# DOES MEMORABILITY GUIDANCE TRANSFER BEYOND THE GUIDING CRITIC? EVIDENCE FROM LATENT DIFFUSION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Image memorability is the likelihood that an image will be remembered after viewing. Neural memorability predictors estimate this likelihood and can guide diffusion models toward images with higher predicted memorability. We call the predictor that provides this guidance the guiding critic. Prior work shows that such guidance can also raise average scores from predictors not used for guidance. What remains unclear is (i) whether improvements in predicted memorability transfer consistently beyond the guiding critic at the individual-image level, (ii) whether these improvements arise from the intended guidance direction rather than generic changes introduced by generation or reconstruction, and (iii) whether gradient-based guidance justifies its added computation compared with simple sampling alternatives. We investigate these questions using training-free universal guidance in latent diffusion for both text-to-image generation and real-image editing. The critic supplies gradients during sampling, while the other predictors serve as held-out evaluators that only score the final images. With AMNet as the critic, guidance raises the mean ResMem, ViTMem, and PerceptCLIP scores by 2.87%, 1.69%, and 1.67% on 200 COCO prompts fixed before generation, while reducing CLIP prompt similarity by 1.88%. These gains hold across four seeds on DrawBench and COCO, and on DrawBench they hold whichever of AMNet, ResMem, or ViTMem acts as the critic. At the image level, however, the held-out predictors often disagree, with all three scores rising together on only 69 of 200 COCO image pairs. To separate the effect of guidance from the changes that noising and denoising make on their own, we edit real SUN photographs and compare each edit with a matched unguided reconstruction. Forward guidance raises and reverse guidance lowers the held-out scores, while a random push of the same size has near-zero effect, so the change follows the direction of the critic's gradient. Generating four unguided images and keeping the one the critic scores highest is competitive with guidance on the held-out scores, keeps prompt similarity higher, and runs faster in our benchmark. A five-prompt diagnostic on Stable Diffusion XL shows that predictor scores can rise even when images are visibly malformed. Together, these results show that memorability guidance produces real but modest gains beyond its critic, which hold on average but not reliably per image, cost prompt fidelity, and offer little over selecting the best of a few samples.

## 1 INTRODUCTION

Some images are more memorable than other images. Image memorability is the intrinsic property of an image to be remembered or forgotten by observers. Memorability can be quantified as a score from 0 to 1 with high values indicate that the image will be remembered more often (Isola et al., 2011b;a; Khosla et al., 2015). This property has been shown to be consistently across different individuals (Bylinskii et al., 2015; Isola et al., 2013). Understanding and modeling this cognitive property holds critical importance across various fields, including education, advertising, data visualization and even human-computer interaction, as it enables the design of visuals that can efficiently capture and retain human attention.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Building upon the foundational understanding of visual persistence, recent advancements in computational vision have shifted focus from merely predicting to actively improving the memorability of images. While Generative Adversarial Networks (GANs) have been used for increasing memorability, they were restricted to narrow domains due to inherent challenges in cross-category generalization e.g. human faces (Sidorov, 2019), style transfer (Siarohin et al., 2017).

Diffusion and flow-matching generators are not tied to one domain and have been used to manipulate memorability (Huang & Simo-Serra, 2024; Dang et al., 2026). Several neural networks trained on human memorability scores predict the score of a new image (AMNet, ResMem, ViTMem, PerceptCLIP; Fajtl et al., 2018; Needell & Bainbridge, 2022; Hagen & Espeseth, 2023; Zalcher et al., 2025). They are differentiable, so universal guidance (Bansal et al., 2024) can use one of them as a frozen critic that steers a frozen latent diffusion model (Rombach et al., 2022). We call the predictor that supplies this signal the *guiding critic*: at each sampling step, the gradient of its score on the predicted clean image shifts the noise estimate.

Optimizing a learned score creates a measurement problem: a gain on the guiding critic shows only that the optimization worked. We therefore audit guidance with *held-out* predictors: in each run, the predictors other than the critic supply no gradients and score only final images. AMNet is the critic in all but one experiment; in that one we rotate the role among AMNet, ResMem, and ViTMem to check that the transfer does not depend on that choice. We ask whether held-out scores rise, and whether the rise follows the critic's direction; whether the held-out predictors agree on which images improve; what guidance costs in prompt fidelity, source preservation, and compute; how it compares with the best of $K$ unguided samples; and where it fails. The four predictors share their training data, so held out does not mean independent; human memory is not measured.

