Category-Aware Fine-Tuning and Cross-Age Transferability in Image Memorability Prediction

Elham Bagheri[1,2], **Johann Cardenas**[2], **Yalda Mohsenzadeh**[1,2*]

[1] Vector Institute for Artificial Intelligence, Toronto, ON M5G 0C6, Canada
[2] Department of Computer Science, Western University, London, ON N6A 3K7, Canada
[*] ymohsenz@uwo.ca

Abstract

Image memorability is highly consistent across observers, yet current vision models achieve only moderate accuracy and remain below human consistency. We study two questions: (i) whether making semantic category structure explicit during training improves prediction, and (ii) whether adult-trained predictors transfer to adolescents, and whether any gains from category-specific adaptation generalize across observers of different age. We compare a mixed-category model (*All*) with per-category fine-tuning (*CatFT*) for two pretrained backbones, MemNet (AlexNet-based CNN) and ViT-B/16 (Vision Transformer), each fine-tuned on MEMCAT under *All* and *CatFT*. Adult-trained models are evaluated on MEMOIR (adolescent labels) without additional training to assess transfer, and Grad-CAM is used to examine which regions drive predictions on the best model. On adults, category-aware training increases Spearman's ρ for both backbones (ViT-B/16: $0.548 \rightarrow 0.592$; MemNet: $0.429 \rightarrow 0.477$). Memorability prediction itself transfers across age even without category-specific fine-tuning (ViT-B/16: $\rho = 0.456$ with *All*), with a small additional adolescent gain from *CatFT* (to $\rho = 0.471$); MemNet remains stable on adolescents ($\rho = 0.405$ with or without *CatFT*). Grad-CAM highlights semantically meaningful regions for highly memorable images and more diffuse patterns for low-memorability images. Overall, incorporating category structure improves adult accuracy, cross-age generalization of memorability prediction is robust, and among the tested backbones, ViT-B/16 performs best, with *CatFT* providing modest transfer gains.

Introduction

Image memorability, a phenomenon where certain images are consistently remembered or forgotten by different individuals, has intrigued the research community because it is robust yet difficult to interpret. In a foundational study by Isola et al. (2013), more than 2,000 scene images were presented to participants in a repeat-detection task that required them to identify recurring images. Memorability was defined as the likelihood of an image being remembered after a single exposure. This study revealed a large variation in image memorability, and a consistency analysis using Spearman's rank correlation confirmed that memorability rankings were stable across participants, suggesting that memorability relates more to the intrinsic properties of images than to the characteristics of observers. Despite progress in understanding attention and memory, the image features that drive memorability remain only partially understood.

Subsequent research extended these findings, showing stability across different memory tasks and retention intervals (Goetschalckx, Moors, and Wagemans 2018), across contextual settings (Bylinskii et al. 2015), under incidental and intentional memory instructions (Goetschalckx, Moors, and Wagemans 2019), and across age groups (Almog et al. 2023). While image memorability appears robust, identifying which visual attributes govern it is challenging. Evidence indicates that memorability reflects multiple factors that include intrinsic content properties and extrinsic context (Bylinskii et al. 2015; Rust and Mehrpour 2020; Bylinskii et al. 2022). Recent work links memorability to how images are represented in learned feature spaces, where single-exposure reconstruction error and latent distinctiveness carry signal about memorability (Bagheri and Mohsenzadeh 2024).

Building on this foundation, we ask whether incorporating semantic category structure during training improves alignment between model predictions and human memorability, and whether any gains persist across age groups. We study two representative backbones under a matched transfer path: MemNet as a convolutional neural network (CNN) and ViT-B/16 as a Vision Transformer (ViT). We contrast a single model trained on all categories with per-category fine-tuning, and we evaluate cross-age transfer by testing adult-trained models on adolescent labels without additional training. Finally, we use Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize which regions drive predictions for images with high and low memorability. Our aim is to clarify the role of semantic structure in memorability prediction and to assess whether the observed trends generalize across architectures and age groups.

