Humans Remember People; Models Remember Pixels: Exposing Blind Spots in Visual Memorability Prediction

Benjamin TenWolde and Virginia de Sa
Department of Cognitive Science, University of California, San Diego
btenwolde@ucsd.edu, desa@ucsd.edu

Abstract

Memorability is highly consistent across viewers, yet the models that predict it remain poorly understood. We analyze two architectures (ResMem, ViTMem) using attribution analysis, synthetic probes, and counterfactual interventions on THINGS. Both show little sensitivity to person presence (r=0.145 for humans vs. r=−0.068 for ResMem) and instead respond to chromatic and frequency statistics not linked to behavioral evidence. Gradient-guided pixel perturbations (ϵ=0.005; $L_\infty \approx 1.3/255$) shift predictions by ∼0.56, more than eight times the person-presence effect, indicating that these models rely on pixel-level statistics rather than the scene-level processing that underlies human memorability.

Introduction

Memorability is highly consistent across observers (Goetschalckx & Wagemans, 2019; Isola et al., 2011) and has been framed as a stimulus-intrinsic property tied to processing efficiency (Bainbridge et al., 2025). Modern predictors such as ResMem (Needell & Bainbridge, 2022) and ViTMem (Hagen & Espeseth, 2023) achieve rank correlations above 0.6 with human annotations, yet remain largely opaque. Neuroscience evidence suggests this opacity may mask mismatches with human processing: intracranial EEG shows memorability encoded in prefrontal cortex dissociably from visual and semantic coding (Wang et al., 2026), memorable images elicit facilitated N300/N400 ERPs (Deng et al., 2024), and memorability depends on surrounding context (Bylinskii et al., 2015) and interference dynamics (Morales-Calva et al., 2025). Current models account for none of these factors. We combine attribution analysis, synthetic probes, and counterfactual interventions on THINGS (Hebart et al., 2019) to identify where these models diverge from human memorability.

Method

Models and target. ResMem (Needell & Bainbridge, 2022) pairs an ImageNet-pretrained ResNet-152 with a regression head fine-tuned on LaMem (Khosla et al., 2015) (∼60k images). ViTMem (Hagen & Espeseth, 2023) fits a regression head on frozen ViT-B/16 features trained on LaMem.

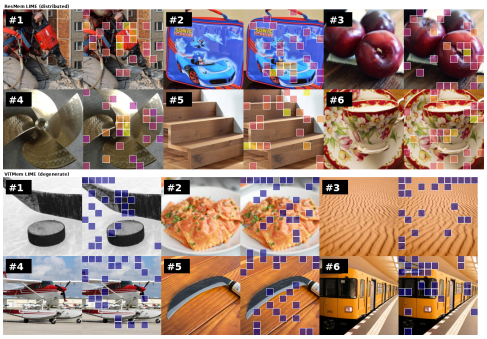

Figure 1: LIME overlays on THINGS images. **Top:** ResMem distributes importance broadly. **Bottom:** ViTMem produces near-uniform attributions, indicating no spatial selectivity.

Attribution analysis. For CNNs we compute Grad-CAM (Selvaraju et al., 2017); for transformers we use attention weights combined with input-gradient maps. Both are cross-checked with LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) over 144 segments (512 images per model). We measure how concentrated each attribution map is by the number of segments needed to account for 50% of attribution mass. Progressive occlusion (5% steps) tests whether the highlighted regions actually drive predictions.

Synthetic probes and interventions. We generate 270 procedural patterns spanning the HSV color space and multiple texture families (Fig. 3). Counterfactual interventions include frequency filters, desaturation, and gradient-guided pixel perturbations (PGD, 10 steps, ϵ=0.005; pixel $L_\infty \approx 1.3/255$) on 512 THINGS images; if such small pixel changes control model output, the learned discriminants are not human-aligned. Person presence is estimated using OpenCLIP ViT-B/16 (Cherti et al., 2023), a vision-language model that scores each image for person content without task-specific training.

Results

Attribution maps reveal contrasting strategies (Fig. 1). ResMem LIME attributions are broadly distributed, with ≈33 of 144 segments covering 50% of attribution mass;

Extended abstract presented at the *9th Conference on Cognitive Computational Neuroscience*, New York, NY, USA, 2026.

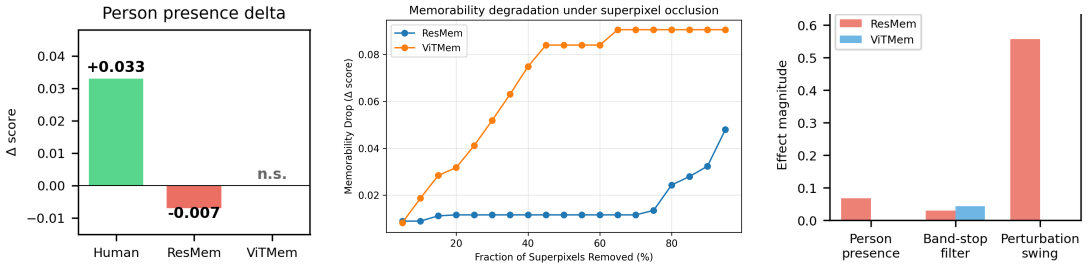

Figure 2: **(A)** Person presence (recognition-accuracy Δ, distinct from the correlation r in Table 1): humans +0.033, models reversed or flat. **(B)** Occlusion reveals divergent degradation. **(C)** Pixel perturbations (ϵ=0.005, $L_\infty \approx 1.3/255$) swing ResMem by 0.56, exceeding all other effects.