Our contributions are:

1. We demonstrate that training-free memorability guidance yields significant average score improvements on held-out predictors across multiple datasets and interchangeable frozen critics. However, paired analyses reveal that these population-level gains mask substantial instance-level disagreement among predictors.
2. Using matched reconstruction, reverse guidance, and norm-matched random controls in real-image editing, we test whether score changes follow the critic's gradient rather than generic perturbation. Forward and reverse guidance move the scores in opposite directions, while matched random perturbations have near-zero effect.
3. We measure text-image alignment, source preservation, and latency, and compare gradient guidance with best-of-four unguided selection. Guidance reduces text–image alignment and, in editing, costs a little source preservation, while best-of-four is competitive on held-out memorability scores and faster in our benchmark.

## 2 Related work

The study of image memorability began with Isola et al. (2011a;b), who introduced the Visual Memorability Game and the SUN dataset to quantify how consistently images are remembered across observers. Early computational approaches relied on handcrafted visual descriptors, such as GIST (Oliva & Torralba, 2004), SIFT and HOG to capture these cognitive properties. The paradigm shifted significantly with the advent of deep learning, marked by Khosla et al.'s introduction of MemNet (Khosla et al., 2015) and the large-scale LaMem dataset, which demonstrated that hierarchical features learned via convolutional networks vastly outperform handcrafted alternatives. Subsequent architectures refined this trajectory through transfer learning (e.g., AMNet (Fajtl et al., 2018)), multi-modal captioning integration (Squalli-Houssaini et al., 2018) and graph learning (Banna et al., 2023). While these works established robust predictive frameworks, research has predominantly focused on measurement and scoring, leaving the automated generation and manipulation of memorable images relatively underexplored.

Our work relies on four LaMem-trained predictors: AMNet (Fajtl et al., 2018) leveraging recurrent attention over convolutional representations, ResMem (Needell & Bainbridge, 2022) utilizing a residual architecture, ViTMem (Hagen & Espeseth, 2023) employing a vision transformer, and PerceptCLIP (Zalcher et al., 2025) operating via a LoRA-adapted CLIP encoder. Prior generative manipulation frameworks have explored style transfer (Siarohin et al., 2017), attribute-specific GAN

modifications (Sidorov, 2019), latent space traversal via GANalyze (Goetschalckx et al., 2019), and targeted scene-generation GANs (Kyle-Davidson et al., 2022), with the latter two validated through human recognition experiments. However, none of these past approaches utilize modern text-to-image diffusion models.

Expanding beyond traditional classifier and classifier-free formulations (Dhariwal & Nichol, 2021; Ho & Salimans, 2021), universal guidance (Bansal et al., 2024) derives its steering signal from arbitrary differentiable functions of the estimated clean image, enabling off-the-shelf networks to control frozen generative models. (Huang & Simo-Serra, 2024)proposed a mehtod to perturb input images to intermediate diffusion states via SDEdit (Meng et al., 2022) before steering the reverse trajectory with a memorability objective. Similarly, CogBlender (Dang et al., 2026) regulates cognitive properties like memorability and emotion in flow-matching models by interpolating velocity fields across prompt variations. While distinct in their technical execution, our study primarily centers on a rigorous audit of these guidance mechanisms.

# 3 METHODOLOGY

## 3.1 PROBLEM DEFINITION

Let a pre-trained latent diffusion model be parameterized by a noise predictor $\epsilon_\theta$, a decoder $D$, and a cumulative noise schedule $\bar{\alpha}_t$, conditioned on a text prompt embedding $c$. Given $c$, standard sampling generates a final image $x = D(z_0)$ that aligns with the textual description and maintains perceptual realism.

Beyond generating images from pure text descriptions, we also investigate real image-to-image editing, where the generation trajectory is initialized from a partially noised encoding of an existing source image $x_{\text{src}}$ rather than an empty or text-conditioned latent.

In both settings, our objective is to steer the generation trajectory toward a higher memorability target. Let $M$ denote a pre-trained memorability critic (e.g., AMNet) that maps an image to a scalar score. Given a prompt $c$ and a fixed diffusion schedule, we aim to modify the intermediate latents $\{z_t\}_{t=T}^{0}$ such that the final decoded output maximizes the predicted memorability score $M(D(\widehat{z}_0))$, while simultaneously preserving semantic alignment with the prompt $c$ and visual fidelity, all under a strictly training-free framework where both the diffusion model and the critic remain frozen.