Related Work

Memorability as an image-computable property. Isola et al. (2013) showed that image memorability rankings are highly consistent across observers after a single exposure using a repeat-detection task. Follow-up work reported robustness over longer delays (Goetschalckx, Moors, and Wagemans 2018), under contextual manipulations (Bylinskii et al.

2015), and across task instructions (Goetschalckx, Moors, and Wagemans 2019). Surveys synthesize evidence that memorability depends on image content, relates to attention and perception, and can be approached as an image-computable quantity (Rust and Mehrpour 2020; Bylinskii et al. 2022). Complementary representational accounts indicate that latent distinctiveness and reconstruction behavior in learned feature spaces relate to memorability (Bagheri and Mohsenzadeh 2024).

Datasets and predictive models. Large annotated datasets and regression models have driven progress on predicting image-level memorability scores. Khosla et al. (2015) introduced LaMem and trained MemNet, a CNN that achieves strong rank correlations with human annotations. With the advent of Vision Transformers (Dosovitskiy et al. 2020), transformer-based predictors have been explored for memorability, and ViT configurations have been reported as effective on multiple benchmarks (Hagen and Espeseth 2023). To avoid dataset leakage when analyzing MEMCAT category effects, we do not use memorability checkpoints trained on MemCat, such as ViTMem, and instead follow an ImageNet to LaMem to MEMCAT transfer path for both CNN and ViT backbones. In this work, these models serve as tools to study semantic training effects rather than as the focus of an architectural comparison.

Semantic structure and category-sensitive training. MEMCAT balances five broad categories and provides per-image memorability scores, enabling controlled tests of semantic effects (Goetschalckx and Wagemans 2019). Prior work shows that category membership explains some variance in memorability, but substantial variability remains *within* categories; in other words, category labels alone do not determine which images are remembered (Bylinskii et al. 2015, 2022). This supports the view that both high-level semantics and distinctive image content contribute to memorability. Most predictive models are trained on mixed-category data, leaving open whether per-category fine-tuning can better capture within-domain cues. We therefore compare a single mixed model (*All*) with per-category fine-tuning (*CatFT*) under matched protocols and assess whether any gains carry across age groups.

Cross-age memorability. Most datasets comprise adult participants. Almog et al. (2023) introduced adolescent memorability scores for a subset of MEMCAT, enabling cross-age evaluation on identical images and reporting high adult to adolescent agreement with content-dependent differences. This makes it possible to test whether adult-trained CNN and ViT predictors preserve rank correlation on adolescent scores without additional training, and whether benefits from category-sensitive training carry over.

Interpretability. Attribution methods, such as Grad-CAM, visualize regions of an image that influence model outputs (Selvaraju et al. 2017). In memorability, highly memorable images often align with semantically significant regions such as faces, animals, and central objects, whereas low memorability images emphasize diffuse backgrounds or less defined subject matter (Bylinskii et al. 2015, 2022). Beyond supervised saliency, recent work associates memorability with reconstruction difficulty and latent distinctiveness, using Integrated Gradients to characterize informative visual attributes on MEMCAT (Bagheri and Mohsenzadeh 2024). These analyses converge on the role of distinctive and semantically meaningful content in shaping memorability.

Materials and Methods

Datasets

MemCat MEMCAT (Goetschalckx and Wagemans 2019) contains 10,000 natural images evenly distributed across five semantic categories (animals, food, landscapes, sports, vehicles; 2,000 each). Memorability was measured with a continuous repeat-detection task in which participants viewed a stream of images and indicated repeats, following the general protocol introduced by Isola et al. (2013). For each image, the dataset provides an image-level memorability score derived from repeat-detection outcomes; where applicable, scores are corrected for false alarms in the dataset pipeline. These false-alarm–corrected scores provide a more accurate estimate of memorability by discounting noise from incorrect responses. The category structure allows either a single model trained on the full set or separate models per category.

Score distribution. Across MEMCAT (n=10,000), false-alarm–corrected memorability scores have mean 0.693 ± 0.137 (range 0.144–0.978); food shows the highest central tendency, landscapes the lowest, and animals, sports, and vehicles lie in between; the median number of respondents per image is 99 (interquartile range (IQR): 93–106), which is stable across categories.

