Correlates of Image Memorability in Vision Encoders: Activations, Attention Entropy, Patch Uniformity and Autoencoder Losses

Ece Takmaz (e.k.takmaz@uu.nl)[1] **& Albert Gatt**[2] **& Jakub Dotlačil**[1]
[1]Institute for Language Sciences, Department of Languages, Literature and Communication, Utrecht University
[2]Department of Information and Computing Sciences, Utrecht University

Abstract

Images vary in how memorable they are to humans. Inspired by findings from cognitive science and computer vision, we explore correlates of image memorability in pretrained transformer-based vision encoders for the first time. Focusing initially on activations, attention distributions, and the uniformity of image patches, we find that these features correlate with memorability to some extent. Additionally, we explore sparse autoencoder loss over the representations of vision encoders as a proxy for memorability, which yields results outperforming past methods using convolutional neural network representations. Our results shed light on the relationship between model-internal features and memorability. They show that some features are informative predictors of what makes images memorable to humans; revealing that, in particular, the reconstruction loss from our autoencoders is a strong correlate of image memorability.

Keywords: image memorability; vision encoders; autoencoders; vision transformers, attention entropy

Introduction

Humans can remember very well whether they saw an image, even after viewing 10,000 pictures (Standing, 1973), thanks to the vast storage capacity of visual long-term memory (Brady et al., 2008). However, some images are more memorable than others; see Figure 1. **Image memorability** is a complex phenomenon; yet, it is consistent across individuals, and seems to be an intrinsic property of images (Bainbridge, 2019; Bylinskii et al., 2022; Rust & Mehrpour, 2020), albeit influenced by extrinsic, contextual factors (Bylinskii et al., 2015). What makes an image memorable has long been a question that has been posed both in research on visual cognition and computer vision (CV) (Bainbridge, 2019; Bylinskii et al., 2022; Isola, Parikh, et al., 2011; Isola, Xiao, et al., 2011; Isola et al., 2014; Rust & Mehrpour, 2020). The former line of research collects data from humans and extracts recognition scores. The latter proposes models to predict memorability and also to edit images to alter their memorability scores.

Various factors have been identified to be associated with image memorability: **stronger responses** elicited in the brain for more memorable images (Jaegle et al., 2019), as well as **deeper levels of processing** during encoding, which benefits memory retention (Craik & Lockhart, 1972). In CV, scene category is identified as a significant factor in whether an image is memorable–e.g., natural landscape images with **large uniform regions** are forgotten more than images of people (especially faces) or salient objects (Bylinskii et al., 2015; Goetschalckx & Wagemans, 2019; Isola, Xiao, et al., 2011; Isola et al., 2014; Khosla et al., 2015). Conversely, other properties such as interestingness, aesthetics, and low-level features (e.g., color and contrast) are not as correlated with memorability (Isola, Xiao, et al., 2011; Isola et al., 2014; Khosla et al., 2015). The **distribution of attention** over images is also connected to memorability, with more focused attention observed in more memorable images (Celikkale et al., 2013; Khosla et al., 2015; Mancas & Le Meur, 2013).

(a) **Mem:** 0.98 **Loss:** 294.95 (b) **Mem:** 0.14 **Loss:** 219.02

Figure 1: Images with the highest (a) and lowest (b) memorability scores (Mem) from the MemCat dataset (Goetschalckx & Wagemans, 2019), along with the autoencoder reconstruction losses (Loss) incurred by them in our autoencoder whose losses best correlate with image memorability scores.

Although earlier research used handcrafted features to predict memorability, the introduction of large-scale datasets of images with empirical memorability scores, such as LaMem (Khosla et al., 2015) and MemCat (Goetschalckx & Wagemans, 2019) facilitated the use of deep learning techniques. Influential works mainly use convolutional neural networks (CNNs) (Khosla et al., 2015; Needell & Bainbridge, 2022; Perera et al., 2019; Praveen et al., 2021; Sadana et al., 2024; Squalli-Houssaini et al., 2018), with more recent work also training vision transformers (Dosovitskiy et al., 2021; Hagen & Espeseth, 2023). Recently, Vogelsang and Heilbron (2025) investigated the magnitudes of image representations obtained from layers of CNNs as potential signs of image memorability across a dataset of images that depict objects (THINGS database; Hebart et al., 2019; Kramer et al., 2023).