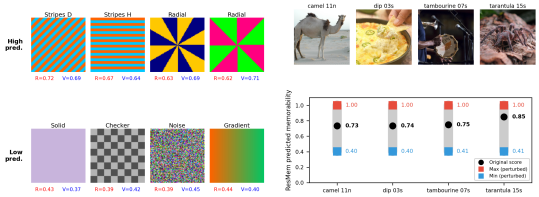

Figure 3: **Left:** Synthetic probes: models respond to chromatic and frequency statistics, not semantic content (R=ResMem, V=ViTMem). **Right:** Small perturbations (ϵ=0.005, $L_\infty \approx 1.3/255$) control the full prediction range, confirming pixel-level rather than scene-level discriminants.

ViTMem attributions are near-uniform, spread across all 144 segments, consistent with global rather than local reliance. Progressive occlusion (Fig. 2B) confirms this: ResMem is stable under occlusion (drop of 0.012 at 30%, reaching 0.048 at 95%) while ViTMem degrades steadily (0.052 at 30%, plateauing at $\sim$0.091 by 65%). This parallels the texture bias in classification CNNs (Geirhos et al., 2019), and may be more pronounced because memorability regression offers no object category to localize, leaving no pressure to attend to specific regions.

Synthetic probes reveal chromatic biases: ResMem peaks in yellow–green; ViTMem drifts toward cyan–blue. Neither matches documented human chromatic preferences.

The clearest divergence is semantic. On THINGS memorability (Kramer et al., 2023), person-containing images are +3.3 pp more memorable to humans (recognition-accuracy Δ; rank correlation r=0.145, $p<0.001$, n=11,998); ResMem correlates *negatively* ($r=-0.068$) and ViTMem shows no reliable association (Fig. 2A). Removing mid-range spatial frequencies (band-stop filtering) shifts predictions by -0.030 (ResMem) and -0.044 (ViTMem), and retaining only high frequencies *increases* ViTMem predictions by +0.038. Intracranial EEG shows memorability arises from prefrontal representations that integrate visual and semantic information over time (Wang et al., 2026); feedforward pixel-to-score regression cannot learn such representations. A perturbation probe confirms this: at ϵ=0.005 ($\approx 1.3/255$ pixel L_∞) gradient-guided pixel changes shift predictions by $\sim$0.56 (above the models' $\approx$0.4 natural range on THINGS), more than eight times the person-presence effect, from a visually subtle change. As in adversarial robustness research on classifiers (Ilyas et al., 2019), the discriminants these models rely on are pixel-level statistics without perceptual relevance to human memory (Table 1).

Table 1: Human–model dissociations.

Diagnostic	Human	ResMem	ViTMem
Person presence r	+0.145	-0.068	n/a
Band-stop Δ	—	-0.030	-0.044
High-pass Δ	—	+0.004	+0.038
50% coverage (of 144)	—	33	144
Perturbation swing	0	$\sim$0.56	$\sim$0

Conclusion

Humans are sensitive to semantic content, especially people, while both architectures respond to chromatic, frequency, and pixel-level statistics not linked to behavioral evidence. That small perturbations shift predictions across the full range confirms these discriminants are not human-aligned. Growing evidence shows memorability depends on context and interference dynamics (Bylinskii et al., 2015; Morales-Calva et al., 2025): an image's memorability changes with the surrounding sequence, yet current models assign a fixed score per image in isolation. Standard repeat-detection paradigms with large inter-item lags may themselves underestimate these contextual effects, suggesting that both the models and the benchmarks used to train them underrepresent the relational nature of human memory.

Acknowledgments

Claude Code (model Opus 4.8) was used to generate the code for this project, and to write the routines for the diagrams. All analyses, claims, and conclusions were designed and verified by the authors.

References

Bainbridge, W. A., Walther, D. B., Fukuda, K., & Goetschalckx, L. (2025). Memorability of visual stimuli and the role of processing efficiency [Online ahead of print]. *Nature Reviews Psychology*, *5*, 47–58. https://doi.org/10.1038/s44159-025-00512-3.

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

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., & Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829. https://doi.org/10.1109/cvpr52729.2023.00276.

Deng, W., Beck, D. M., & Federmeier, K. D. (2024). Image memorability is linked to facilitated perceptual and semantic processing. *Imaging Neuroscience*, *2*. https://doi.org/10.1162/imag_a_00281.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *Proceedings of the International Conference on Learning Representations*.

Goetschalckx, L., & Wagemans, J. (2019). MemCat: A new category-based image set quantified on memorability. *PeerJ*, *7*, e8169. https://doi.org/10.7717/peerj.8169.

Hagen, T., & Espeseth, T. (2023). Image memorability prediction with vision transformers. https://doi.org/10.48550/arXiv.2301.08647.

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. https://doi.org/10.1371/journal.pone.0223792.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*, *32*.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 145–152. https://doi.org/10.1109/cvpr.2011.5995721.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. *Proceedings of the IEEE International Conference on Computer Vision*, 2390–2398. https://doi.org/10.1109/iccv.2015.275.

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.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4768–4777.

Morales-Calva, F., Velgekar, A., Sekili, M., & Leal, S. L. (2025). Image memorability depends on interference in memory. *Scientific Reports*, *15*, 36597. https://doi.org/10.1038/s41598-025-21937-z.

Needell, C. D., & Bainbridge, W. A. (2022). Embracing new techniques in deep learning for estimating image memorability. *Computational Brain & Behavior*, *5*, 168–184. https://doi.org/10.1007/s42113-022-00126-5.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. https://doi.org/10.1109/iccv.2017.74.

Wang, Y., Brunner, P., Willie, J. T., Cao, R., & Wang, S. (2026). Characterization of the spatiotemporal representations of visual, semantic, and memorability features in the human brain [Published online 2 January 2026]. *PLOS Biology*, *24*(1), e3003614. https://doi.org/10.1371/journal.pbio.3003614.