LaMem We also use LaMem (Khosla et al. 2015), a large-scale image memorability dataset of over 60,000 diverse images, each annotated with a memorability score obtained through online repeat-detection experiments. LaMem provides broad coverage of visual content and serves as an intermediate stage for transfer learning: both convolutional neural network (CNN) and Vision Transformer (ViT) backbones are fine-tuned on LaMem before being further fine-tuned on MemCat.

Memoir MEMOIR (Almog et al. 2023) quantifies adolescent memorability (ages 10–18) using the same repeat-detection paradigm, enabling cross-age comparison with adult data. The published dataset contains 1,000 images sampled from the five MEMCAT categories and scored in an online memory game; scores are provided per image. We use MEMOIR to evaluate whether models trained on adult labels generalize to adolescent memorability without additional training. Both adult and adolescent evaluations use the exact 1,000 MEMCAT images that appear in MEMOIR (195 animals, 242 sports, 209 food, 162 vehicles, 192 landscapes): adult metrics use the MEMCAT scores for these images, and adolescent metrics use the MEMOIR scores.

Backbones and Training Regimes

Naming. We refer to training regimes as *All* (single model trained on the full MEMCAT training set) and *CatFT* (sep-

arate fine-tuning within each MEMCAT category). We combine these with the backbone name in text and tables, e.g., MemNet (All), MemNet (CatFT), ViT-B/16 (All), ViT-B/16 (CatFT).

Convolutional backbone (MemNet) MemNet (Khosla et al. 2015) is obtained by fine-tuning the Hybrid-CNN (AlexNet architecture) that was pretrained jointly on ILSVRC-2012 ImageNet and Places, replacing the task head with a single real-valued regression output and optimizing mean squared error (MSE) on LaMem. We initialize from the released MemNet weights (Hybrid-CNN pretrained on ImageNet+Places, then fine-tuned on LaMem) and further fine-tune on MEMCAT under the *All* and *CatFT* regimes.

Transformer backbone (ViT-B/16) For the transformer backbone, we adopt a Vision Transformer Base with 16×16 patches (ViT-B/16) (Dosovitskiy et al. 2020). We initialize from ImageNet-1k pretrained weights (Deng et al. 2009; Russakovsky et al. 2015) using the `timm` implementation (Wightman 2019), and replace the classification head with a single-unit regression output followed by a sigmoid (bounding outputs to $[0, 1]$), training with MSE. No memorability-specific pretrained checkpoints are loaded, avoiding leakage from MemCat-trained models such as ViTMem (Hagen and Espeseth 2023). To parallel MemNet's transfer path, we fine-tune this model first on LaMem and then on MEMCAT under the *All* and *CatFT* regimes.

Training Procedure

For both backbones (MemNet and ViT-B/16), we adopt a sequential transfer-learning path: ImageNet-1k pretraining $\rightarrow$ LaMem fine-tuning $\rightarrow$ MEMCAT fine-tuning. On MemCat, we train either a single model across all categories (All) or separate models per category (CatFT). Models are always trained to predict the dataset-provided memorability scores (false-alarm–corrected where available), using MSE as the loss.

Data splits. The test set is the 1,000-image overlap between MEMCAT and MEMOIR (195 animals, 242 sports, 209 food, 162 vehicles, 192 landscapes) and is used for both adult (MEMCAT scores) and adolescent (MEMOIR scores) evaluations to ensure per-image comparability. The remaining 9,000 MEMCAT images are split 80/20 into training and validation; for *CatFT*, this split is performed within each category. All checkpoints are selected by the lowest validation MSE, and the 1,000-image test set is never used for training or validation.

Optimization. We use Adam. Weight decay is applied for ViT but not for MemNet. Learning-rate scheduling uses `ReduceLROnPlateau` (factor 0.5) triggered after 3 epochs without validation improvement, with a minimum learning-rate floor. Early stopping with a patience of 30 is used to prevent overfitting, restoring the checkpoint with the lowest validation MSE.