## 3.2 MEMORABILITY GUIDANCE AND EVALUATION

At each denoising step, $\epsilon_{\text{cfg}}$ is the classifier-free-guided noise estimate and $\widehat{z}_0(z_t)$ is the clean latent it implies. The critic $M$ scores the decoded clean estimate after its own differentiable preprocessing $P$; we call this score $m_t$. Its gradient with respect to the noisy latent $z_t$ shifts the noise estimate with guidance weight $w$:

$$g_t = \nabla_{z_t} M\big(P(D(\widehat{z}_0(z_t)))\big), \tag{1}$$

$$\widetilde{\epsilon}_t = \epsilon_{\text{cfg}} - w\sqrt{1-\bar{\alpha}_t}\, g_t. \tag{2}$$

The DDIM step (Song et al., 2021) then uses $\widetilde{\epsilon}_t$ in place of $\epsilon_{\text{cfg}}$. We take $g_t$ end-to-end through the critic, the decoder, and the U-Net; no weights are updated. With $w = 0$ the sampler is exactly the unguided one, and $w < 0$ reverses guidance. We use only the forward term of universal guidance (Bansal et al., 2024), once per step. Appendix A.1 gives the standard equations.

Our evaluation framework utilizes four established memorability predictors: AMNet, ResMem, ViTMem, and PerceptCLIP. AMNet, ResMem, and ViTMem alternately serve as the guiding critic $M$ during generation, whereas PerceptCLIP is reserved strictly for evaluation. During any given sampling run, only the active critic provides the gradient signal for Eq. 1. The remaining models act as held-out predictors: they neither influence the generation trajectory nor participate in image selection, evaluating only the final saved output (Fig. 1). Unless specified otherwise, AMNet functions as the default critic, with rotating roles explored in Section 5.1.

Because optimizing the active critic merely confirms that the gradient descent successfully altered the score toward the critic's objective, we treat the critic's own score increase purely as a manipulation check. True evidence of generalizable improvement is instead measured through the responses of the

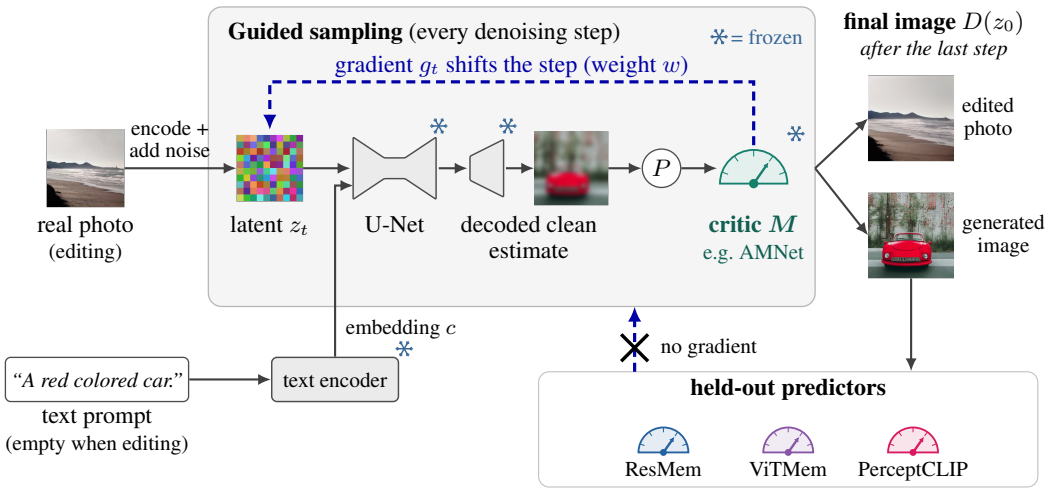


Figure 1: **One frozen critic steers every denoising step through its gradient; the other predictors never steer and only judge the final image.** Text-to-image sampling starts from noise and a prompt; editing starts from a real photo, encoded and partly noised, with an empty prompt. The blue dashed path is the gradient of the critic's score, taken back through $P$, the decoder, and the U-Net to the latent.

held-out predictors. However, because all four models are trained on the same underlying human dataset, they share common inductive biases and error profiles that optimization can inadvertently exploit. Finally, standard metrics such as CLIP similarity, SSIM, and LPIPS are tracked exclusively as post-hoc evaluations to quantify text alignment and image fidelity, as they receive no gradient feedback during generation.

## 4 Experimental Details

### 4.1 Datasets

Our text-to-image experiments utilize all 200 prompts from DrawBench (Saharia et al., 2022) and a fixed sample of 200 COCO val2017 captions (Lin et al., 2014). To mitigate hyperparameter tuning bias, the guidance weight $w = 100$ was determined via a grid search on an initial subset of 20 DrawBench development prompts, with a secondary confirmation set of 200 locally fixed COCO captions verifying generalizability. For real-image editing, we use the SUN Memorability dataset (Isola et al., 2011b), containing 2,222 images.

### 4.2 Implementation Details

The core generative model across our primary experiments is Stable Diffusion v1.4 (SD1.4) operating at a native $512^2$ resolution, utilizing 20 DDIM steps ($\eta = 0$) and a classifier-free guidance (CFG) scale of $s = 7.5$. All memorability predictors utilize their official pre-trained checkpoints. Notably, the AMNet preprocessing function resizes images without antialiasing during active gradient optimization, whereas offline final scoring applies antialiased resizing. Consequently, the optimization and evaluation phases share identical network weights but differ slightly in their spatial resampling operations.

