Using AI to Improve Real Photos Before They Are Taken

Originally published at https://www.unite.ai/using-ai-to-improve-real-photos-before-they-are-taken/

Instead of using GenAI to fix photos after *you shoot them, researchers have trained a system that tells you how to move, pose and frame the shot beforehand, using studied knowledge of what makes pictures memorable.*

Fixing photos after-the-fact has been getting easier for quite some time, as manufacturers and tech platforms increasingly offer in-camera editing that allows users to change images as soon as they have taken them. Popular systems of this kind include Google's conversational editing, and Samsung's generative edit, among others.

However, a nascent trend that favors 'authenticity' over AI-'improved' results could mean that many of the consumers such systems are aimed at begin to regard 'altered' photos as AI slop.

Perhaps this is what inspired Google to create an AI-trained 'camera coach' informed by Gemini, which is capable of giving direct instruction to improve a photo *during* the process of taking it:

Google's Camera Coach tells the user how to reframe a photo, among other basic pieces of advice. Source

Being a proprietary system, and with practically zero information available online in regard to it, Camera Coach appears to leverage Gemini to help users improve framing (see image above) or to make minor changes in stance (such as moving closer together, or looking directly at the camera).

So as far as anyone can tell, the product pushes the composition towards the median, presumably based on millions of uploaded content data-points likely to have contributed to Gemini's training data. In this sense, the uploading users have created the AI's calibration by rejecting unsatisfactory shots and uploading the ones they like – an effective (and free) form of dataset curation!

That said, photos which are *averaged-out* in terms of composition do not necessarily possess the same aesthetic values or viewer-impact as photos that are *memorable*.

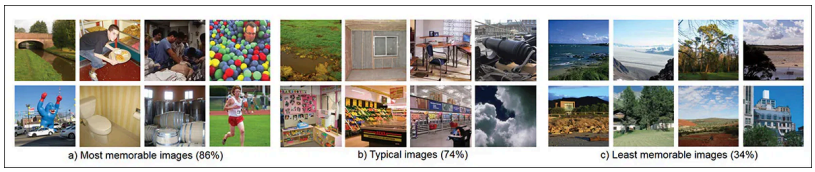
Beyond ‘Cheese!’ and the Rule of Thirds

To this end, and towards a system that is more accessible across platforms, new research from Italy offers a Coach-style system that’s based on prior knowledge of *what makes photos stick in the mind*:

Far-ranging examples of advice from the authors’ new system. Source

In the examples above, we see advice given by the authors’ new system – dubbed *MemCoach* – which it is hard to imagine a composition-centric AI such as Camera Coach providing. In the first (leftmost) instance, the advice to remove the headdress is particularly specious; in the second picture, it’s hard to imagine what conventional context a composition-based AI could draw from the general scenario (i.e., an ‘artistic’ picture of a young woman lying on the floor with her eyes closed).

The core understanding about memorability in photography, used to develop the three-part Italian system, is drawn from various prior works, including the 2015 outing *What makes an object memorable?*, and the 2013 paper *What makes a photograph memorable?*.



From the 2013 paper What makes a photograph memorable?, *representative examples of good, medium and bad photos, in terms of memorability.* Source

Anyone, like me, with a negative Unix birth-date, will probably recognize the template for ‘least memorable images’ (top right in image above), from the endless slide nights that cursed our childhoods. As the authors state*:

> *‘These works identified key intrinsic factors such as the presence of people, indoor scenes, or emotional expressions, rather than objects and panoramic views, as well as extrinsic ones, including context and the observer. ‘*

The project centers on ‘memorability feedback’ (*MemFeed*), which is expressed in the MemCoach tutor application, and a benchmark (titled *MemBench*) based on the PPR10K dataset.

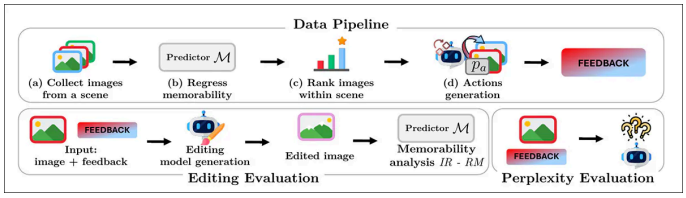
From the paper PPR10K: A Large-Scale Portrait Photo Retouching Dataset with Human-Region Mask and Group-Level Consistency*, diverse samples from the da images, bottom row shows expert-retouched versions along with corresponding human-region masks. The original photos vary widely in viewpoint, background, lig whereas the retouched results display improved visual quality and stronger consistency within each group.* Source

The paper observes that memorability is quantifiable in photos, rather than a register of subjective judgments, and the authors further note that the property has been identified both for photos (in various works) and videos (in various others).