Hyperparameter search. We conduct a grid search over batch sizes $\{16, 32, 64, 128\}$, learning rates in the range $[10^{-2}, 10^{-7}]$, and dropout in $[0, 0.5]$. For MemNet, dropout is applied at the fully connected layers (`fc6` and `fc7`); for ViT, dropout is applied via the implementation's `drop_rate` parameter. The final configuration for each model/backbone is selected by the lowest validation MSE.

Augmentation and normalization. For both backbones, we resize to 256×256, apply a single augmentation block wrapped in `RandomApply` with $p = 0.7$ (horizontal flip, blur, sharpness, color jitter, affine, and perspective), then center-crop to 224×224, convert to tensors, and normalize using the ImageNet mean and standard deviation. At evaluation time, no augmentations are used: we resize, center-crop, and apply ImageNet normalization only.

Evaluation Protocol

We report Spearman's rank correlation coefficient (ρ) between predicted and ground-truth memorability scores. We report overall and per semantic category performance to quantify gains from category-specific adaptation. We also monitor validation MSE during training for learning-rate scheduling and early stopping, and we evaluate using the checkpoint with the lowest validation MSE. To assess cross-age generalization, models are evaluated without additional training on MEMOIR (adolescents), and we report ρ on the full MEMOIR set as well as per category corresponding to the five MEMCAT categories.

Interpretability Procedure

To analyze which regions drive the predictions, we use Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al. 2017). We employ the `pytorch-grad-cam` implementation with a configuration matched to each backbone. For MemNet, the target layers are the convolutional blocks `conv1` to `conv5`. For ViT, the target layer is the final transformer block's pre-attention normalization (`blocks[11].norm1`), and we apply a reshape transform that discards the class token and maps patch tokens to a 14×14 spatial grid. We use augmentation smoothing and eigen smoothing to stabilize attribution. For each image, we compute Grad-CAM with respect to the memorability regression output, upsample the heatmap to input resolution, normalize it to $[0, 1]$, and overlay it on the corresponding RGB image after resizing and center cropping (i.e., on an unnormalized copy).

For qualitative analysis on MemCat, we display examples from the top 5% (high memorability) and bottom 5% (low memorability) of each category's score distribution, using the dataset-provided scores. Each panel shows the original high-memorability image and its Grad-CAM map, followed by the original low-memorability image and its Grad-CAM map. Visualizations are produced using the best-performing checkpoint.

Results

Performance on MEMCAT (Adults)

We first evaluate MemNet and ViT-B/16 on MEMCAT under a single mixed-category model (*All*) and separate category fine-tuning (*CatFT*), reporting Spearman's rank correlation ρ overall and per category.

As shown in Table 1 and Figure 1 (panel a), category-specific fine-tuning yields consistent gains relative to a single mixed model. The effect is largest for ViT-B/16: overall ρ rises from 0.548 (*All*) to 0.592 (*CatFT*; Δ=+0.044, about 8% relative). Per-category gains are strongest in Food (+0.085, 0.519→0.604) and present in Animals (+0.044), Vehicles (+0.044), Landscapes (+0.026), and Sports (+0.022). MemNet also improves under *CatFT* (overall 0.429→0.477, Δ=+0.048) with notable per-category boosts in Food (+0.098) and Landscapes (+0.076).

Two additional observations emerge. First, ViT-B/16 outperforms MemNet in nearly every category under both training regimes (Table 1), consistent with the view that transformers capture global structure helpful for memorability prediction. Second, the magnitude of achievable correlation varies by category and roughly tracks human-consistency ceilings: categories with higher human consistency (e.g., Landscapes) tend to yield stronger model correlations; Vehicles are also strong relative to lower-consistency categories such as Food and Sports.

Cross-Age Generalization

We next evaluate the adult-trained models directly on adolescent scores from MEMOIR, without any adolescent fine-tuning. Table 2 and Figure 1 (panel b) show that memorability signals transfer across age: both backbones retain moderate correlations on adolescents. ViT-B/16 again leads overall (0.456→0.471 with *CatFT*; Δ=+0.015), while MemNet shows little aggregate change (0.405→0.405). Category-wise patterns are heterogeneous: for ViT-B/16, Animals (+0.063) and Food (+0.028) gain under *CatFT*, while Sports dips slightly (−0.036). These shifts indicate that the relative weighting of semantic cues changes by age, but category-aware gains largely persist, especially for ViT.

