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

Supplementary Material

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

In the supplementary material, we present an expanded set of results and analyses to better understand of our work. Appendix A provides details of our eye-tracking setup, methodology, and apparatus. Appendix B examines the performance of our model on image memorability tasks and impact of transfer learning. Appendix C presents additional experiments and results: (i) model ablation results and comparison to state-of-the-art on VideoMem [12]; (ii) detailed qualitative analysis on both Memento10K [33] and VideoMem datasets including human gaze and model attention maps; (iii) additional similarity metrics and assessment of the impact of video complexity; and (iv) results comparing human gaze *vs*. model attention on the FIGRIM [8] image memorability dataset. Appendix D explains the integration of text captions into our model, and the corresponding results. Finally, Appendix E discusses the results of applying panoptic segmentation to better understand the semantic concepts in the scene.

A. Eye-tracking Setup

A.1. Experiment Setup Details

The eye-tracking experiment is structured in the form of a continuous recognition experiment, where we present participants with a series of videos and instruct them to press the SPACEBAR when they recognize a video as being a repeat of one they had seen earlier in the sequence. As feedback for participants, we change the background color of the display to GREEN in case of a true positive and RED in case of a false positive.

We recruit 20 participants to watch 200 videos each from the Memento10K and VideoMem datasets, each participant watching videos exclusively from one dataset.

We select participants based on a strict criterion relating to their visual acuity, only considering individuals with a refractive error (eyeglass power) within the range of $[-1, +1]$ diopters. We establish this criterion in order to maintain a standard level of natural visual acuity among participants. Additionally, we require all participants to view the videos without the aid of eyeglasses, ensuring that any corrective lenses did not affect the pupil tracking device.

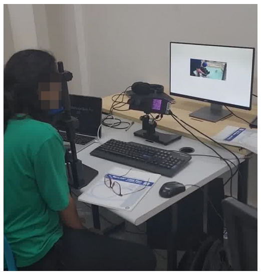

Figure 8. Participant watching videos from Memento10K during the eye-tracking experiment (face anonymized).

For the participants watching videos from Memento10K we display videos in their original size and aspect ratio on a screen of size 1024×768. For participants watching videos from VideoMem, we display videos in their original aspect ratio, resized to fit the screen width. For example, we convert videos with size 1920×1080 to 1024×576, maintaining the aspect ratio of 1.77.

We calibrate and validate pupil positions after every 20 videos for Memento10K and 10 videos for VideoMem (approximately 1 minute). Participants use a mounted chin-rest while viewing videos, placed at a distance of 35 cm from the screen.

The primary interest is in capturing the participants' fixations while engaged in a memory game similar to the original studies of Memento10K and VideoMem.

The eye-tracking study involving human participants was reviewed and approved by the Institute Review Board (IRB). The participants provided their written informed consent to participate in the study.

A.2. Eye-tracking Procedure

The main procedure of the experiment (sequence in which videos are shown) is presented in Fig. 9. An instance of a participant watching the videos can be seen in Fig. 8.

Video selection. We select 200 videos each from the validation sets consisting of 1500 videos in Memento10K and

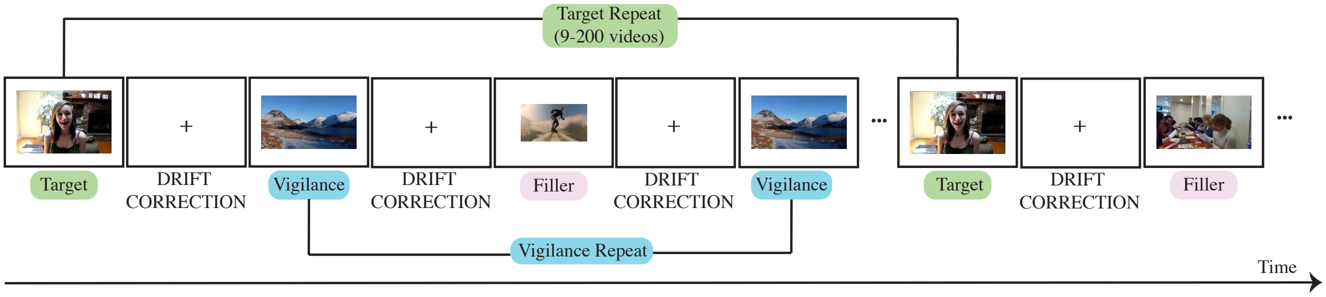

Figure 9. Design of eye-tracking experiment. A subject watches alternating videos and drift correction fixation crosses (typically between 0.5 s to 1 s). A vigilance video (one of 40) is repeated in a short interval of 2-3 videos to ensure that the subject is alert, while the target videos (one of 20) have a lag of at least 9 videos. Filler videos (80) are not repeated.

1000 in VideoMem. To ensure a representative and varied selection of videos, we use a two-step process:

1. **Clustering:** We initially cluster videos based on their visual features. We extract the average CLIP ResNet [36] embeddings from selected frames of each video — T=5 linearly spaced frames for videos from the Memento dataset and T=7 linearly spaced frames from Videomem. We then group these videos into 28 distinct clusters using K-Means Clustering, providing a structured framework for subsequent selection. We choose K=28 by visually inspecting the quality of clusters (generated from hierarchical clustering) for values around 30.

2. **Binning:** Following clustering, we bin videos based on their ground truth memorability scores, creating 10 distinct bins. This stratification allows for a balanced representation of memorability levels within the selected videos.

We select videos for the experiment through the following sampling strategy: Initially, we sample one video from each cluster-bin combination, ensuring broad coverage across all memorability levels and visual characteristics. In instances where the initial sampling does not yield 200 videos, we sample for a second iteration. This round involves selecting an additional video from some cluster-bin combinations, again governed by the availability of videos within each category. To adhere to the desired total of 200 videos, we uniformly remove any excess videos from the sampled pool. We randomly select and remove these excess videos from the cluster-bin combinations, ensuring an even distribution across all categories.