The new paper is titled *How to Take a Memorable Picture? Empowering Users with Actionable Feedback*, and comes from four researchers across the University of Trento, the University of Pisa, and Fondazione Bruno Kessler. The accompanying project page suggests that GitHub code and Hugging Face-hosted data will be available next month (March 2026).

Method

To curate the MemBench dataset from the source PPR10K portrait dataset, the researchers grouped photos from the same scene and scored each image for memorability using a trained predictor based on CLIP features. They then ranked the photos within each scene from less to more memorable and paired them accordingly:



Overview of MemBench construction and evaluation. Top row depicts the data pipeline, from grouping images by scene and predicting memorability, to ranking photos and generating memorability-aware action feedback. Bottom row illustrates evaluation, measuring feedback quality through editing-based memorability gai and perplexity scoring.

For each pair, natural language descriptions were generated with the InternVL3.5 model to explain the visible differences between the less memorable version and the more memorable version; and these descriptions would constitute the training signal for the memorability feedback system.

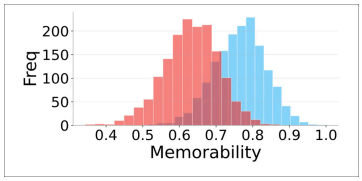
In contrast to the kind of logic that underpins Google's Camera Coach, the researchers sought a more subtle set of interpretations[†]:

> *'Contrary to computational photography adjustments focusing on post-hoc corrections (e.g., "make the image brighter"), we focus on semantic actions that a user can take on-the-fly for a better shot, e.g., "Face each other".'*

The final MemBench collection comprises around 10,000 images grouped into 1,570 scenes, averaging 6.5 images for each scene. The word-cloud the authors generated (see image below), suggests a wide range of semantic categories in the dataset:

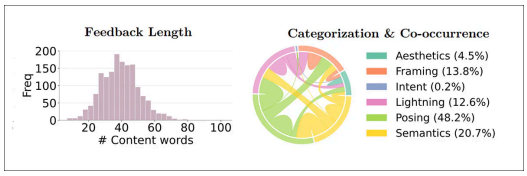
A word cloud of the most frequent terms in MemBench.

Source photos averaged a memorability score of 0.63, while the most memorable shots from the same scene stretched from 0.51 up to 1.0, with noticeable overlap between the two groups:



Memorability score distributions comparing the least and most memorable images within each scene.

The feedback itself ranged from short seven-word notes, to notably longer instructions (left, in image below). Each piece of advice was then broken down into small action types using GPT-5 Mini (right, in image below):



Feedback length distribution measured in content words, and categorization of atomic sub-actions with chord widths indicating co-occurrence frequency across categories.

The authors note that most suggestions focused on how the subject was posed, followed by changes to meaning or scene content, with framing often linked to posing, and lighting adjustments frequently tied to semantic changes.

Flux Capacitor

To evaluate whether memorability was increased by the feedback, user compliance was simulated through the use of the FLUX.1 Kontext generative model, as a proxy for the photographer. Given a source image and a piece of textual feedback, an edited version was generated by Flux that simulated the suggested changes:

The images on the left are real, from the dataset, and the images on the right (in each case) are created by Flux, based on the prompt (in yellow, below). In this way, the effectiveness of prompts could be evaluated without extensive human involvement. This knowledge would feed back finally into the MemCoach framework, and in fact represents a workflow that could iteratively improve a system of this kind (i.e., ultimately with real-world rather than Flux examples).

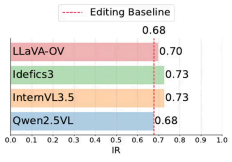
Both the original and edited images were then passed through a memorability predictor, enabling measurement of how often the edited version achieved a higher score – termed the *Improvement Ratio* – and how large the gain was relative to the starting image, termed *Relative Memorability*.

