UC San Diego

Humans Remember People; Models Remember Pixels

Evidence from person, texture, composition, and perturbation probes

Benjamin TenWolde (btenwolde@ucsd.edu) · Virginia de Sa (desa@ucsd.edu) · UC San Diego

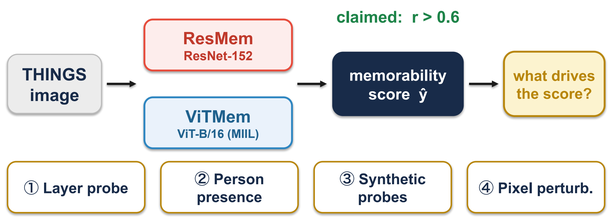
Abstract

Image memorability, the probability that an image is later remembered, is highly consistent across observers. Deep models (ResMem, ViTMem) reach rank correlations above 0.6 on benchmark datasets, interpreted as evidence they capture visual factors associated with human memory.

We evaluate this interpretation across THINGS, MemCat, FIGRIM, and synthetic stimuli. The models often agree more strongly with each other than with human judgments, respond to person presence without consistently ranking person-containing images as humans do, assign high scores to synthetic textures, and remain comparatively stable after scene composition is disrupted. These results suggest the predictors rely heavily on low-level image statistics, with limited evidence that they capture the semantic factors underlying human memorability.

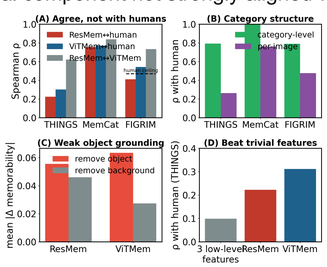
Method

ResMem = ResNet-152 backbone (ImageNet-1k) with a memorability regression head trained on **LaMem**. **ViTMem** = ViT-B/16 (**ImageNet-21k**, fine-tuned on ImageNet-1k) with a head trained on LaMem. We evaluate both without additional training, out of distribution on **THINGS**, **MemCat**, and **FIGRIM**, plus synthetic and perturbation-based probes; person presence is estimated with OpenCLIP.



1. Model Agreement Exceeds Human Alignment

ResMem and ViTMem correlate with human memorability on MemCat (0.77) and FIGRIM (0.54), but the two models often **correlate more strongly with each other (0.62–0.84)** than either does with human judgments; on THINGS, model-human correlations are lower (0.23–0.30). Layer-wise linear readout to human memorability peaks at about ρ=0.32, indicating a shared representational component not strongly aligned with human memorability.

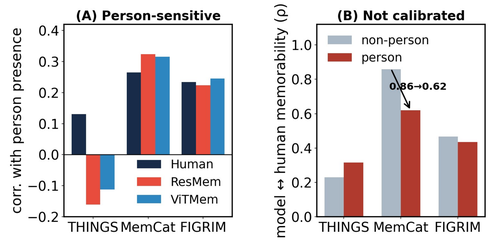


(A) model-model agreement exceeds model-human; (B) alignment is largely category-level; (C) limited object grounding (background contributes); (D) predictions exceed a simple low-level baseline.

Caveat: the datasets are not directly comparable, so correlations are read within each dataset; all interventions are performed within-dataset.

2. Sensitivity to Person Presence Does Not Ensure Human Alignment

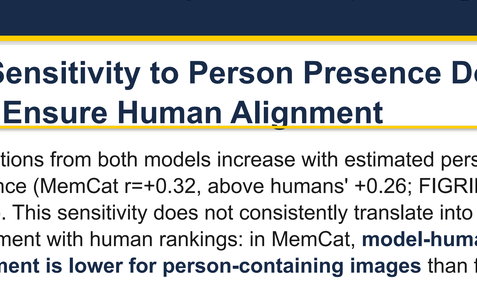
Predictions from both models increase with estimated person presence (MemCat r=+0.32, above humans' +0.26; FIGRIM +0.22). This sensitivity does not consistently translate into agreement with human rankings: in MemCat, **model-human alignment is lower for person-containing images** than for images without people (ρ 0.62 vs 0.86). The models detect person-related features, but their rankings of person-containing images are less aligned with human judgments. (THINGS, an object set with few person images, is an exception.)



(A) predictions increase with person presence; (B) model-human alignment is lower for person-containing images.

3. Synthetic Textures Receive High Memorability Predictions

Checkerboards, stripes, and uniform color fields receive scores comparable to real object images (0.87/0.81 vs 0.85/0.84). Because these stimuli contain no identifiable objects or scenes, the result demonstrates that color and texture statistics are sufficient to produce high predicted memorability in these models.