For image-to-image editing, we adopt an SDEdit-style (Meng et al., 2022) approach wherein the source photograph is encoded into the latent space, corrupted with noise, and iteratively denoised. The editing strength dictates the noise injection level: at the primary balanced setting (strength 0.35), the sampler executes the final 7 of 20 diffusion steps, while the secondary aggressive setting (strength 0.50) runs the final 10 steps. Because these edits operate without a text prompt—relying entirely on the critic's gradient ($w = 200$), we isolate the pure effect of the guidance using a matched reconstruction baseline. This baseline uses the exact same encoded latent, noise vector, and diffusion

Table 1: **Held-out predictors rise and CLIP falls with AMNet as the critic.** Relative change (%) of the guided mean ($w = 100$) over the unguided mean, 200 prompts per set. COCO confirmation: captions fixed before generation, one seed; DrawBench and COCO: mean ± SD over 4 seeds.

| Score | COCO confirmation | DrawBench | COCO |
|---|---|---|---|
| AMNet (critic) | $+5.46$ | $+5.32 \pm 0.31$ | $+6.24 \pm 0.33$ |
| ResMem | $+2.87$ | $+3.63 \pm 0.19$ | $+3.46 \pm 0.73$ |
| ViTMem | $+1.69$ | $+2.33 \pm 0.27$ | $+2.37 \pm 0.46$ |
| PerceptCLIP$^{\dagger}$ | $+1.67$ | $+2.34 \pm 0.21$ | $+2.16 \pm 0.39$ |
| CLIP | $-1.88$ | $-1.62 \pm 0.95$ | $-1.48 \pm 0.25$ |

trajectory, but disables guidance ($w = 0$). By measuring source preservation against the original photograph for both the guided edit and its matched reconstruction, we strictly isolate the degradation caused by the guidance from the inherent baseline reconstruction loss.

Building on the established experimental design, the control arms are implemented mechanically as follows: reverse guidance inverts the objective weight ($w = -200$), while the norm-matched random control forces a perturbation of identical magnitude to the active guidance at every step, but projected into a random orthogonal direction. The untrained critic control utilizes an uninitialized AMNet architecture, leaving its gradient norms unmatched. The best-of-four baseline passively evaluates four unguided generation trajectories, selecting the output with the highest AMNet score.

To strictly isolate guidance effects, all comparative evaluations utilize paired arms initialized from identical latents $z_T$ under a shared deterministic DDIM schedule. This ensures variations in prompt and initialization cancel out in paired differences, though standard GPU non-determinism during guided backward passes inherently introduces minor variance not present in unguided runs. Methodologically, editing hyperparameters (strength and $w$) were finalized on 20 held-out development photos prior to any test-set generation. Conversely, the rotating-critic experiment was conducted retrospectively on initial development prompts, sequentially applying pre-fixed weights across AMNet, ResMem, and ViTMem.

### 4.3 Evaluation Metrics

Post-hoc validation utilizes CLIP ViT-B/32 cosine similarity to quantify text-image alignment. Source preservation during real-image editing is strictly quantified using structural similarity (SSIM), learned perceptual image patch similarity (LPIPS), and image-to-image CLIP cosine distance (CLIP-I).

## 5 Results Analysis

### 5.1 Transfer to held-out predictors

Our primary text-to-image experiments demonstrate that memorability guidance successfully transfers to predictors excluded from the gradient computation. On the COCO confirmation set using AMNet as the guiding critic ($w = 100$), all three held-out predictors exhibit significant mean score increases relative to the unguided baseline: ResMem improves by 2.87%, ViTMem by 1.69%, and PerceptCLIP by 1.67% (all $p < 1.4 \times 10^{-4}$, Holm-corrected). As expected, the guiding critic, AMNet, achieves the largest gain (+5.46%). Notably, these memorability improvements incur a minor degradation in prompt fidelity, with CLIP text-image similarity falling by 1.88%, though this remains within a 5% non-inferiority margin ($p_{\mathrm{NI}} = 1.82 \times 10^{-7}$).

PerceptCLIP's $p$ is unadjusted ($6.28 \times 10^{-5}$). Over four seeds (Table 1), all three held-out predictors rise on DrawBench and COCO, their mean change is positive at every seed, and CLIP again falls (every four-seed change has $p < 3.4 \times 10^{-5}$). Without the 20 development prompts, the four-seed DrawBench gains are $3.49 \pm 0.35\%$ (ResMem) and $2.29 \pm 0.28\%$ (ViTMem).