Similarity to memorability-focused reference advice was also measured by calculating perplexity against the ground-truth descriptions, and an 80–20 split was applied at the scene level so that testing was carried out only on scenes that had not been used during training.

State of the Art

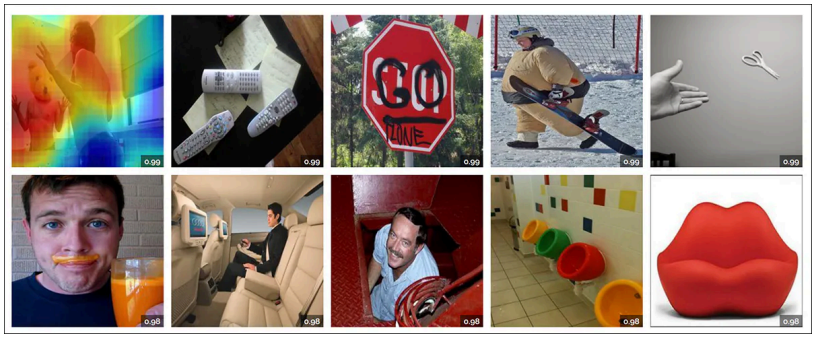
The memorability awareness of current multimodal large language models was tested. Images from the LaMem dataset were shown to several leading models, which were asked if the image was memorable. The model's confidence estimate was then compared with the scores assigned by human viewers in the original study:

Model	Spearman Rank (↑)
Inter-annotator*	0.68
QWEN2.5VL [4]	-0.06
INTERNVL3.5 [58]	-0.01
IDEFICS3 [33]	-0.07
LLAVA-OV [2]	0.08



Tests indicating that baseline multimodal models do not capture memorability. Left, Spearman rank correlation between model predictions and LaMem ground-truth scores, with inter-annotator agreement from LaMem shown for reference. Right, improvement ratio achieved by zero-shot feedback relative to the editing baseline, showing only marginal gains.

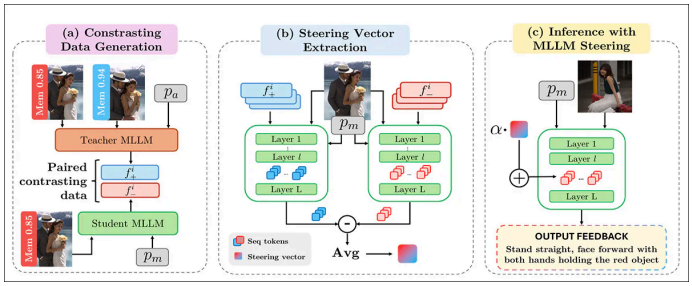
Almost no meaningful correlation with human judgments was found, and, despite large-scale pretraining, the authors assert that the models did not track what people consistently remember.



Examples from the LaMem dataset. Upper-left, we also see a heat map depicted for that image. Source

MemCoach

MemCoach focuses on semantic, on-the-fly instructions that can be carried out before the shutter is pressed – for example, adjusting pose, altering interactions between subjects, or modifying scene elements. The feedback provided by MemCoach varies from 7-102 content words. Memorability, the paper posits, appears to be driven more by subject configuration and narrative cues than by simple compositional tweaks:



Overview of the MemCoach pipeline, in which memorability-aware guidance from a teacher MLLM is paired with neutral student responses to form contrasting dat activation differences across layers are averaged to derive a memorability steering vector; and that vector is injected at inference to shift student activations towar producing improved, memorability-oriented feedback, without additional training.

Tests

Seven Multimodal Large Language Models (MLLMs) were used in the testing phase for the new system: Qwen2.5V.L; InternVL3_5-8B; Idefics3-8B; and LLaVA-OneVision-1.5. Additionally GPT-5 Mini was included as a representative of proprietary, closed-source models, along with the aesthetics-specialized Q-Instruct and AesExpert models. The MLLMs operated variously as zero-shot and teacher oracles.