Closer to our work, Bagheri and Mohsenzadeh (2025) and Lin et al. (2024) propose the use of **reconstruction loss from an autoencoder** and **reconstruction residual from a sparse coding model**, respectively, as a proxy for memorability. Both studies utilize image representations extracted from CNNs and observe higher reconstruction errors for images that require

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

more complex computations or levels of processing, which are linked to higher memorability (Bagheri & Mohsenzadeh, 2025; Lin et al., 2024). These approaches carry theoretical and applied relevance. In computational neuroscience, sparse coding has been proposed to explain the sparse activations observed in the primary visual cortex in response to natural images (Olshausen & Field, 1996, 1997). In artificial neural network interpretability research, sparse autoencoders have been utilized to represent internal activations as a sparse linear combination of dictionary elements, facilitating the extraction of interpretable features from models (Cunningham et al., 2024; Elhage et al., 2022; Sharkey et al., 2023).

In this paper, for the first time, we investigate the extent to which state-of-the-art **transformer-based vision encoders** capture **correlates of image memorability** in a way compatible with the findings in visual cognition, without being trained specifically for this purpose. We first explore the internals of these encoders, inspecting activations, distributions of the attention applied over the image regions and the uniformity of the image regions. Secondly, we train sparse autoencoders built on the representations from vision encoders to investigate the correlation between reconstruction loss and memorability. Our results indicate that model-internal features correlate with memorability to some extent, although the patterns across layers and models are not very consistent. On the other hand, our sparse autoencoders yield losses that strongly correlate with memorability in the MemCat dataset, outperforming previous approaches, as well as generalizing to the LaMem dataset. The features encoded by the autoencoders also allow us to interpret what properties affect the memorability of images the most, supporting previous findings regarding across- and within-scene category variation.[1]

Methodology

We explore potential proxies for properties connected to memorability in 2 parts: (1) extracting representations and statistics from state-of-the-art vision encoders and conducting Spearman's correlation analysis and linear regression between model-internal features and memorability; (2) training autoencoders on top of the extracted representations to explore and propose novel setups to investigate the correlation between autoencoder loss and memorability.

Data. We primarily use images and memorability scores from the **MemCat** dataset (Goetschalckx & Wagemans, 2019), which consists of 10,000 images each belonging to one of 5 categories: animal, food, vehicle, landscape, and sports. MemCat includes crowdsourced annotations from participants who viewed image sequences and indicated whether a given image had been shown earlier in the sequence. MemCat includes raw data as well as false-recall corrected memorability scores. We use the corrected score, ranging between 0.14 and 0.98, with a mean of 0.693, standard deviation of 0.137 and median of 0.714. Per category means and standard deviations in MemCat are as follows in increasing order of mean: **Landscape**: $\mu = 0.526$, $\sigma = 0.138$; **Vehicle**: $\mu = 0.695$, $\sigma = 0.093$; **Sports**: $\mu = 0.713$, $\sigma = 0.098$; **Animal**: $\mu = 0.729$, $\sigma = 0.094$; **Food**: $\mu = 0.802$, $\sigma = 0.081$.

We also utilize **LaMem** (Khosla et al., 2015), which is a larger dataset consisting of 58,741 images. The dataset provides 5 random split versions consisting of train-validation-test sets with disjoint images and participants. The authors use a random half of the participants for the train and validation scores, and the other half is used to compute the test scores. As a result, an image can have different scores in different versions. To have a single memorability score, we take the average of the scores for an image in all 5 splits. These aggregated scores range between 0.266 and 1.0, have a mean of 0.756, standard deviation of 0.116, and median of 0.769.

Models. We use the vision encoder from two image-text models, CLIP (Radford et al., 2021) and SigLIP2 (Tschannen et al., 2025), as well as a vision-only model DINOv2 (Oquab et al., 2024).[2] All three vision encoders have 12 layers and are based on the Vision Transformer (ViT; Dosovitskiy et al., 2021), which is the architecture of choice for the visual backbone of many state-of-the-art vision-language models, as well as computer vision models.

One important point to note is that these models were trained with respect to different objectives. CLIP has a contrastive image-text learning scheme, optimizing the alignment between matching images and texts (Radford et al., 2021). SigLIP2, on the other hand, uses a sigmoid-based learning regime when matching images and texts, in addition to captioning, self-distillation and masked prediction objectives (Tschannen et al., 2025). DINOv2 is not exposed to language, and is trained to predict masked patches in images in a self-supervised manner (Oquab et al., 2024). For CLIP and SigLIP2, supervision with the text encoder was intended to enhance the capabilities in the visual modality. In this way, we explore the effects of having vision encoders trained in coordination with language.

Darcet et al. (2024) identify a problem in several vision transformer models including DINOv2, showing the existence of outlier image tokens with very high norms. These tokens, called *registers*, mainly correspond to patches with low information value such as background regions. The authors claim that registers are repurposed to store global image information. Therefore, when we inspect patch uniformity, we also test DINOv3 (Siméoni et al., 2025), which includes extra register tokens to avoid the problem of such artifacts.[3]

Extracting Model-Internal Features

Informed by findings from cognitive science and CV, we investigate a set of model-internal features explained below.