From these selected 200 videos, the experiment design requires a refined set of 140 unique videos (20 target repeats, 40 vigilance repeats, and 80 fillers). We, therefore, randomly select videos for each category (vigilance, target, and regular) from the pool of 200 videos, ensuring that each category had a distinct set of videos. The target and vigilance repeats are the same across all participants. For each experimental run, we use a unique order of video presentations. This involves mixing regular videos with the vigilance and target videos and then randomly shuffling this combined set. We constrain the placement of repeated vigilance videos to a lag of $2-3$ videos, while for target videos, we maintain a minimum lag of 9 videos (similar to VideoMem and Memento10k).

A.3. Details of Gaussian blur

To account for the visual field of a participant, we apply a Gaussian blur to fixation maps obtained from the experiment. The standard deviation (σ) of the Gaussian blur is calculated using the formula:

$$\sigma = \frac{\text{Pixels Per Degree}}{2.355}. \quad (4)$$

Here, 2.355 is a constant derived from the assumption that the visual angle corresponds to the Full Width at Half Maximum (FWHM). The Pixels Per Degree (PPD) is computed as follows:

$$\text{PPD} = \frac{2 \times d \times \tan(\frac{\theta}{2})}{h \times y}, \quad (5)$$

where, d is the distance of the participant to the screen (13.77 inch or 35 cm), θ is the visual angle (assumed to be 1°), h is the height of the screen (23.5 inch), and y is the vertical resolution of the screen (768 pixels in our case).

A.4. Metric: AUC-Percentile

For a video V_k and video frame f_{ik}, the true frame similarity score is computed using AUC-Judd between the model's attention map α_{ik} and the corresponding gaze fixation density map G_{ik}. Then, for each video, we compute a true video similarity score by averaging the true frame similarity scores.

We perform a permutation test by comparing the attention map α_{ik} of frame f_{ik} against the fixation map G_{il} from frame f_{il} of a randomly chosen different video $V_l, l \neq k$.

This process is repeated 100 times, yielding a distribution of 100 video-level similarity AUC-Judd scores under the null hypothesis of no specific relationship between model attention and human fixations.

We denote AUC-Percentile for each video as the percentile of the true video similarity score within the distribution of permuted AUC-Judd scores. A high AUC-Percentile indicates a strong alignment between the model's attention map and the human gaze fixation density map for the same video, relative to a null distribution of comparisons between different videos. For example, an AUC-Percentile of 80 implies that there is a less than 20% chance that the observed alignment between the model's attention map and the human gaze fixation density map could be attributed to chance or general center-bias in the data.

B. Transferring from/to Image Memorability

To ascertain the reliability of our simple approach, we evaluate on image memorability tasks by considering the image as a "video" of $T{=}1$. As seen in Tab. 4, on the LaMem dataset [25] we match SoTA results (0.720 RC [41]). On the FIGRIM dataset [8], we achieve results close to human performance (0.74 RC [8]). Previous studies [33] pretrain models on image memorability datasets and then fine-tune them for video memorability prediction. Tab. 4 R2 *vs*. R4 shows a small improvement in Memento10k RC score from 0.706 to 0.718 with LaMem pretraining. However other results do not improve. We also observe that training on one dataset and evaluating on another (rows 1-3) usually leads to significant degradation and is an important problem for future work.

C. Additional Results and Qualitative Analysis

In this section, we present the results for video memorability prediction on the VideoMem dataset, followed by a qualitative analysis of the model's performance. Finally, we explore the alignment between human gaze and model attention through various analyses on both video and image memorability datasets.

C.1. Video Memorability prediction for Videomem

Expanding on the model ablations for Memento10k in Sec. 4.1 (of the main paper), Tab. 5 shows results for VideoMem, which generally follows similar trends, with Row 1 (R1) achieving the best results. However, random sampling during training does not improve performance and including or predicting captions has no impact, perhaps due to the noise in the captions.

SoTA comparisons are shown in Tab. 6. As the test set memorability scores (labels) for VideoMem are not available, no previous work apart from the creators of the dataset have evaluated on a held-out test set. Instead, all approaches likely overfit on the validation set with RC scores much higher than the human-human consistency RC at 0.481. Our scores are lower than other SoTA methods, likely due to the challenges discussed in Sec. 4.2. However, we suspect that other models that leverage multiple modalities are strongly overfitting on this dataset.

C.2. Qualitative Analysis

We provide a qualitative analysis of the model's predictions and the alignment of its attention maps with human gaze, highlighting the model's successes and failures.

Best, worst, over, and under predictions. A few qualitative examples of different predictions of our model across both datasets can be seen in Fig. 10. The model seems to perform well on videos with a clear subject (face, a man playing with their dog, *etc*.). Worst predictions (over and under) are observed on underexposed (dark) videos. The model tends to over-predict on certain videos with clutter, while under-predict on scenic videos.

Visualizing gaze and attention maps. The human gaze fixation maps and model attention maps across multiple videos can be seen in Fig. 13 for Memento10k and Fig. 14 for VideoMem. In both cases, model attention maps appear to be more similar to human gaze maps in higher memorability (GT) videos compared to lower memorability ones. Note, in Sec. 4.3, we rule out the possibility that this alignment between model attention and human gaze is driven by center-bias.

C.3. Additional Results Comparing Human Gaze *vs*. Model Attention (Video Memorability)

We expand on the evaluation of human gaze and model attention alignment using additional metrics and explore how video complexity affects this alignment.

Additional similarity metrics. To compare human gaze fixation maps to the model's attention maps, we use standard metrics used in saliency evaluation such as AUC-Judd, NSS, CC, KLD. Additionally, we develop and apply a novel shuffle-based metric, the AUC-Percentile.

While Fig. 5 from the main paper shows results only on AUC-Judd and NSS due to space restrictions, we now extend this to all metrics in Fig. 11. We observe a common trend of greater match between human gaze and model attention maps with increasing memorability scores across most metrics, indicating that memorable videos attract both human and model attention to the same regions of the video frames.

Impact of video complexity on gaze/attention alignment We split the videos in each dataset at the median of the average number of objects per frame to get one group of simpler and one group of more complex videos. We computed model attention-human gaze (M-H) and human-human (H-

Table 4. Results of transferring an image/video memorability model to images/videos. Datasets: LM: LaMem [25], M10k: Memento10k [33], VM: VideoMem [12], and FG: FIGRIM [8]. Training strategy: P for pretraining and F for fine-tuning. Results reported on validation set.