InternVL3.5 was used for both the teacher and student models, with the MemBench training split used to create contrasting examples:

Model	Editing		Perplexity (↓)
	IR (↑)	RM% (↑)	
Edit model	0.68	3.72	*n.d.*
Teacher oracles			
LLaVA-OV [2]	0.74	5.93	5.73
Idefics3 [33]	0.80	9.84	29.21
Qwen2.5VL [4]	0.83	10.16	2.34
InternVL3.5 [58]	0.85	11.92	2.40
Aesthetics specialized			
AesExpert [20]	0.73	6.67	5.97
Q-Instruct [60]	0.73	5.31	5.36
Zero-shot baselines			
GPT-5 Mini [43]	0.75	7.03	*n.d.*
LLaVA-OV [2]	0.70	5.87	7.58
Idefics3 [33]	0.73	6.64	20.19
Qwen2.5VL [4]	0.68	4.26	10.23
InternVL3.5 [58]	0.73	5.47	5.49
MemCoach (Ours)	**0.80**	**7.21**	**4.99**

MemCoach performance compared with state-of-the-art MLLMs across teacher oracles, aesthetics-specialized models, and zero-shot baselines, showing higher Improvement Ratio and competitive Relative Memorability together with the lowest perplexity, indicating more consistent and memorability-oriented feedback.

In the table for the first test (shown above), we see that MemCoach appears to delivers more effective memorability advice than any of the comparison models – and the steered InternVL3.5 model raises memorability more often, and by a larger amount, with a 5% Improvement Ratio gain over GPT-5 Mini, and a 31.81% jump in Relative Memorability over its unsteered version.

It also outperforms aesthetics-focused systems, despite requiring no extra training. Lower perplexity, the paper contends, further suggests that its feedback follows the same linguistic patterns that human memorability judgments tend to reward:

Model	Editing		Perplexity (↓)
	IR (↑)	RM% (↑)	
LLaVA-OV [2]	0.70	**5.87**	**7.58**
***MemCoach*-LLaVA**	**0.73** (+4.29%)	5.04 (-14.14%)	14.05 (+85.36%)
Idefics3 [33]	0.73	6.64	20.19
***MemCoach*-Idefics**	**0.75** (+2.74%)	**6.69** (+0.75%)	**19.81** (-1.88%)
Qwen2.5VL [4]	0.68	4.26	**10.23**
***MemCoach*-Qwen**	**0.74** (+8.82%)	**5.49** (+28.87%)	13.90 (+35.87%)
InternVL3.5 [58]	0.73	5.47	5.49
***MemCoach*-InternVL**	**0.80** (+9.59%)	**7.21** (+31.81%)	**4.99** (-9.11%)

Generalization results showing that MemCoach improves memorability-oriented feedback across multiple multimodal backbones, consistently raising Improvement Ratio and Relative Memorability while also reducing perplexity for most models.

A further test (see table above) indicates that adding MemCoach boosted memorability-aware feedback across every tested multimodal backbone, with consistent gains in Improvement Ratio and the largest jumps appearing for Qwen2.5VL and LLaVA-OV.

A qualitative evaluation was then conducted, analyzing examples of MemCoach feedback in which the source image, the natural-language suggestion, and the imagined improved result were examined side-by-side:

Qualitative examples of memorability-oriented feedback generated by MemCoach. Each triplet shows the source image, the natural-language instruction, and the Relative Memorability (RM) indicating the measured change. The guidance ranges from pose and gaze adjustments to semantic interventions such as object remo gains and cases where removing unusual elements reduces memorability.

Of these results, the authors state:

> *'The examples highlight the variety of suggestions the model proposes, ranging from fine-grained compositional adjustments, such as altering gaze direction, pose, or hand position, to semantic interventions involving object removal or face expression change.*
>
> *'Feedback is naturally interpretable and actionable, expressed in concise textual instructions (mostly involving verbs "Bring", "Stand", "Remove") that can be directly implemented, effectively verbalizing how to take a memorable picture.'*

Conclusion

It would be most interesting to compare the methodology of Google's closed-box approach to the MemBench project – not least to know what central standards, references and databases Google used to define the system's aesthetic standards.

The negative aspect of systems of this kind, open or closed source, is that at scale they risk to enforce uniform standards that are destined to end as memes and clichés – a kind of visual equivalent of the AI em-dash debates, wherein the 'correct' procedure has become somewhat cursed in casual usage.

* *My conversion of the authors' inline citations to hyperlinks, if the link is not presented elsewhere in the article.*

† *The paper refers here, as in several other places, to 'supplementary material' that I am unable to locate, either from the paper, the core Arxiv listing, or the project site.*

First published Thursday, February 26, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More