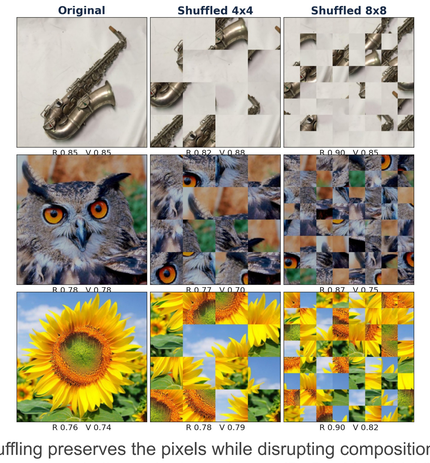
Synthetic stimuli with each model's score (R=ResMem, V=ViTMem).

Additional low-contrast, low-saturation synthetic textures, containing no objects, also receive scores near the upper end of the model range (0.92–0.99).

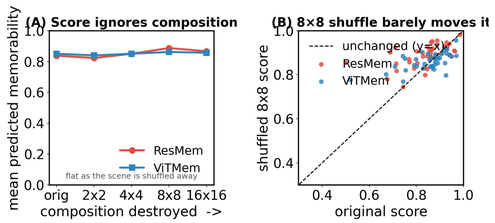
Synthetic low-contrast, low-saturation textures scored near the top of the model range.

4. Predictions Are Weakly Sensitive to Spatial Shuffling

We shuffle each image into spatial tiles while preserving its pixel values and global color statistics. At an 8×8 shuffle, the mean prediction **changes by about 0.05**, while the backbone representation changes substantially: the ViTMem feature moves roughly 65% of the way toward a different image (ResMem ~81%). This indicates the memorability readout is considerably less sensitive to compositional disruption than the underlying representation.



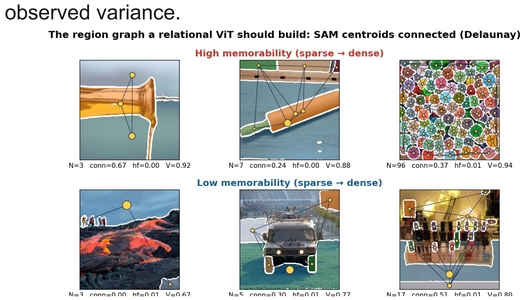
Spatial shuffling preserves the pixels while disrupting composition; shuffled tiles receive scores similar to the intact image (R=ResMem, V=ViTMem).



(A) mean prediction changes little across shuffle granularities; (B) per image, an 8×8 shuffle changes the score by about 0.05.

Memorability Is Weakly Associated with Scene-Graph Density

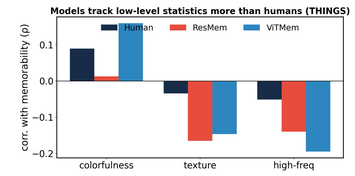
We represent each image using SAM segment centroids connected by Delaunay triangulation, a coarse measure of relational scene structure. High- and low-memorability images span similar ranges of graph density; graph size is weakly correlated with ViTMem predictions (ρ=0.14) and human memorability (0.11), so this measure explains little of the observed variance.



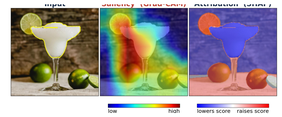
High- (top) and low-memorability (bottom) images span similar ranges of graph density; predictions are weakly associated with graph size.

Model Predictions Are More Strongly Associated with Texture

On THINGS, model predictions are more strongly associated with **edge density and high-frequency energy** than human memorability judgments are (model ρ≈−0.15 to −0.20 vs human ≈−0.04), roughly three to four times larger in magnitude. This suggests the models overweigh texture and frequency statistics relative to human observers.



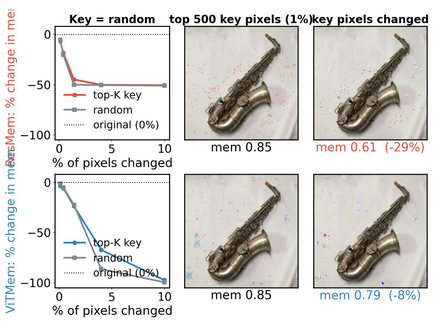
Model predictions correlate with edge density and high-frequency energy more strongly than human judgments do.



Grad-CAM highlights the object, but SHAP attributions for the score are often outside it and can be negative within it; localization does not imply the object drives the prediction.

5. Attribution-Ranked Pixels Do Not Outperform Random Pixels

Perturbing the pixels ranked most important by the tested attribution methods changes the score **no more than an equal number of random pixels**. Predictions begin to shift after about 3% of pixels are modified. Under this perturbation procedure, the score appears to depend on distributed image statistics rather than a small set of localized pixels.



Attribution-ranked pixels shift the score no more than an equal number of random pixels; about 3% of pixels must change before it moves. (ResMem plateaus near its ~0.32 output floor.)

Conclusion

Across multiple datasets and controlled probes, ResMem and ViTMem show substantial sensitivity to color, texture, frequency, and person-related cues. Their predictions remain comparatively stable under compositional disruption and can be high for synthetic stimuli without recognizable content. These findings suggest current predictors rely heavily on low-level and category-associated image statistics. **Better alignment with human memorability may require representations that explicitly capture objects, relations, and semantic scene structure.**