Activations. In ViTs, each image is divided into a grid of

[1] Code and models available at https://github.com/ecekt/image_memorability_memvis

[2] HuggingFace models for CLIP: 'openai/clip-vit-base-patch32', SigLIP2: 'siglip2-base-patch16-224', DINOv2: 'dinov2-base'.

[3] DINOv3-base: 'facebook/dinov3-vitb16-pretrain-lvd1689m', DINOv3-large: 'facebook/dinov3-vitl16-pretrain-lvd1689m'

equal-sized patches. The models encode these patches to form image tokens. Prepended to the image tokens is a token called [CLS]. [CLS] captures a 'summary' of the image in a vector with 768 dimensions; therefore, we utilize its activations to explore correlates of memorability. We extract the [CLS] representation per layer from the vision encoders. We obtain each [CLS] representation's mean activation, maximum activation, and maximum absolute activation for each of the 12 layers, except the embedding layer, as the initial [CLS] embedding is the same for every image. Note that SigLIP2 does not have a [CLS] token and instead uses a multihead attention pooling head at the final layer. For SigLIP2, we report findings using the first token, though not directly comparable, for compatibility with our layer-wise analysis of the models. One can regard these findings as an analysis of a contextualized image token or a potential *register* token that might capture global image information, as explained earlier (Darcet et al., 2024). Our use of [CLS] in what follows includes this first token for SigLIP2. Later, we complement this analysis using the pooled output from SigLIP2 to train our autoencoders.

[CLS] delta. Inspired by the idea that deeper levels of processing lead to more memorability (Craik & Lockhart, 1972), we quantify the magnitude of the change the [CLS] representation goes through, by calculating the cosine distance between [CLS] at adjacent layers. This is a proxy for measuring how levels in image processing relate to memorability.

Attention entropy. As human attention seems to differ based on memorability (Celikkale et al., 2013; Khosla et al., 2015; Mancas & Le Meur, 2013), we are interested in the distribution of the attention applied by the model to the image patches. The [CLS] token captures global image information by paying attention to patches differently. Therefore, we investigate the attention applied by the [CLS] to the image patches. We use the entropy of the attention distribution as a metric that quantifies how spread-out attention is at every layer.

Patch uniformity. As less memorable images tend to have no salient, central object (Khosla et al., 2015), we also investigate image patch tokens. Specifically, we compute how varied the image patches of an image are. For each image patch at a given layer, we calculate its average cosine similarity to the rest of the patches for the same image. The mean over all patches of the image gives the image's patch uniformity score.

Autoencoders

Among prior works, Bagheri and Mohsenzadeh (2025) use reconstruction losses from a VGG16-based autoencoder fine-tuned on all MemCat images for 1 epoch, using batch size 1, and reaches 0.45 Spearman's correlation coefficient with memorability. Lin et al. (2024) also explore a sparse coding approach again using VGG16 and report 0.33 Pearson's correlation on the dataset introduced by Isola et al. (2014).

We experiment with **ViTMAE**, a pretrained masked autoencoder based on ViT, which was trained on ImageNet1k (Deng et al., 2009). We use its base and large versions ('facebook/vit-mae-base' and 'facebook/vit-mae-large'). We compute autoencoder reconstruction loss for images with 75% of image patches masked, as this ratio was shown to be optimal (He et al., 2022). As MemCat includes 5617 images from ImageNet, we also conduct experiments where we discard this subset and report the correlation for the rest of MemCat ($N = 4383$).

Additionally, we train **our own sparse autoencoders** that take in image representations from CLIP and SigLIP2.[4] For the CLIP-based autoencoder, we use the [CLS] from the last layer of the vision encoder, while for SigLIP2, we opt for the pooled output as mentioned previously.

Our autoencoder consists of a linear layer, W_{enc}, that projects an image representation, y, to a vector of dimensionality 100, ReLU, and a second linear layer, W_{dec}, that projects back to the dimensions of the image representation:

$$\hat{y} = \mathbf{W}_{dec}(ReLU(\mathbf{W}_{enc}y + b_{enc})) + b_{dec} \quad (1)$$

We use Mean-Squared Error Loss (MSE) as the autoencoder's reconstruction loss and a weighted sparsity penalty (L1 loss) to encourage the model to create sparse latent representations z, constraining the sum of absolute values in the bottleneck of the autoencoder (to aid interpretability, not correlation). In our memorability correlation analyses, we only use the reconstruction loss.

$$\mathcal{L} = \underbrace{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}_{\text{MSE reconstruction loss}} + \underbrace{\lambda \sum_{j=1}^{M}|z_j|}_{\text{sparsity loss}} \quad (2)$$

We use 80% of MemCat for training and 20% for validation, after shuffling the whole dataset. Following a hyperparameter search, we train each setup for 5 epochs with batch size 4, learning rate 5e-4, sparsity weight λ = 1e-5, and hidden dimensions 100 (best memorability correlations among 50, 100, 200, 500, 768, 1000). Before training, we apply z-score normalization to image representations per vision encoder.