Aggregate View and Practical Takeaways

Across datasets, the strongest configuration is ViT-B/16 with category-specific fine-tuning, reaching *MemCat* ρ=0.592 and *Memoir* ρ=0.471 (Figure 1). In practice, when category labels are available, category-aware training is a safe default because it delivers steady gains, especially in categories with distinctive within-class cues; transformers are preferable for this task, as they consistently outperform a strong CNN baseline; and cross-age robustness is the norm, since adult-trained models predict adolescent memorability well and the *CatFT* advantage is mostly preserved. Overall, *MemCat* yields ρ=0.429/0.477 for MemNet (*All*/*CatFT*) and 0.548/0.592 for ViT-B/16; *Memoir* yields ρ=0.405/0.405 for MemNet and 0.456/0.471 for ViT-B/16.

Interpretability Analysis

To characterize how models use image content, we applied Grad-CAM to the best-performing configuration (ViT-B/16, *CatFT*) and visualized high- and low-memorability exemplars per category (Figure 2). High-memorability images consistently elicit focused attributions on semantically meaningful regions (for example, faces, iconic silhouettes, central objects), whereas low-memorability images yield diffuse, background-oriented maps. Category-specific tendencies align with domain priors: Animals emphasize heads and contours; Food prioritizes central high-contrast items over clutter; Landscapes highlight distinctive near-field landmarks rather than distant horizons; Sports track human figures; Vehicles emphasize large recognizable shapes. Qualitatively, MemNet maps are more diffuse, mirroring their lower correlations.

Discussion

Our findings demonstrate that semantic categories play a crucial role in improving the performance of image memorability prediction models. Fine-tuning on specific categories in the MEMCAT dataset consistently increased correlations with human memorability scores compared to training on the dataset as a whole, for both convolutional and transformer-based models (Table 1, Figure 1a). This supports the view that semantic structure provides useful constraints for learning memorability-relevant features, and the magnitude of improvements varies across categories. Quantitatively, overall correlation increases from 0.548 to 0.592 for ViT-B/16 (Δ=+0.044, about 8% relative) and from 0.429 to 0.477 for MemNet (Δ=+0.048). The largest ViT category gain is Food (+0.085), with positive shifts also in Animals, Vehicles, Landscapes, and Sports.

Across both datasets, the Vision Transformer consistently outperformed the convolutional baseline. In MemCat, ViT-B/16 achieved higher correlations than MemNet in nearly every category, particularly in Food ($\rho = 0.60$ vs. $\rho = 0.37$) and Landscapes ($\rho = 0.62$ vs. $\rho = 0.52$) (Table 1). On Memoir, ViT also outperformed MemNet, with a notable advantage in Vehicles ($\rho = 0.60$ vs. $\rho = 0.49$) (Table 2, Figure 1b). This pattern aligns with evidence that transformer architectures capture long-range dependencies and holistic scene information more effectively than CNNs, and it is consistent with recent links between memorability and representational properties in learned feature spaces. To ensure a fair comparison and avoid dataset leakage, the ViT models were initialized from ImageNet, fine-tuned on LaMem, and then fine-tuned on MemCat, rather than loading memorability checkpoints already trained on MemCat.

Evaluation on the adolescent dataset (Memoir) revealed both consistency and divergence relative to adult memorability. Models trained on adult-based datasets predicted adolescent memorability with moderate correlations, supporting the robustness of memorability as an intrinsic image property across age groups. On adolescents, ViT-B/16 also benefits from category fine-tuning but more modestly (0.456 to 0.471, Δ=+0.015), while MemNet shows no net change (0.405 to 0.405). Cross-age transfer holds even without category fine-tuning: a single mixed-category ViT model already reaches ρ=0.456 on adolescents; *CatFT* adds a small, consistent boost. Category-level differences are more pronounced than global shifts, indicating that the relative weighting of semantic cues can change with development. For example, *CatFT* benefits are not uniform across adolescent categories: ViT Sports dips slightly (−0.036), while Animals improves (+0.063).