To verify that these gains are not an artifact of AMNet's specific architecture, a rotating-critic experiment on DrawBench alternated AMNet, ResMem, and ViTMem as the active guiding critic. Across all 18 critic-to-predictor permutations ($w \in \{10, 50, 100\}$), the held-out scores consistently increased with statistical significance (largest $p_{\mathrm{Holm}} = 0.0020$).

Table 2: **Rotating the critic on DrawBench: the held-out scores rise for every critic.** Each row is one guided run, defined by its critic and guidance weight $w$. The *update ratio* shows how hard the critic pushes: at each denoising step, the size of the guidance shift (Eq. 2) divided by the size of the model's own noise estimate $\epsilon_{\text{cfg}}$, averaged over steps and images; $0.041$ means the critic changes each step by about 4%. We set $w$, not this ratio, so the same $w$ gives ResMem about twice the push of AMNet or ViTMem, and the rows do not rank the critics. Entries are relative changes (%) from the unguided image, mean $\pm$ SD over 4 seeds (200 prompts); the critic's own column is a manipulation check. The AMNet $w = 100$ run is a new, deterministic run, so it differs slightly from Table 1.

| Critic | $w$ | Update ratio | AMNet | ResMem | ViTMem | CLIP |
|---|---|---|---|---|---|---|
| AMNet | 10 | 0.005 | $+0.69 \pm 0.11$ | $+0.91 \pm 0.54$ | $+0.37 \pm 0.23$ | $-0.14 \pm 0.50$ |
| AMNet | 50 | 0.023 | $+3.04 \pm 0.41$ | $+2.16 \pm 0.41$ | $+1.21 \pm 0.58$ | $-0.77 \pm 0.35$ |
| AMNet | 100 | 0.041 | $+5.24 \pm 0.40$ | $+3.49 \pm 0.23$ | $+2.37 \pm 0.48$ | $-1.59 \pm 0.67$ |
| ResMem | 10 | 0.012 | $+0.70 \pm 0.19$ | $+3.80 \pm 0.29$ | $+0.57 \pm 0.39$ | $+0.16 \pm 0.55$ |
| ResMem | 50 | 0.049 | $+2.45 \pm 0.09$ | $+11.05 \pm 0.52$ | $+2.16 \pm 0.39$ | $-1.03 \pm 0.87$ |
| ResMem | 100 | 0.080 | $+4.40 \pm 0.39$ | $+14.33 \pm 0.55$ | $+3.26 \pm 0.52$ | $-4.11 \pm 0.98$ |
| ViTMem | 10 | 0.006 | $+0.55 \pm 0.21$ | $+0.69 \pm 0.53$ | $+1.13 \pm 0.18$ | $-0.04 \pm 0.21$ |
| ViTMem | 50 | 0.024 | $+1.89 \pm 0.20$ | $+2.51 \pm 0.34$ | $+3.96 \pm 0.42$ | $-0.34 \pm 0.30$ |
| ViTMem | 100 | 0.043 | $+3.11 \pm 0.30$ | $+4.19 \pm 0.71$ | $+6.41 \pm 0.38$ | $-1.96 \pm 0.71$ |

At $w = 100$ the held-out gains (Table 2) range from 2.37% to 4.40%, and at $w = 50$ and 100 each critic's own score rises more than either held-out score. CLIP falls for every critic at $w = 100$ (Holm-significant), most under ResMem, whose update at equal $w$ is about twice the others' (update-ratio column). Without the 20 development prompts, the largest transfer $p_{\text{Holm}}$ is 0.0026.

However, average positive transfer conceals substantial instance-level disagreement among the predictors. On the COCO confirmation set, all three held-out models simultaneously register a score increase on only 34.5% of the prompts (69 of 200). Figure 4 shows this proportion for all four studies, together with the Spearman correlations between predictors' paired differences, which on the confirmation set are moderate (0.334 to 0.455). This indicates that while the overall population mean rises, the predictors frequently disagree on which specific images have become more memorable.

**Comparison with best-of-four sampling.** Guidance is not the only way to raise the scores: keeping the best of four unguided samples also does. Best-of-four raises all four memorability predictors over one unguided sample without lowering CLIP (Table 3). Against guidance, no held-out difference is significant (all $p \geq 0.053$); AMNet favors guidance and CLIP favors best-of-four on both datasets (COCO CLIP: $p = 0.0049$). Which method ranks higher thus depends on the score, and nonsignificance is not equivalence. Best-of-four is also faster: at batch size one on an RTX 5090, a guided request takes 6.14 s, best-of-four 4.66 s, and one unguided sample 1.16 s. Parallel sampling could change this ratio (Appendix F).