Results

Model-Internal Features

Figure 2 depicts the three encoders' **activations** in terms of how much they correlate with memorability over the layers. CLIP's mean activation starts from a moderate correlation and displays a trend that goes in the negative direction. DINOv2 Max Abs is a good correlate but fluctuates. SigLIP2's activations are weak correlates, likely because they originated from the first patch. The plot on the left in Figure 3 depicts that, for CLIP and DINOv2, the deltas show an oscillating correlation to memorability, with the correlation increasing in specific layer intervals and subsequently decreasing, unlike SigLIP2, which has a more stable pattern.

The middle plot in Figure 3 illustrates the correlation between **attention entropy** and memorability. More visibly for DINOv2 and SigLIP2 than for CLIP, we see an increasing

[4] Using DINOv2 representations as input to our autoencoders yielded only weak correlation (0.26), suggesting that language supervision helps capture memorability-related semantics better.

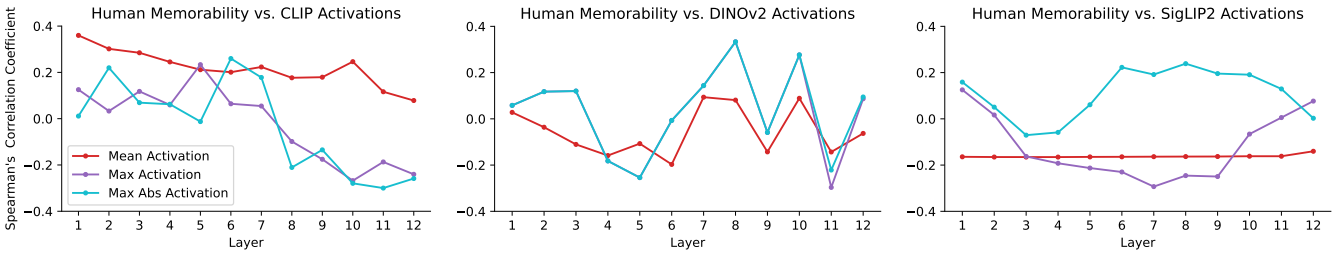

Figure 2: Correlation coefficients between human memorability and features of [CLS] activations over the layers of vision encoders (except layer 0, because it is the same for all images). *Left:* CLIP, *Middle:* DINOv2, *Right:* SigLIP2.

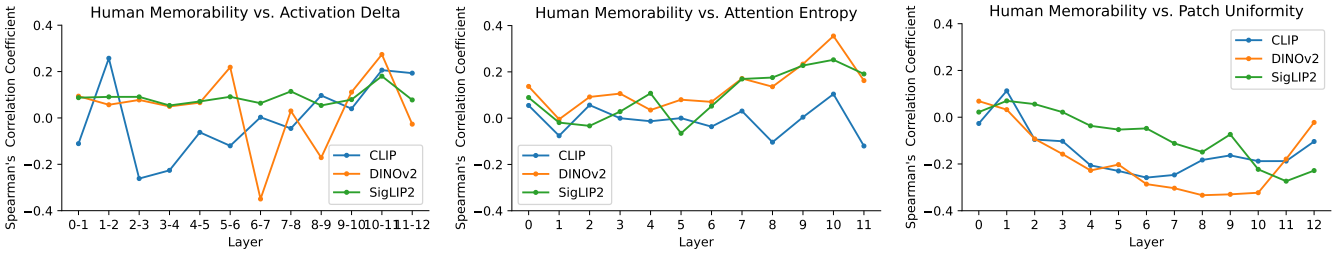

Figure 3: Correlation between image memorability and *Left:* the cosine distance between adjacent layers' [CLS] activations (delta), *Middle:* entropy of the attention applied by the [CLS] token to the image tokens, and *Right:* patch uniformity.

trend towards the later layers where the attention distribution reflects information correlating with memorability.