Method	All	Animals	Food	Landscapes	Sports	Vehicles
Human Consistency	0.781	0.671	0.591	0.771	0.600	0.641
MemNet (All)	0.429	0.448	0.268	0.448	0.470	0.509
MemNet (CatFT)	0.477	0.455	0.366	0.524	0.502	0.537
ViT-B/16 (All)	0.548	0.506	0.519	0.595	0.539	0.582
ViT-B/16 (CatFT)	0.592	0.550	0.604	0.621	0.561	0.626

Table 1: Category-wise Spearman correlations (ρ) on MEMCAT (adults). We compare MemNet and ViT-B/16 trained as *All* vs. *CatFT*. Human-consistency values are from Goetschalckx and Wagemans (2019).

Method	All	Animals	Food	Landscapes	Sports	Vehicles
Human Consistency	0.810	0.751	0.781	0.760	0.801	0.740
MemNet (All)	0.405	0.420	0.379	0.355	0.385	0.509
MemNet (CatFT)	0.405	0.314	0.398	0.455	0.387	0.489
ViT-B/16 (All)	0.456	0.348	0.448	0.511	0.426	0.578
ViT-B/16 (CatFT)	0.471	0.411	0.476	0.522	0.390	0.595

Table 2: Category-wise Spearman correlations (ρ) on MEMOIR (adolescents). We compare MemNet and ViT-B/16 trained as *All* vs. *CatFT*. Human-consistency values are from Almog et al. (2023).

Category-wise performance patterns also align with reported human-consistency ceilings in MemCat. Landscapes and Vehicles have relatively high human consistency, and the models achieve strong correlations in these categories, suggesting that shared salient structure, canonical silhouettes, and high signal-to-noise cues facilitate both human and model agreement. In Food, where human consistency is lower, ViT benefits more from category-specific fine-tuning than MemNet, which may reflect the importance of color and texture contrasts that are better captured by global attention. In Animals and Sports, performance differences narrow, consistent with the prominence of faces and bodies as strong, localized cues.

For model explanation, we used Grad-CAM on the best-performing models, specifically ViT-B/16 under category fine-tuning, and visualized adult MEMCAT exemplars from the top and bottom five percent within each category (Figure 2). High-memorability images elicit focused attribution on semantically meaningful regions such as faces, central objects, and iconic silhouettes, while low-memorability images produce diffuse or background-oriented maps. Category-specific tendencies are clear: for Animals, attention concentrates on heads and contours; for Food, it highlights central high-contrast items rather than surrounding clutter; for Landscapes, it spreads across distinctive near-field landmarks rather than distant horizons; for Sports, it tracks human figures; and for Vehicles, it emphasizes large recognizable shapes. Qualitatively, MemNet saliency maps are more diffuse and less localized than ViT, which aligns with MemNet's lower correlations.

Taken together, these results advance understanding of image memorability in several ways. They confirm that memorability is not uniform across semantic domains and that category-specific modeling yields substantial gains. They highlight transformers as well suited for capturing global scene-level context for memorability prediction. They show that while memorability generalizes across developmental stages, the balance of semantic cues can shift with age, which is relevant for applications that target diverse user groups. Finally, the interpretability results connect computational models with psychological theory, bridging machine and human perspectives on visual memory. Practically, if category labels are available, category fine-tuning is recommended; if not, a single ViT model trained on all categories is a strong default that generalizes across age.

Limitations and Future Directions

We did not report confidence intervals or significance tests; future work should include bootstrapped confidence intervals and paired tests across model regimes. Analyses are restricted to five semantic categories and rely on a single attribution method for visualization; expanding category coverage and complementing Grad-CAM with additional tools would further test generality. Exploring age-aware or domain-adaptation fine-tuning, calibration on small adolescent subsets, and additional interpretability tools could clarify how semantic cues shift with age and enhance cross-age robustness. Per-category fine-tuning increases the number of trained heads (i.e., effective capacity) relative to a single mixed model, so some gains may reflect capacity/partitioning effects; parameter-matched controls (e.g., multi-head or adapter baselines) are a priority for future work.