	Train on				LaMem		Memento10k		VideoMem		FIGRIM	
	LM	M10k	VM	FG	RC	MSE	RC	MSE	RC	MSE	RC	MSE
1	F	-	-	-	**0.729**	0.0074	0.526	0.0220	0.382	0.0233	0.647	0.0168
2	-	F	-	-	0.547	0.0273	*0.706*	0.0061	0.439	0.0165	0.351	0.0525
3	-	-	F	-	0.549	0.0147	0.525	0.0089	**0.513**	0.0060	0.501	0.0355
4	P	F	-	-	0.679	0.0161	**0.718**	0.0568	0.446	0.0144	0.634	0.0318
5	P	-	F	-	0.688	0.0090	0.459	0.0096	0.504	0.0059	0.627	0.0237
6	P	-	-	F	0.678	0.0113	0.507	0.0130	0.392	0.0191	**0.742**	0.0123
7	P	F	F	-	0.664	0.0135	0.689	0.0058	0.483	0.0062	0.626	0.0273

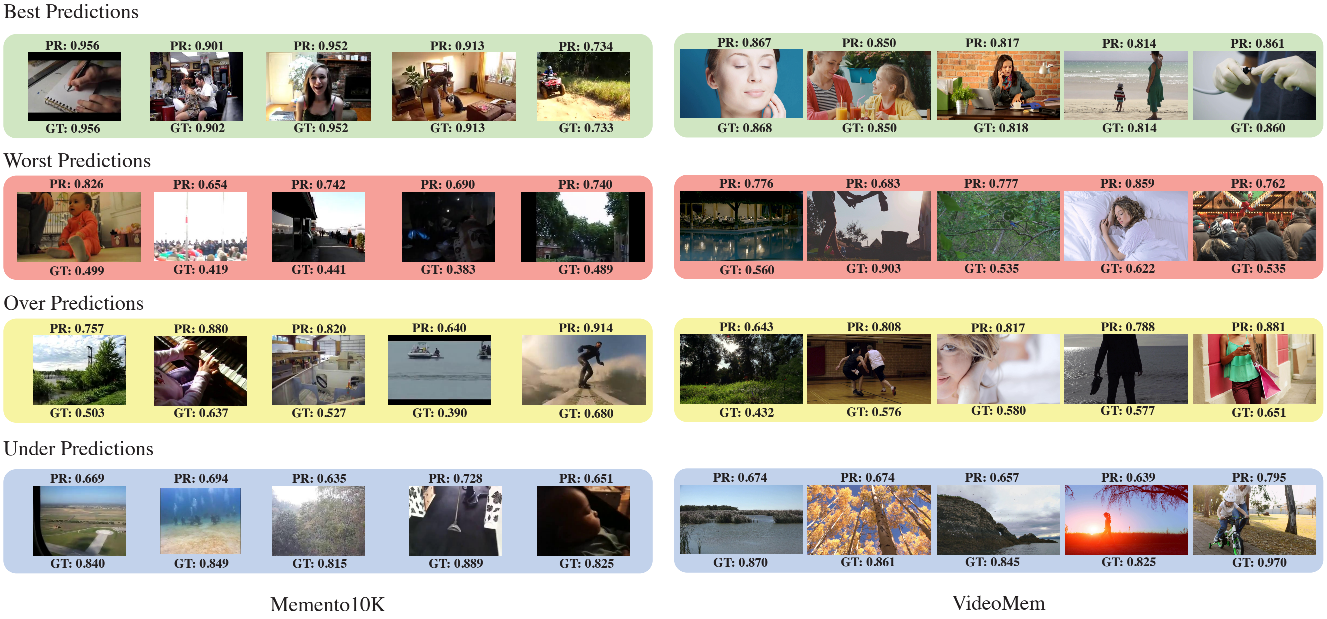

Figure 10. Qualitative analysis of different predictions of our model over Memento10K *(Left)* and VideoMem *(Right)*. Ground-truth (GT) and predicted (PR) memorability scores are annotated at the bottom and top of each frame, representative of the videos. Best viewed on screen by zooming in.

H) gaze alignment scores for these groups of videos. The alignment metrics are presented in Tab. 7 and indicate that in both datasets, humans gaze patterns tend to agree with those of other humans as well as model attention patterns with no statistically significant differences between simple and complex videos except in the M-H NSS metric for Memento10k. Therefore, the results presented in the main paper are unlikely to be explained by complexity of the videos.

C.4. Human Gaze *vs.* Model Attention (Image Memorability)

Next, to establish the general trend of similarity between model attention and human gaze with increasing memorability, we also present results on the FIGRIM dataset, which provides gaze data along with memorability scores for images. While Appendix B provides quantitative results on memorability prediction, Fig. 12 illustrates a similar trend of increasing human gaze and model attention agreement with increasing memorability scores on the FIGRIM dataset.

D. Modeling with Captions

Building upon Sec. 3.1 where we presented the vision-only model, we now explain how captions can be easily integrated into the existing modeling framework. We consider two paradigms. In the first, the caption is assumed available, both during training and inference. This may be

		Embedding				Memento10k (val)		VideoMem (val)	
	CLIP	Time	Space	Sampling	Caption	RC ↑	MSE ↓	RC ↑	MSE ↓
1	Spatio-Temporal	Fourier	-	Random	-	**0.706**	0.0061	*0.513*	*0.0060*
2	Temporal	Fourier	-	Random	-	0.687	0.0062	0.508	0.0064
3	Spatio-Temporal	Learnable	-	Random	-	0.696	0.0059	0.502	0.0060
4	Spatio-Temporal	Fourier	1D	Random	-	*0.703*	*0.0057*	0.506	**0.0059**
5	Spatio-Temporal	Fourier	2D	Random	-	0.701	**0.0056**	0.505	*0.0060*
6	Spatio-Temporal	Fourier	-	Middle	-	*0.703*	0.0066	**0.515**	**0.0059**
7	Spatio-Temporal	Fourier	-	Random	Original	**0.745**	**0.0050**	0.505	0.0061
8	Spatio-Temporal	Fourier	-	Random	Predicted	*0.710*	*0.0056*	0.508	0.0061

Table 5. **Model ablations.** Column 1 (C1) compares the impact of using spatio-temporal features versus temporal features with global average pooling. C2 and C3 specify the types of temporal and spatial position embedding used. C4 is the frame sampling method used during training. C5 indicates whether the video 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 varying visual choices: features, position-encoding, and frame sampling methods. R7 presents results with original captions 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*.