The plot on the right in Figure 3 reveals the negative correlation in the middle-to-later layers between **patch uniformity** and memorability. The patch uniformity scores per category are as follows in increasing order: **Vehicle**: 0.191, $\sigma = 0.125$; **Food**: 0.236, $\sigma = 0.114$; **Sports**: 0.257, $\sigma = 0.158$; **Animal**: 0.265, $\sigma = 0.151$; **Landscape**: 0.354, $\sigma = 0.138$. In MemCat, landscape images have the lowest memorability as well as higher patch uniformity. This is in line with previous findings about images with no salient and central objects having lower memorability (Khosla et al., 2015).

Checking different versions of the DINOv3 model, we see that register tokens make patches more meaningful; see Figure 4. DINOv3-Base model has a stronger correlation compared to DINOv2-Base in the first half of the layers, potentially due to the help of the register tokens. DINOv3-Large model yields weaker correlations in the first half, suggesting that large models may be worse at predicting human signals. We continue using DINOv2 for comparability as the other models do not have register tokens.

We also fit **linear regression models** to investigate the relationship between model-internal features and memorability as the dependent variable in a single model. To exemplify, in Table 1, we report the outcomes for DINOv2's layer 10 features, as its activations, attention entropy, and patch uniformity show consistently stronger correlations with memorability. The results suggest that activation mean and attention entropy are strong, positive predictors of memorability. On

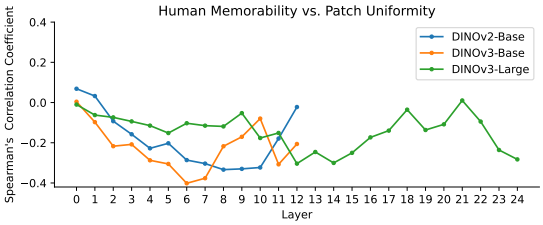

Figure 4: Correlation between patch uniformity and memorability in different versions of the DINO model.

the other hand, as patch uniformity increases, memorability decreases. These findings indicate that, although there does not seem to be a clear pattern among the models and across the layers, there exist model-internal features that correlate with memorability to some extent ($R^2 = .238$, for Table 1).

The variation in these results could be due to the architectures, training sets and training objectives of each model. However, compared to metrics based on activations, the results on two measures reflecting information distribution, i.e., **attention entropy** and **patch uniformity** show a clearer pattern across layers, which seem to be valid for all the models.

Autoencoder Reconstruction Loss

Table 2 reports the correlation between memorability and reconstruction losses obtained from **ViTMAE** models. The models yield very weak correlations, a result that resonates with findings in the linguistic modality: *larger* pre-trained lan-

Table 1: Linear regression model for DINOv2 layer 10. *** indicates significance at $p < .001$. Independent variables were z-score normalized.

	Coef	Std Err	Z
Intercept	0.6932	0.001	577.806***
Activation max	0.0144	0.001	23.556***
Activation mean	0.0264	0.001	20.171***
Activation max abs	0.0144	0.001	23.556***
Patch uniformity	-0.0170	0.002	-10.574***
Attention entropy	0.0457	0.002	27.745***

guage models cannot account for human reading signals and diverge from human-like surprisal estimations (Oh & Schuler, 2023). This could be because these models are also trained on large amounts of data, what is unique or surprising for humans is not so for such models. The results are slightly higher, yet still very weak, when we discard the ImageNet data, indicating that these autoencoders cannot be used as off-the-shelf proxies for memorability prediction.

Table 2: Correlation between image memorability and ViTMAE loss. Very weak correlation, although $p < .001$.

	ViTMAE - base	ViTMAE - large
Full MemCat	0.073***	0.056***
No ImageNet	0.118***	0.097***

On the other hand, the losses from **our sparse autoencoders** show strong positive correlations with memorability; see Fig. 5. CLIP generally correlates better than SigLIP2. At epoch 3, it reaches 0.57 in one run, outperforming previous work using autoencoders (Bagheri & Mohsenzadeh, 2025; Lin et al., 2024). As done by Bagheri and Mohsenzadeh (2025), to emulate human data collection settings, we also experiment with training on the whole MemCat instead of dividing it into splits. We test a single-exposure setup similar to theirs (batch size 1 and learning rate 1e-4). Our best-correlating setup is a SigLIP2-based autoencoder, reaching **0.58** at epoch 5, after which it drops. Using this autoencoder, we obtain reconstruction losses for images. We depict illustrative examples in Figure 1: the most memorable image in MemCat incurs a loss of 294.954, while the least memorable one incurs 219.022. This is in line with findings showing that higher reconstruction loss is paired with higher memorability.