### 5.2 Qualitative results

These examples show what guidance does to single images; our conclusions rest on the averages over 200 prompts or photos (Tables 1 and 4). In text-to-image generation (Fig. 2), guidance visibly

Table 3: **Best-of-four selection is competitive with guidance.** Entries are the percentage difference between the best-of-four mean score and the comparator's, relative to the comparator's mean (200 prompts per dataset); positive favors best-of-four. The comparator is one unguided image or the guided image (AMNet critic, $w = 100$). P-CLIP is PerceptCLIP.

| Dataset | Comparator | AMNet | ResMem | ViTMem | P-CLIP$^{\dagger}$ | CLIP |
|---|---|---|---|---|---|---|
| DrawBench | One unguided | $+4.20$ | $+2.22$ | $+2.91$ | $+2.60$ | $+1.07$ |
| DrawBench | Guided | $-1.36$ | $-1.16$ | $+0.19$ | $+0.21$ | $+1.85$ |
| COCO | One unguided | $+4.32$ | $+3.80$ | $+2.35$ | $+3.01$ | $+0.41$ |
| COCO | Guided | $-1.68$ | $-0.34$ | $-0.22$ | $+0.33$ | $+1.81$ |

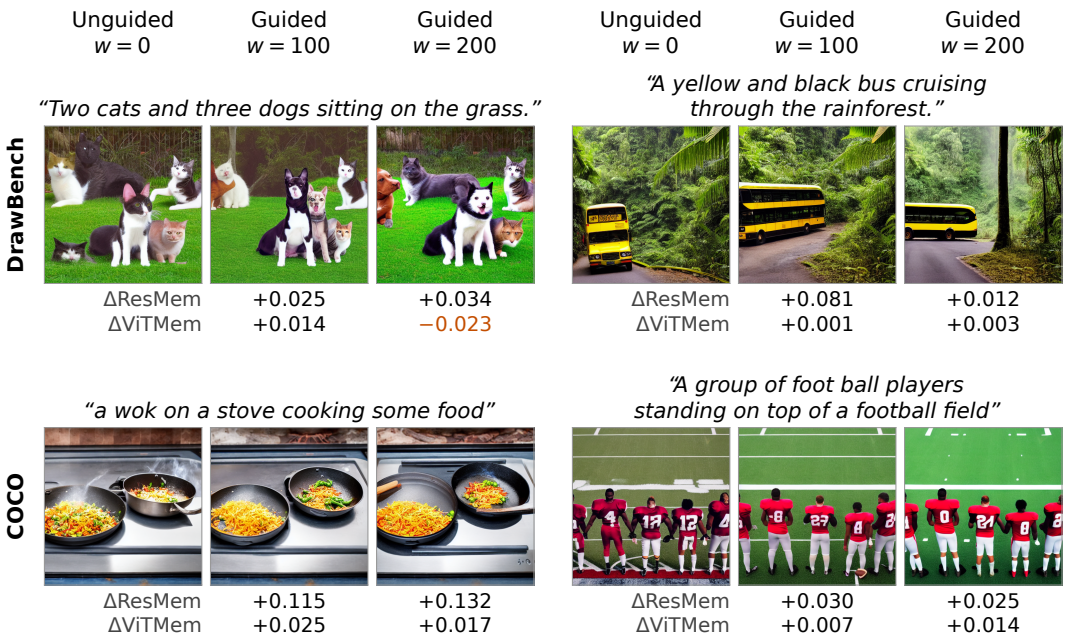


Figure 2: **Guidance changes layout, viewpoint, and color.** SD1.4 with AMNet as the critic. The three images of a prompt start from the same latent. Prompts, verbatim, are at indices 50 (left) and 150 (right) of the DrawBench and COCO prompt lists, chosen by index, not by appearance or score. $\Delta$: change from the unguided image.

changes the image: the bus turns to a side view, the pans move, and the colors get brighter. The held-out scores rise for almost every image, but not for all (ViTMem falls for the cats at $w = 200$), and CLIP even rises for six of the eight guided images, although it falls on average. In editing (Fig. 3), the changes are small and sit mostly on edges and fine detail. Reverse guidance moves ResMem and ViTMem the opposite way in all three photos, and the random direction barely changes the photo. Single photos are noisy, though: for source 11, forward guidance even lowers ResMem. Appendix C shows the same photos at strength 0.35.

### 5.3 Image Editing

We edit the 200 SUN test photos at strength 0.35 with AMNet as the critic and ask whether the held-out scores follow the direction of the critic's gradient. The fair baseline is the matched reconstruction, not the original photo, because encoding and denoising alone already change a photo's scores: reconstruction by itself raises ViTMem by $0.00901 \pm 0.00164$, so only $0.01333 \pm 0.00076$ of the $0.02234 \pm 0.00181$ gain over the source comes from guidance (Table 4).