		Memento10k (test)		VideoMem (test)		Memento10k (val)		VideoMem (val)	
Methods	Caption	RC	MSE	RC	MSE	RC	MSE	RC	MSE
VideoMem ICCV19	No	-	-	0.494	-	-	-	0.503	-
SemanticMemNet ECCV20	No	0.659	-	-	-	-	-	0.555	-
M3-S CVPR23	No	-	-	-	-	0.670	0.0062	0.563	0.0046
Ours (R1 Tab. 5)	No	0.662	0.0065	-	-	0.706	0.0061	0.513	0.0060
SemanticMemNet	Yes	0.663	-	-	-	-	-	0.556	-
Sharingan arXiv	Yes	-	-	-	-	0.72	-	0.6	-
Ours (R7 Tab. 5)	Yes	0.713	0.0050	-	-	0.745	0.0050	0.505	0.0061

Table 6. Comparison against SoTA for video memorability on both test and validation sets for Memento10k and VideoMem. Baselines considered are VideoMem [12], SemanticMemNet [33], M3-S [16], and Sharingan [20]. Human-human split-half consistency scores are 0.73 for Memento10k and 0.481 for VideoMem.

achieved using recent advances in vision-language models (VLMs). In the second, we consider experiments where the caption is predicted simultaneously with the estimation of the video memorability score (similar to [33]).

D.1. Assuming Caption is Available

When the caption is given, we first extract token-level representations through a BERT encoder and append them to the spatio-temporal video tokens for memorability prediction.

Text encoder. We extract textual embeddings for the captions from the last hidden state of the BERT [14] model ψ:

$$\{\mathbf{g}_l\}_{l=1}^{N} = \psi(\{g_l\}_{l=1}^{N}), \tag{6}$$

where $\mathbf{g}_l \in \mathbb{R}^d$, N is the number of tokens, and d is the dimensionality of the embeddings, equal to the reduced dimensionality of images after the linear layer.

Changes to the video encoder. We append N text tokens to the THW visual tokens fed to the Transformer encoder. To distinguish between text and image, we append modality specific embeddings to both the visual (from Eq. 2) and text tokens. We also add position embeddings indicating order to the text tokens.

$$\mathbf{f}'_{ij} = \mathbf{W}^d \mathbf{f}_{ij} + \mathbf{E}^t_i + \mathbf{E}^s_j + \mathbf{E}^m_1, \tag{7}$$

$$\mathbf{g}'_l = \mathbf{g}_l + \mathbf{E}^c_l + \mathbf{E}^m_2, \tag{8}$$

where $i = [1, \dots, T]$, $j = [1, \dots, HW]$, $l = [1, \dots, N]$, $\mathbf{E}^t_i$ is the i^{th} row of the temporal embedding matrix (learnable or Fourier) for images, $\mathbf{E}^c_l$ is the l^{th} row of the temporal embedding matrix for the caption, $\mathbf{E}^s_j$ is the j^{th} row of the spatial embedding matrix, and $\mathbf{E}^m_{[1,2]}$ are the modality embeddings, one for visual tokens, another for text.

We combine the CLS token (with learnable parameters $\mathbf{h}_{\text{CLS}}$), image and text tokens to create a sequence of $1 + TWH + N$, apply LayerNorm, feed it to the TE.

$$[\tilde{\mathbf{h}}_{\text{CLS}}, \tilde{\mathbf{f}}_{11}, \dots, \tilde{\mathbf{f}}_{THW}, \tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_N] = \text{TE}([\mathbf{h}_{\text{CLS}}, \mathbf{f}'_{11}, \dots, \mathbf{f}'_{THW}, \mathbf{g}'_1, \dots, \mathbf{g}'_N]). \tag{9}$$

As before, $\tilde{\mathbf{h}}_{\text{CLS}}$ is used to predict the memorability score.

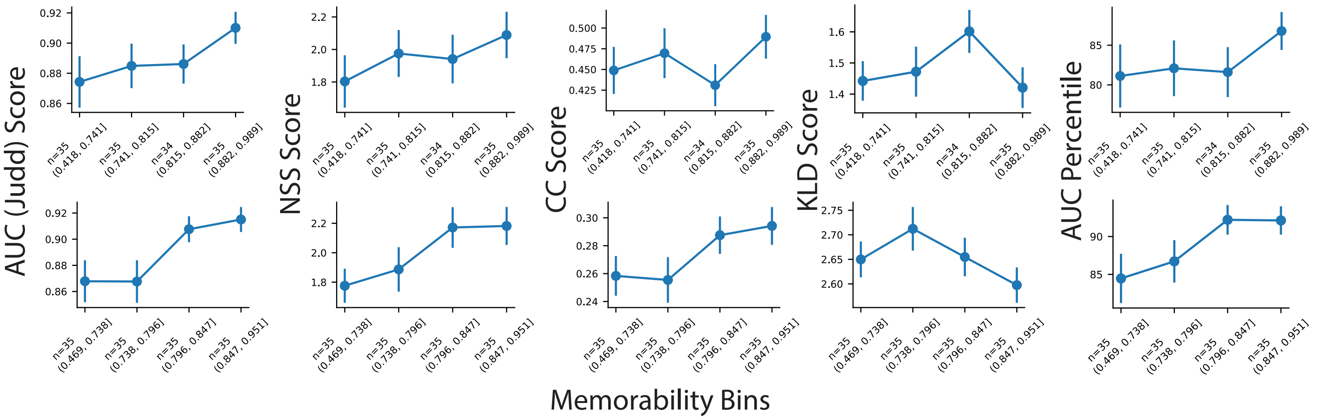

Figure 11. *Top:* Memento10K, *Bottom:* VideoMem. Performance across different similarity/distance metrics while comparing the human gaze fixation maps with model attention maps. The metrics are indicated with arrows to indicate whether higher or lower scores are better: AUC (Judd) ↑; NSS ↑; CC ↑; KLD ↓; and AUC Percentile (ours) ↑. Results are presented for n=139 videos, binned into 4 percentiles based on ground-truth memorability scores.