Generalization. Using the *aggregated* LaMem scores, we test whether our best autoencoder generalizes to images from a different dataset. The best CLIP-based autoencoder yields a correlation of 0.449, and the SigLIP2-based one yields 0.423 correlation, both $p < .001$. We then employ the 5 *test* set splits of LaMem. The best CLIP-based autoencoder reaches 0.414 correlation, and the SigLIP2-based one reaches 0.388 correlation, both $p < .001$. The MemNet model proposed together with LaMem (Khosla et al., 2015), which is a CNN finetuned to predict memorability scores, achieves 0.64 correlation. Our results show that our autoencoder losses can generalize and significantly correlate with the memorability scores of novel images from a larger dataset.

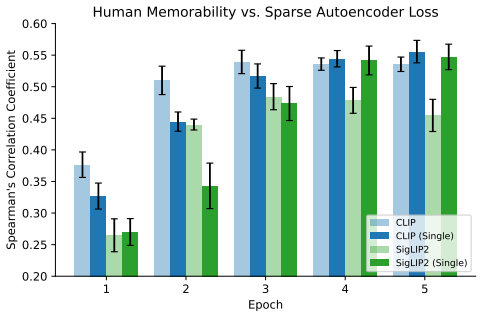

Figure 5: Correlation between reconstruction losses obtained from our sparse autoencoders and memorability on the whole dataset. Error bars indicate standard deviation for 5 random seeds. 'Single' is when we train on the whole MemCat with a batch size of 1, the others follow our initial setting with splits.

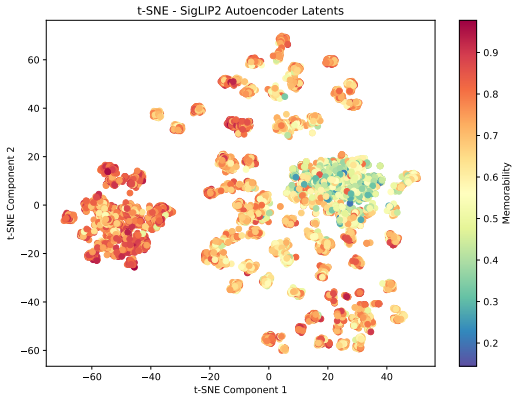

Figure 6: t-SNE visualization of the latent representations from the encoder of the best-correlating SigLIP2-based autoencoder. Red - higher memorability, blue - lower.

Latent Representation Space. To have a better picture of the best-correlating autoencoder, we focus on the latent representations extracted from its *encoder*. We find that latents of high-memorability images have larger activations, with mean absolute activations correlating significantly with memorability (0.41 for CLIP, 0.32 for SigLIP2, $p < .001$). In Figure 6, we illustrate the spatial distribution of the latents, color-coded by their memorability scores. Many clusters consisting of middle and high memorability are spread around the plot. However, the least memorable images seem to be clustered separately. Virtually all of these correspond to landscape images. Figure 7 depicts the same plot annotated with image categories, which shows that the images of different categories are well separated in the autoencoder's latent space.

Within categories, there is some variation in memorability, as previously observed (Bylinskii et al., 2015). As an example, we zoom into the representations of animal images. In the an-

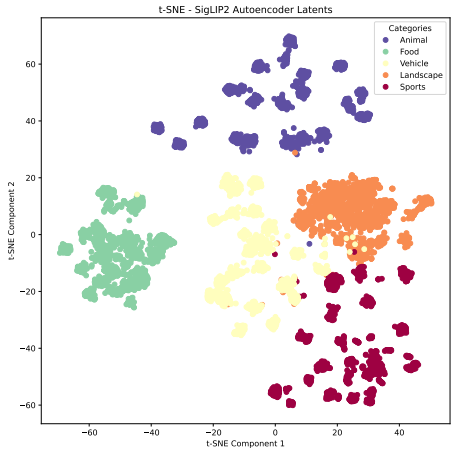

Figure 7: t-SNE visualization of the latent representations from the encoder of the best-correlating SigLIP2-based autoencoder, annotated with image categories.

Figure 8: Images that activate features 97 (left) and 54 (right) the most. These features have the strongest **positive** correlation with memorability ($\rho_{97} = 0.28, \rho_{54} = 0.24$; $p < .001$).

imal cluster, there clearly exist images with low memorability, especially as the image gets closer to the landscape cluster. To check this quantitatively, we compute a mean landscape representation by taking the average of the SigLIP2 representations of images belonging to the landscape category. We find that the closer an animal image's SigLIP2 representations to the mean landscape representation in terms of cosine similarity, the lower its memorability score ($r = -0.452, p < .001$). Additionally, cosine similarity between autoencoder latents of animal images and the mean latent representation of landscape images yields similar results ($r = -0.314, p < .001$).