Forward guidance raises all three held-out predictors, each on at least 182 of the 200 photos (ResMem: $d_z = 1.14$, $p = 7.26 \times 10^{-38}$). Reverse guidance lowers all three. A random push of the same size changes none of them significantly; its only nominally significant change is a tiny drop in AMNet ($-0.00010 \pm 0.00013$, $p = 0.037$). The gain therefore comes from where the critic points, not from disturbing the sampler. As expected, the critic's own score rises most (AMNet $+13.68 \pm 0.28\%$), which only confirms that the optimization worked. The human scores describe only the original photos; in an exploratory one-seed run on all 2,222 photos at strength 0.50, AMNet loses more rank agreement with them after editing than any held-out predictor (Appendix E).

**Cost to the photo.** Guidance changes the photo slightly more than reconstruction does: LPIPS to the source rises by $0.00294 \pm 0.00010$ and CLIP-I falls by $0.00487 \pm 0.00135$. A pass/fail count hides this cost, because the same 154 of 200 photos meet all three preservation thresholds with and without guidance. Photos with larger ResMem gains also lose more of the source (Spearman 0.364

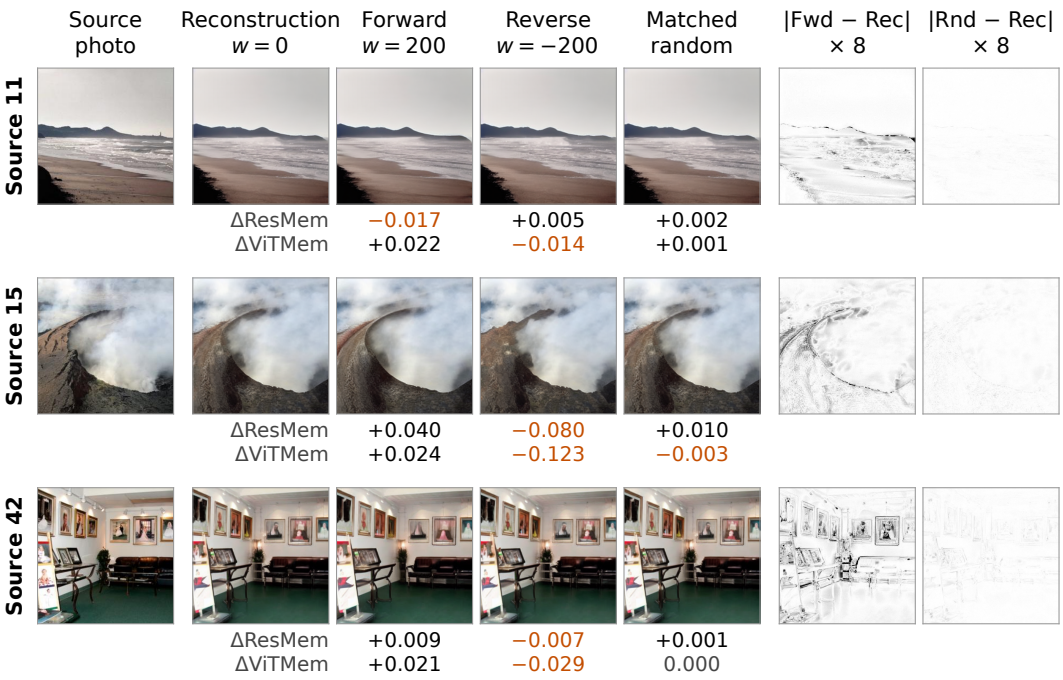


Figure 3: **Forward guidance changes the reconstruction far more than the matched random direction does.** SD1.4, AMNet as the critic. Each row's outputs start from one noised source ; rows are the first three SUN test photos by ID. Last two columns: absolute difference of the forward (Fwd) and matched random (Rnd) outputs from the reconstruction (Rec), averaged over RGB and amplified $8\times$ (darker is larger).