Metrics	Memento10K						VideoMem					
	M-H			H-H			M-H			H-H		
	Simple	Complex	t	Simple	Complex	t	Simple	Complex	t	Simple	Complex	t
AUC-J ↑	0.89 $\pm$0.01	0.88 $\pm$0.01	0.60	0.90 $\pm$0.01	0.89 $\pm$0.01	0.17	0.89 $\pm$0.01	0.89 $\pm$0.01	−0.19	0.83 $\pm$0.01	0.80 $\pm$0.01	1.53
AUC-P ↑	84.41 $\pm$2.18	81.10 $\pm$2.48	0.99	-	-	-	89.45 $\pm$1.92	88.37 $\pm$1.72	0.41	-	-	-
NSS ↑	1.89 $\pm$0.09	1.60 $\pm$0.07	**2.1**	3.02 $\pm$0.17	3.07 $\pm$0.17	−0.20	2.08 $\pm$0.14	1.94 $\pm$0.11	1.05	3.85 $\pm$0.46	3.79 $\pm$0.40	0.11
CC ↑	0.58 $\pm$0.02	0.52 $\pm$0.02	1.66	0.48 $\pm$0.02	0.49 $\pm$0.02	−0.26	0.29 $\pm$0.02	0.26 $\pm$0.01	1.46	0.28 $\pm$0.02	0.28 $\pm$0.01	0.14
KLD ↓	1.08 $\pm$0.02	1.16 $\pm$0.02	−1.41	2.19 $\pm$0.11	2.16 $\pm$0.11	0.23	2.64 $\pm$0.05	2.67 $\pm$0.04	−0.64	4.01 $\pm$0.13	4.16 $\pm$0.10	−0.89

Table 7. Comparing gaze fixation maps against model's attention map via different metrics for simple and complex videos, along with human-human alignment scores(split by half, averaged over 10 random iterations) for Memento10K and Videomem datasets. ↑ (↓) indicates higher (lower) is better. M-H: Model-human; H-H: Human-human; and t: t-test significance. Significant t-statistics are shown in bold ($p < 0.05$).

We report results when using the ground-truth caption in this approach in Tab. 1, row 7 of the main paper (w original captions as input). For Memento10k, we see a 0.04 points increase in Spearman correlation (0.706 to 0.745), however, captions do not seem to assist VideoMem.

D.2. Joint Prediction of Caption and Memorability

When the caption is not available, we consider predicting the caption along with the memorability scores. In particular, we adapt CLIPCap [32], a recent approach that connects CLIP visual features with the GPT-2 decoder using a Transformer mapping layer.

Specifically, we use a mapping network (a Transformer decoder) to convert the THW visual tokens at the output of the Transformer encoder $\tilde{\mathbf{f}}_{ij}$ to a set of P prefix tokens. The mapping network of L_D=6 layers consists of P query learnable tokens and uses visual inputs as memory, P=30. The outputs of this mapping network are fed as prefix tokens to the GPT-2, and captions are generated in an auto-regressive manner.

We train the model jointly, to predict both the memorability score (using L1 regression loss) and the caption (using cross-entropy loss). Results of this approach are presented in Tab. 2, row 3. A small increase of 0.004 is observed in the RC score (0.706 to 0.710) for Memento10k, while VideoMem continues to not benefit from captions.

We conclude that generating captions separately with a VLM and using them (as shown above) may be a better course of action than training a joint model.

E. Panoptic Segmentation

We present additional experiments and results from the semantic *stuff vs. things* analysis obtained through panoptic segmentation.

Pixel count, human gaze, and model attention across all labels. In Fig. 15, we show the distributions for all stuff and things labels. Row 1 is the probability distribution of pixel

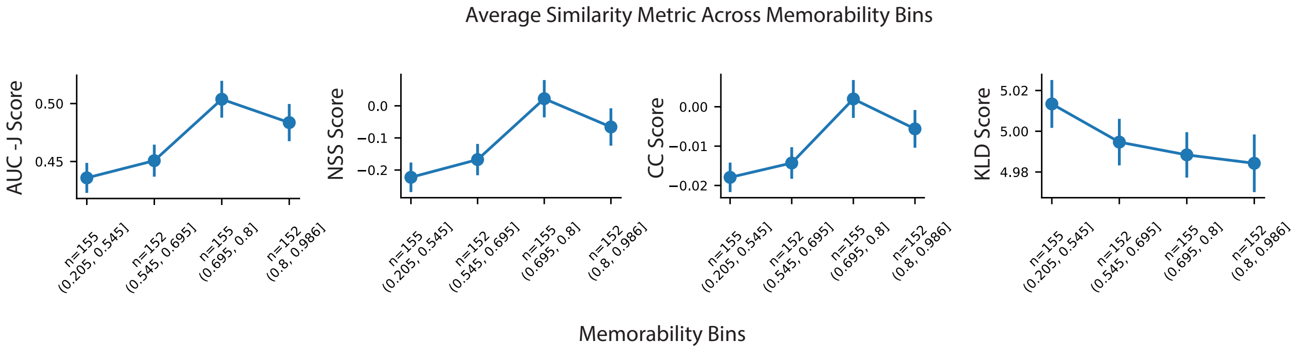

Figure 12. Performance across different similarity/distance metrics while comparing the human gaze fixation maps with model attention maps for the FIGRIM dataset. The metrics are indicated with arrows to indicate whether higher or lower scores are better: AUC (Judd) ↑; NSS ↑; CC ↑; and KLD ↓. Results are presented for n=614 images, binned into 4 percentiles based on ground-truth memorability scores.

counts and gaze/attention weighted counts for stuff labels (plotted in semilog scale). In row 2, we normalize these counts by the pixel count (blue), highlighting dynamic stuff labels such as *light, food, platform* receiving higher attention weighted scores, while other mundane labels such as *wall, sky, road* receiving lower scores.