Interpreting Latent Dimensions. Inspired by studies on the mechanistic interpretability of latent spaces of sparse autoencoders for representations from vision-language models (Bhalla et al., 2024; Olson et al., 2025; Papadimitriou et al., 2025; Stevens et al., 2025), we also look deeper into dimensions of the latent representations from the autoencoder. To assess the dimensions, i.e., features, in terms of their relation to memorability, we extract latent representations of all images. Then, we create 100 lists of 10,000 items corresponding to each dimension's activation per image. Afterwards, we compute the correlation between each list and the memorability scores. In this way, we identify the features that significantly correlate with memorability.

Then, we find the images that activate these features the most. We depict the top images for the features with the strongest positive and negative correlations with memorability in Figures 8 and 9, respectively. Features that are activated by images containing beaches, forests, and mountains go together with the lowest memorability scores. On the other hand, features activated by images with objects such as hamsters, limousines, and butternut squash correlate with higher memorability. We observe the influence of semantic categories on memorability (Isola et al., 2014; Khosla et al., 2015; Kramer et al., 2023) also in the latents of our autoencoder.

Figure 9: Images that activate features 63 (left) and 34 (right) the most. The features have the strongest **negative** correlation with memorability ($\rho_{63} = -0.49, \rho_{34} = -0.48$; $p < .001$).

General Discussion

Image memorability is a multifaceted phenomenon influenced by intrinsic and extrinsic factors, and what makes an image memorable has been studied in cognitive science and computer vision. In this paper, for the first time, we investigated correlates of image memorability in a set of model-internal features obtained from transformer-based vision encoders, which were not trained to predict memorability. We find features that are reasonably good predictors of image memorability. Significantly, reconstruction loss from our sparse autoencoders is a strong predictor. Our findings resonated with those by Lin et al. (2024) and Bagheri and Mohsenzadeh (2025), confirming that difficult-to-reconstruct images are more memorable. Combined with our results on the negative correlation between patch uniformity and memorability, this points to image complexity and distinctiveness as common drivers of both reconstruction difficulty and human memorability. The outcome regarding the higher activations in the latent representations for memorable images also mirrors the findings about stronger neural responses to memorable images (Jaegle et al., 2019). We observed across-category and within-category variation in memorability, as reflected in their representations and losses (Bylinskii et al., 2015; Goetschalckx & Wagemans, 2019; Isola et al., 2014; Khosla et al., 2015). We also identified and interpreted features that correlate strongly with memorability, which can be further studied to gain insight into fine-grained attributes making an image memorable. Future work can explore a wider range of models, datasets, and other factors linked to memorability such as emotions and context (Bylinskii et al., 2022). In addition to providing pointers to correlates that would inform models of visual memory, there is potential for transferring our approaches to study processes related to the memorability of other modalities, such as language.

Acknowledgments

The research reported in this paper was supported by the European Research Council (ERC), grant 101088098 - MEMLANG. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. An initial version of this work was presented at the ICCV 2025 Workshop MemVis: The 1st Workshop on Memory and Vision. We thank the reviewers and the audience for their valuable comments.

References

Bagheri, E., & Mohsenzadeh, Y. (2025). Modeling visual memorability assessment with autoencoders reveals characteristics of memorable images. https://arxiv.org/abs/2410.15235

Bainbridge, W. A. (2019). Chapter one - memorability: How what we see influences what we remember. In K. D. Federmeier & D. M. Beck (Eds.), *Knowledge and vision* (pp. 1–27, Vol. 70). Academic Press. https://doi.org/https://doi.org/10.1016/bs.plm.2019.02.001

Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F., & Lakkaraju, H. (2024). Interpreting CLIP with sparse linear concept embeddings (spliCE). *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=7UyBKTFrtd

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Bylinskii, Z., Goetschalckx, L., Newman, A., & Oliva, A. (2022). Memorability: An image-computable measure of information utility. In B. Ionescu, W. A. Bainbridge, & N. Murray (Eds.), *Human perception of visual information: Psychological and computational perspectives* (pp. 207–239). Springer International Publishing. https://doi.org/10.1007/978-3-030-81465-6_8

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability [Computational Models of Visual Attention]. *Vision Research*, *116*, 165–178. https://doi.org/https://doi.org/10.1016/j.visres.2015.03.005

Celikkale, B., Erdem, A., & Erdem, E. (2013). Visual attention-driven spatial pooling for image memorability. *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 976–983. https://doi.org/10.1109/CVPRW.2013.142

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/https://doi.org/10.1016/S0022-5371(72)80001-X

Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2024). Sparse autoencoders find highly interpretable features in language models. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=F76bwRSLeK

Darcet, T., Oquab, M., Mairal, J., & Bojanowski, P. (2024). Vision transformers need registers. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=2dnO3LLiJ1

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy models of superposition. https://arxiv.org/abs/2209.10652

Goetschalckx, L., & Wagemans, J. (2019). Memcat: A new category-based image set quantified on memorability. *PeerJ*, *7*, e8169.