Conclusion

We studied how semantic structure influences image memorability prediction by fine-tuning two architectures, MemNet (AlexNet-based) and ViT-B/16 (Vision Transformer), on category-specific subsets of MemCat. Under category-specific fine-tuning, both backbones improve over training on the full dataset, with ViT-B/16 achieving the strongest

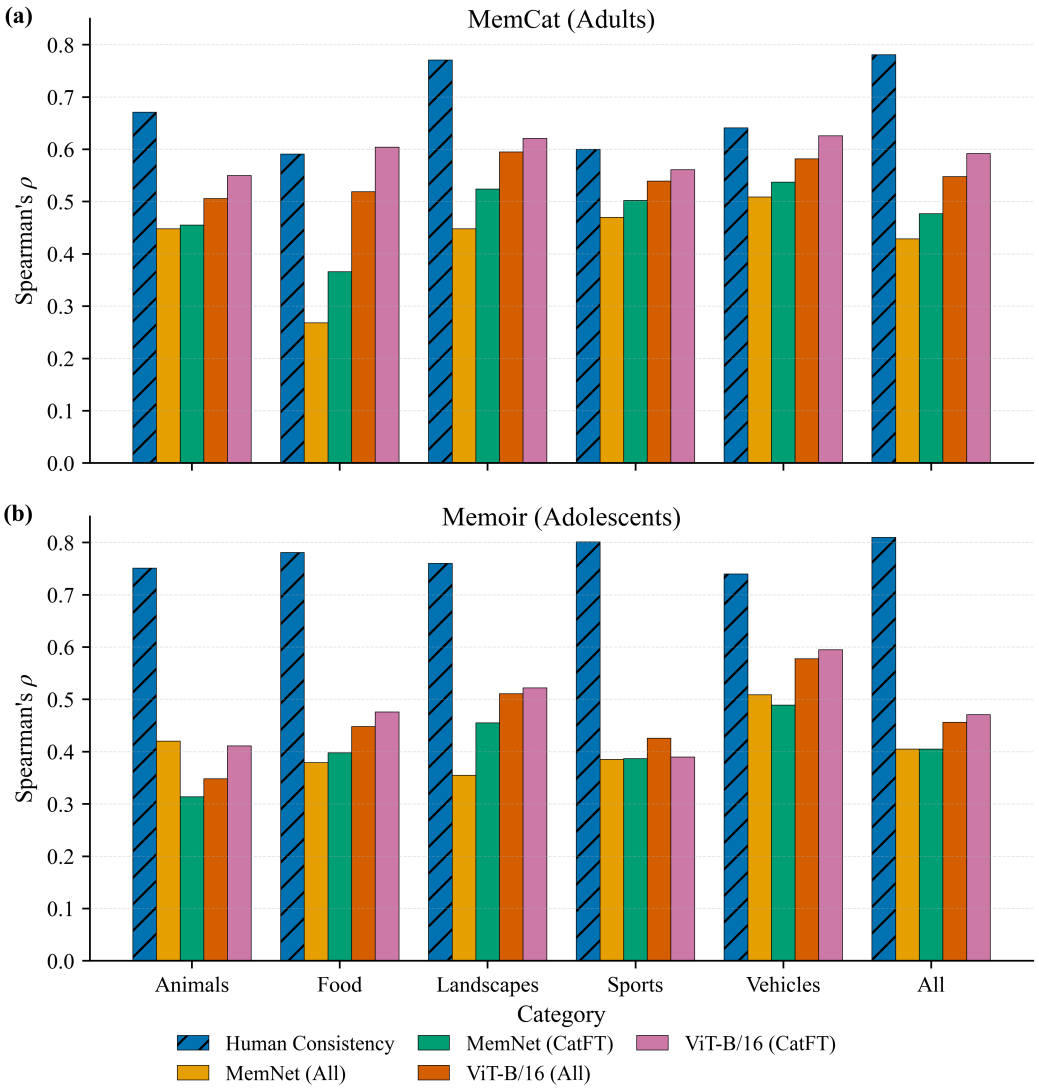

Figure 1: Spearman's ρ by category on MEMCAT (adults; panel a) and MEMOIR (adolescents; panel b). We compare MemNet and ViT-B/16 trained as *All* vs. *CatFT*; human consistency shown as reference. "All" denotes overall correlation; human "All" is the pooled split-half reliability.

overall performance across categories. On adults (MemCat) correlation rises from 0.548 to 0.592 for ViT-B/16 (Δ=+0.044, about 8% relative) and from 0.429 to 0.477 for MemNet (Δ=+0.048).