A similar analysis is shown for *things* in rows 3 and 4. Here too, we observe that daily objects such as *bed, car, toilet* receive less human and model attention to account for memorability, while dynamic or interesting objects such as *person, dog, bird, wine glass, banana* (among others) receive higher attention. This confirms that not all objects are interesting.

Note, while this analysis is also subject to accuracy of Maskformer [11] (the panoptic segmentation approach), qualitatively, we find this to be quite reliable as seen in Fig. 16.

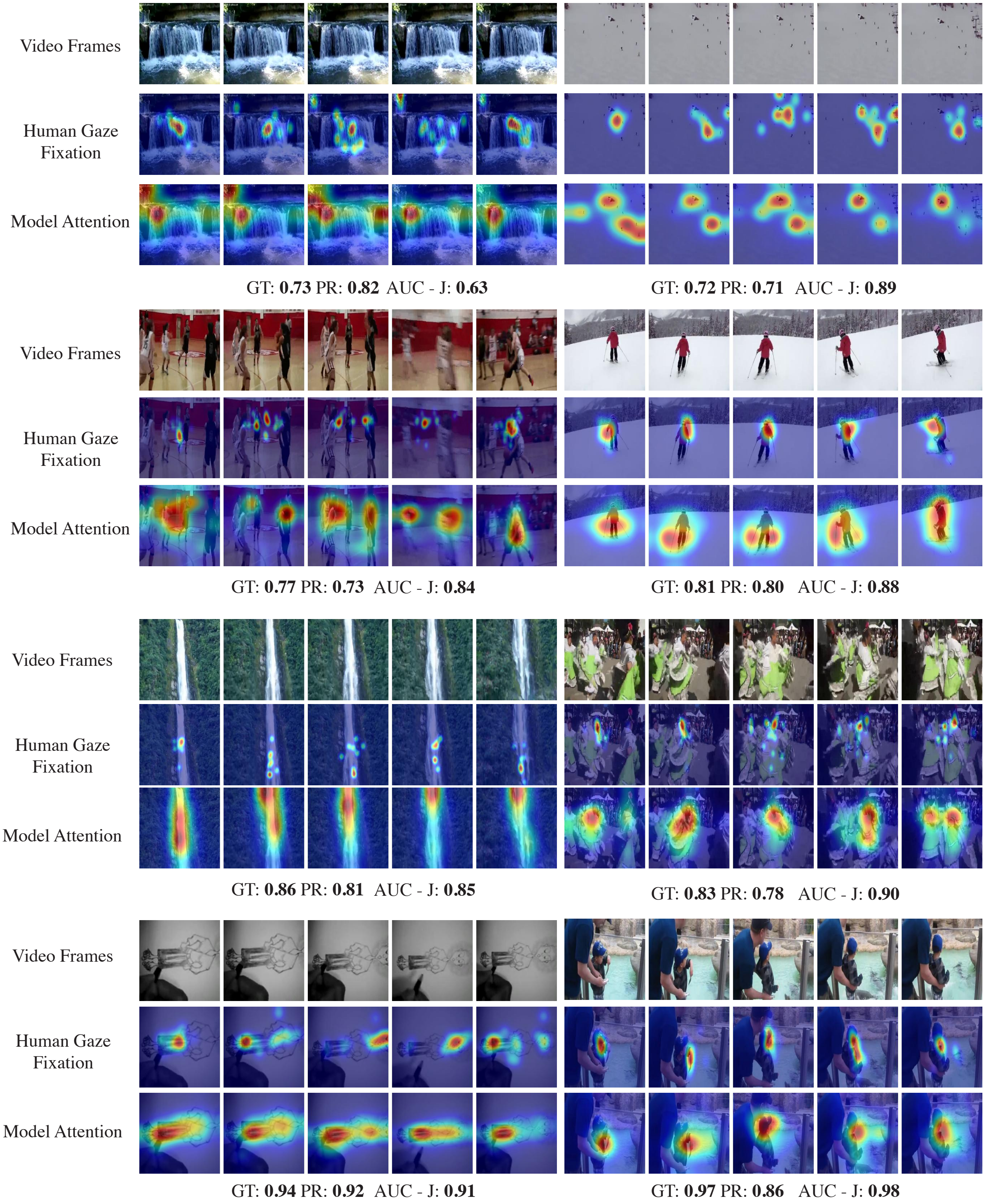
Video Frames
Human Gaze Fixation
Model Attention
GT: 0.73 PR: 0.82 AUC - J: 0.63
GT: 0.72 PR: 0.71 AUC - J: 0.89
Video Frames
Human Gaze Fixation
Model Attention
GT: 0.77 PR: 0.73 AUC - J: 0.84
GT: 0.81 PR: 0.80 AUC - J: 0.88
Video Frames
Human Gaze Fixation
Model Attention
GT: 0.86 PR: 0.81 AUC - J: 0.85
GT: 0.83 PR: 0.78 AUC - J: 0.90
Video Frames
Human Gaze Fixation
Model Attention
GT: 0.94 PR: 0.92 AUC - J: 0.91
GT: 0.97 PR: 0.86 AUC - J: 0.98

Figure 13. Comparison of original video frames, gaze fixation maps, and model attention maps on the Memento10K dataset. We also indicate the ground-truth and predicted memorability scores, and the AUC Judd score measuring similarity between saliency maps.

