fernandez, Alan F Smeaton, et al. Overview of the mediaeval 2022 predicting video memorability task. *arXiv preprint arXiv:2212.06516*, 2022. 4

[44] Benjamin Tatler, Roland Baddeley, and Iain Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision research*, 2005. 6

[45] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. *CoRR*, abs/2003.12039, 2020. 8

[46] Antonio Torralba, Aude Oliva, Monica Castelhano, and John Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review*, 113, 2006. 2

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[48] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008. 7

[49] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal Relational Reasoning in Videos. In *ECCV*, 2018. 4

[50] Tong Zhu, Feng Zhu, Hancheng Zhu, and Leida Li. Aesthetics-Assisted Multi-task Learning with Attention for Image Memorability Prediction. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2020. 3