We assessed cross-age generalization by evaluating adult-trained models on adolescent labels from Memoir. Memorability signals transfer with moderate correlations, and the most salient differences appear at the category level, suggesting potential shifts in the relative weighting of semantic cues with development while preserving core regularities across age groups. Cross-age transfer holds even without category fine-tuning: a single mixed-category ViT model reaches ρ=0.456 on adolescents; *CatFT* adds a modest boost to 0.471 (+0.015), while MemNet shows no net change (0.405 to 0.405).

Interpretability analyses using Grad-CAM on the best-performing models (ViT-B/16 under CatFT) show that high-memorability images elicit focused attribution on semantically meaningful regions, whereas low-memorability images produce more diffuse or background-oriented maps. Qualitative inspection of MemNet attributions is more diffuse, aligning with its lower correlations.

Overall, the results establish that category-sensitive fine-tuning yields consistent gains on adults, cross-age generalization from adults to adolescents is stable with small additional gains from *CatFT* for ViT, transformer-based models outperform a CNN baseline, and attribution maps connect model predictions to psychologically meaningful cues. When category labels are available, *CatFT* is recommended; otherwise, a single ViT model trained on all categories is a strong default that generalizes across age.

Figure 2: Grad-CAM on MEMCAT using the best-performing model (ViT-B/16, *CatFT*). Rows are semantic categories (Animals, Food, Landscapes, Sports, Vehicles). Each pair shows the original image and its Grad-CAM map; memorability scores are overlaid on originals.

References

Almog, G.; Alavi Naeini, S.; Hu, Y.; Duerden, E. G.; and Mohsenzadeh, Y. 2023. Memoir Study: Investigating Image Memorability Across Developmental Stages. *PLOS ONE*, 18(12): e0295940.

Bagheri, E.; and Mohsenzadeh, Y. 2024. Modeling Visual Memorability Assessment with Autoencoders Reveals Characteristics of Memorable Images. *arXiv preprint arXiv:2410.15235*.

Bylinskii, Z.; Goetschalckx, L.; Newman, A.; and Oliva, A. 2022. Memorability: An Image-Computable Measure of Information Utility. In *Human Perception of Visual Information: Psychological and Computational Perspectives*, 207–239. Springer.

Bylinskii, Z.; Isola, P.; Bainbridge, C.; Torralba, A.; and Oliva, A. 2015. Intrinsic and Extrinsic Effects on Image Memorability. *Vision Research*, 116: 165–178.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Goetschalckx, L.; Moors, J.; and Wagemans, J. 2019. Incidental Image Memorability. *Memory*, 27(9): 1273–1282.

Goetschalckx, L.; Moors, P.; and Wagemans, J. 2018. Image Memorability Across Longer Time Intervals. *Memory*, 26(5): 581–588.

Goetschalckx, L.; and Wagemans, J. 2019. MemCat: A New Category-Based Image Set Quantified on Memorability. *PeerJ*, 7: e8169.

Hagen, T.; and Espeseth, T. 2023. Image memorability prediction with vision transformers. *arXiv preprint arXiv:2301.08647*.

Isola, P.; Xiao, J.; Parikh, D.; Torralba, A.; and Oliva, A. 2013. What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7): 1469–1482.

Khosla, A.; Raju, A. S.; Torralba, A.; and Oliva, A. 2015. Understanding and Predicting Image Memorability at a Large Scale. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2390–2398.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Rust, N. C.; and Mehrpour, V. 2020. Understanding Image Memorability. *Trends in Cognitive Sciences*, 24(7): 557–568.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Wightman, R. 2019. PyTorch Image Models (timm). https://github.com/rwightman/pytorch-image-models.