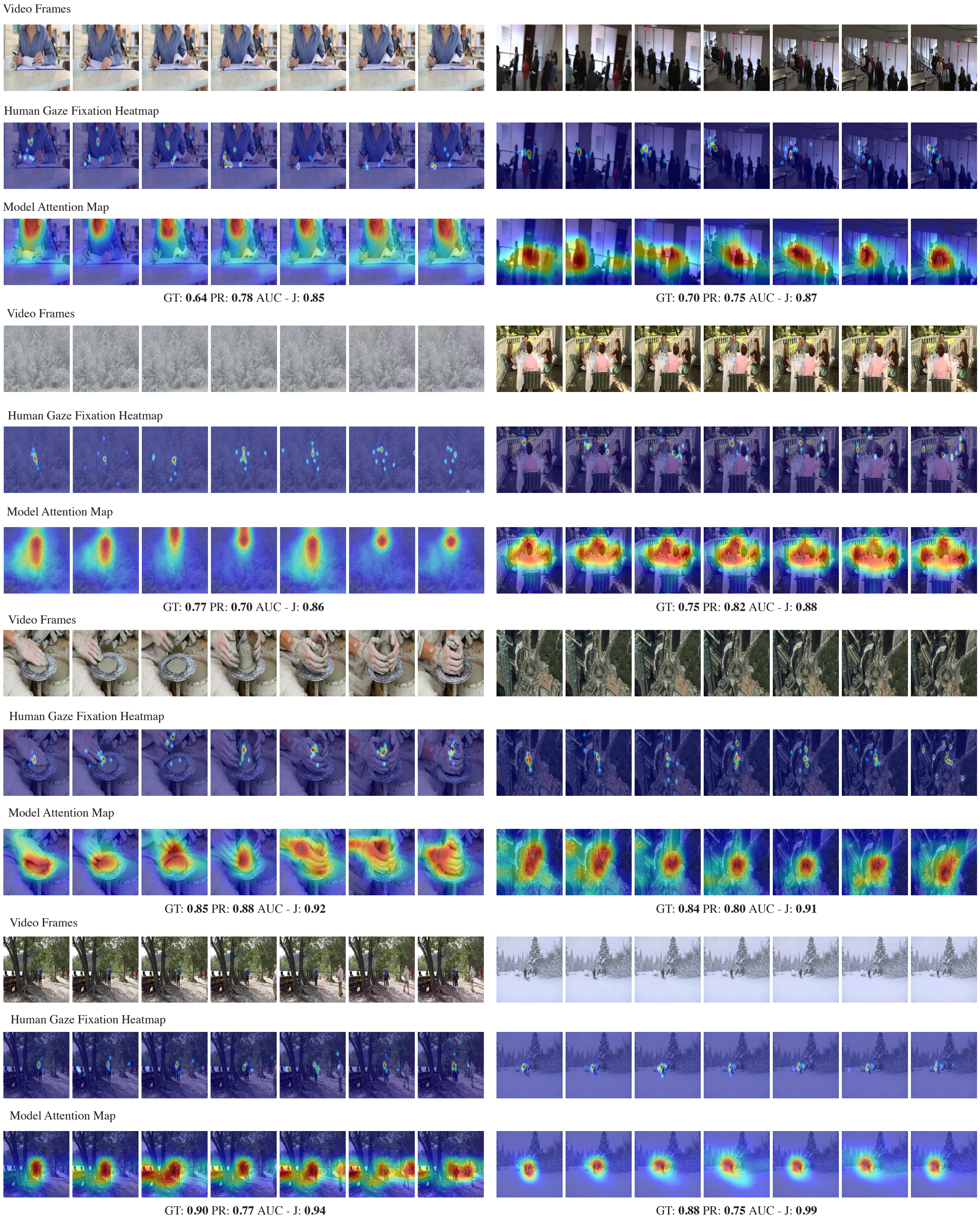
Video Frames
Human Gaze Fixation Heatmap
Model Attention Map
GT: 0.64 PR: 0.78 AUC - J: 0.85
GT: 0.70 PR: 0.75 AUC - J: 0.87
Video Frames
Human Gaze Fixation Heatmap
Model Attention Map
GT: 0.77 PR: 0.70 AUC - J: 0.86
GT: 0.75 PR: 0.82 AUC - J: 0.88
Video Frames
Human Gaze Fixation Heatmap
Model Attention Map
GT: 0.85 PR: 0.88 AUC - J: 0.92
GT: 0.84 PR: 0.80 AUC - J: 0.91
Video Frames
Human Gaze Fixation Heatmap
Model Attention Map
GT: 0.90 PR: 0.77 AUC - J: 0.94
GT: 0.88 PR: 0.75 AUC - J: 0.99

Figure 14. Comparison of original video frames, gaze fixation maps, and model attention maps on the VideoMem dataset. We also indicate the ground-truth and predicted memorability scores, and the AUC Judd score measuring similarity between saliency maps.