Hagen, T., & Espeseth, T. (2023). Image memorability prediction with vision transformers. https://arxiv.org/abs/2301.08647

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners [Publisher Copyright: © 2022 IEEE.; 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022 ; Conference date: 19-06-2022 Through 24-06-2022]. *Proceedings - 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). Understanding the intrinsic memorability of images. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 24). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2011/file/286674e3082feb7e5afb92777e48821f-Paper.pdf

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *CVPR 2011*, 145–152.

Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability (T. E. Behrens, T. Pasternak, & E. A. Buffalo, Eds.). *eLife*, *8*, e47596. https://doi.org/10.7554/eLife.47596

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. *International Conference on Computer Vision (ICCV)*.

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science Advances*, *9*(17), eadd2981. https://doi.org/10.1126/sciadv.add2981

Lin, Q., Li, Z., Lafferty, J., & Yildirim, I. (2024). Images with harder-to-reconstruct visual representations leave stronger memory traces. *Nature Human Behaviour*, *8*, 1–12. https://doi.org/10.1038/s41562-024-01870-3

Mancas, M., & Le Meur, O. (2013). Memorability of natural scenes: The role of attention. *2013 IEEE International Conference on Image Processing*, 196–200. https://doi.org/10.1109/ICIP.2013.6738041

Needell, C., & Bainbridge, W. (2022). Embracing new techniques in deep learning for estimating image memorability. *Computational Brain & Behavior*, *5*. https://doi.org/10.1007/s42113-022-00126-5

Oh, B.-D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350. https://doi.org/10.1162/tacl_a_00548

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, *37*(23), 3311–3325. https://doi.org/https://doi.org/10.1016/S0042-6989(97)00169-7

Olson, M. L., Hinck, M., Ratzlaff, N., Li, C., Howard, P., Lal, V., & Tseng, S.-Y. (2025). Probing the representational power of sparse autoencoders in vision models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 6167–6177.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., ... Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision [Featured Certification]. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=a68SUt6zFt

Papadimitriou, I., Su, H., Fel, T., Kakade, S. M., & Gil, S. (2025). Interpreting the linear structure of vision-language model embedding spaces. *Second Conference on Language Modeling*. https://openreview.net/forum?id=qPsmGjpq1j

Perera, S., Tal, A., & Zelnik-Manor, L. (2019). Is image memorability prediction solved? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Praveen, A., Noorwali, A., Samiayya, D., Khan, M., Vincent, D., Bashir, A., & Alagupandi, V. (2021). Resmem-net: Memory based deep cnn for image memorability estimation. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.767

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021, 18–24 Jul). Learning transferable visual models from natural language supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (pp. 8748–8763, Vol. 139). PMLR. https://proceedings.mlr.press/v139/radford21a.html

Rust, N. C., & Mehrpour, V. (2020). Understanding image memorability. *Trends in Cognitive Sciences*, *24*(7), 557–568. https://doi.org/https://doi.org/10.1016/j.tics.2020.04.001

Sadana, A., Thakur, N., Poria, N., Anand, A., & Seeja, K. R. (2024). Comprehensive literature survey on deep learning used in image memorability prediction and modification. In A. E. Hassanien, O. Castillo, S. Anand, & A. Jaiswal (Eds.), *International conference on innovative computing and communications* (pp. 113–123). Springer Nature Singapore.

Sharkey, L., Braun, D., & Millidge, B. (2023). Taking features out of superposition with sparse autoencoders. 2022. *URL https://www. alignmentforum. org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition*.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., ... Bojanowski, P. (2025). DINOv3. https://arxiv.org/abs/2508.10104

Squalli-Houssaini, H., Duong, N. Q. K., Gwenaelle, M., & Demarty, C.-H. (2018). Deep learning for predicting image memorability. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2371–2375. https://doi.org/10.1109/ICASSP.2018.8462292

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207–222. https://doi.org/10.1080/14640747308400340

Stevens, S., Chao, W.-L., Berger-Wolf, T., & Su, Y. (2025). Interpretable and testable vision features via sparse autoencoders. https://arxiv.org/abs/2502.06755

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., & Zhai, X. (2025). Siglip 2: Multilingual vision-language en-

coders with improved semantic understanding, localization, and dense features. https://arxiv.org/abs/2502.14786

Vogelsang, D. A., & Heilbron, M. (2025). Representational magnitude as a geometric signature of image and word memorability. *bioRxiv*. https://doi.org/10.1101/2025.09.01.673067