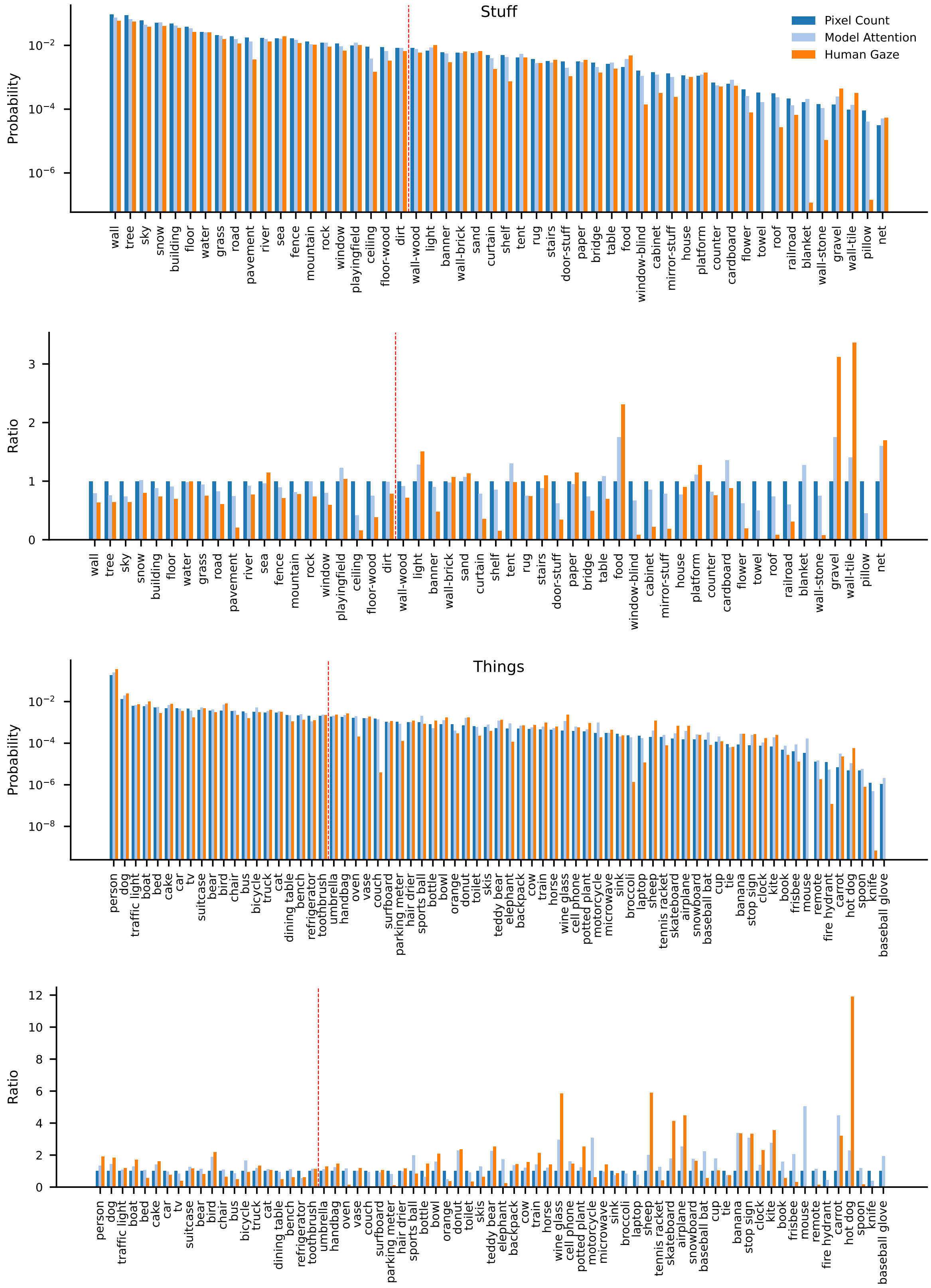

Figure 15. Analysis of panoptic segmentation results. The vertical red line marks the top-20 labels within these categories. **First** and **Third**: Raw, attention-, and gaze-weighted pixel probabilities for *stuff* and *things*, respectively (plotted in semilog scale); **Second** and **Fourth**: Highlights how model attention-weighted and human gaze-weighted pixel counts are higher or lower relative to normalized raw pixel counts for *stuff* and *things*.

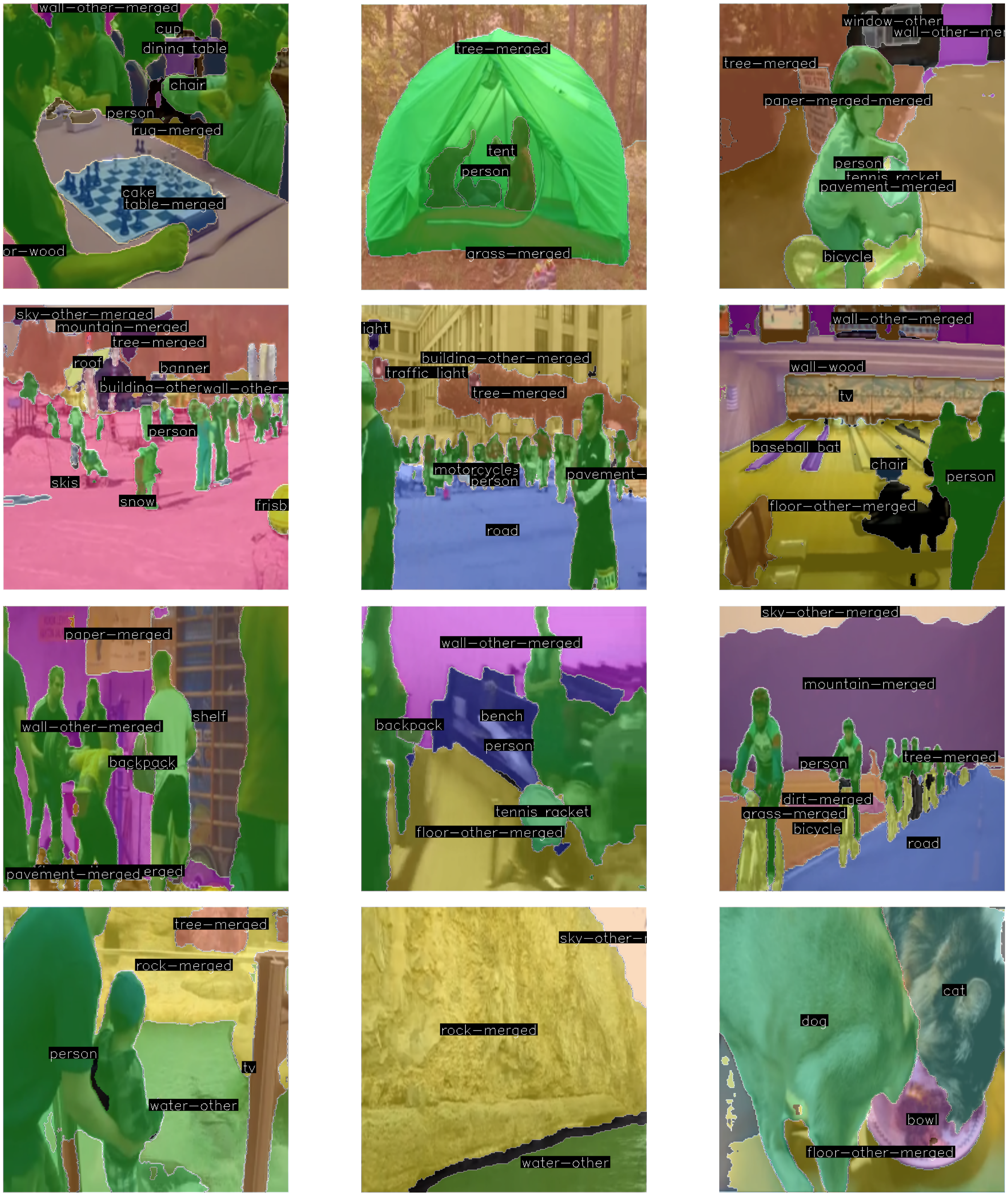

Figure 16. Visualisation of panoptic segmentation predictions on Memento10k dataset.