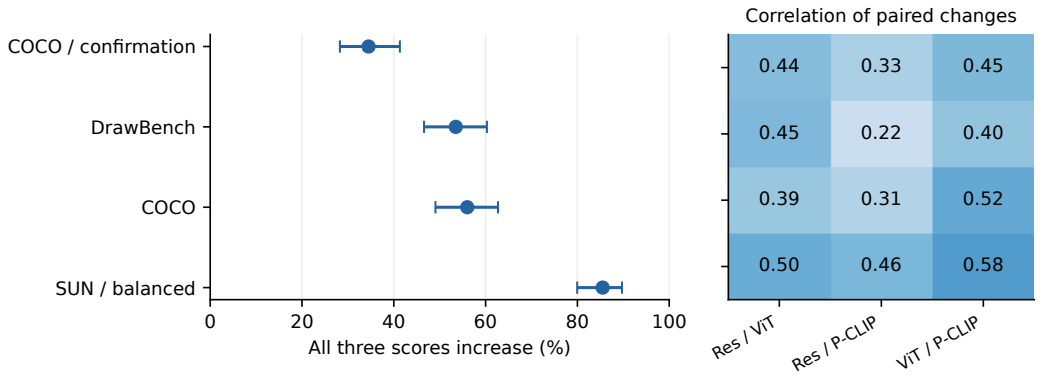


Figure 4: **The held-out predictors agree on the average gain, but less on which images gain.** Using AMNet as critic: Left: share of the 200 prompts or photos on which ResMem, ViTMem, and PerceptCLIP (Res, ViT, P-CLIP) all rise under guidance, with 95% Wilson intervals. Right: Spearman correlation between two predictors' paired differences. Rows differ in task and seed count; "SUN / balanced" is editing at strength 0.35.

between the ResMem gain and the LPIPS increase). At strength 0.50, scores rise more and more of the source is lost (Appendix C).

**Agreement with the human rankings.** Guidance also changes how well the predictors rank the SUN photos as people did. In an exploratory run over all 2,222 photos, AMNet's Spearman correlation with the source-image human scores falls from 0.626 to 0.264 after guided editing, while the drops for ResMem, ViTMem, and PerceptCLIP are all near $-0.11$ (Table 10 in Appendix E).

Table 4: **Reverse guidance reverses the effect; a random direction changes almost nothing.** Held-out predictor scores for 200 SUN test photos edited at strength 0.35 with AMNet as the critic. $\Delta$ is the change from matched reconstruction and Relative is $\Delta$ over the reconstruction mean, both mean $\pm$ SD over 3 seeds. Forward and Reverse use $w = \pm 200$; Matched random keeps the per-step gradient norm in a random direction. W/L counts photos whose seed-averaged score rose/fell.

| Direction | Predictor | $\Delta$ | Relative (%) | W/L |
|---|---|---|---|---|
| Forward | ResMem | $+0.02468 \pm 0.00074$ | $+3.64 \pm 0.11$ | 191/9 |
| Forward | ViTMem | $+0.01333 \pm 0.00076$ | $+1.98 \pm 0.11$ | 195/5 |
| Forward | PerceptCLIP$^{\dagger}$ | $+0.00956 \pm 0.00044$ | $+1.49 \pm 0.07$ | 182/18 |
| Reverse | ResMem | $-0.02038 \pm 0.00270$ | $-3.01 \pm 0.39$ | 15/185 |
| Reverse | ViTMem | $-0.01254 \pm 0.00037$ | $-1.86 \pm 0.05$ | 5/195 |
| Reverse | PerceptCLIP$^{\dagger}$ | $-0.01010 \pm 0.00011$ | $-1.57 \pm 0.02$ | 16/184 |
| Matched random | ResMem | $+0.00003 \pm 0.00008$ | $0.00 \pm 0.01$ | 100/100 |
| Matched random | ViTMem | $-0.00005 \pm 0.00012$ | $-0.01 \pm 0.02$ | 86/114 |
| Matched random | PerceptCLIP$^{\dagger}$ | $+0.00008 \pm 0.00030$ | $+0.01 \pm 0.05$ | 102/98 |

Reconstruction alone accounts for part of each drop; measured against the matched reconstruction, AMNet still falls by $0.337$ and the held-out predictors by $0.043$ to $0.074$. On the 200 test photos at strength 0.35, only AMNet's agreement falls significantly, from $0.634$ to $0.385$. The human scores describe only the original photographs, so this is not a behavioural evaluation of the edited images, but it suggests that optimizing the critic can move its ranking away from the human-grounded one.

## 6 Limitations

Our experimental results rest on memorability predictors rather than on human recognition, which we did not measure for the generated or edited images. The predictors are held out in the sense that they supply no gradients during guidance, but they are not fully independent, since all four are trained on LaMem. We therefore read their agreement as consistency among learned scores, and treat the critic's own gain only as a manipulation check.

The most prominent constraint of memorability guidance is the inherent trade-off between score optimization and visual fidelity. Pushing the guidance weight to aggressive levels yields higher memorability scores but precipitates a steeper decline in prompt alignment. Furthermore, transferring the guidance mechanism from Stable Diffusion 1.4 to a higher-resolution architecture (SDXL at $1024^2$) reveals severe structural fragility. Even with perturbation capping, antialiased critic inputs, and delayed guidance schedules, the SDXL diagnostic generated visibly malformed outputs (e.g., aggressive subject tiling) while paradoxically reporting increased memorability scores.

Methodologically, our analyses were not preregistered. The COCO confirmation set was fixed locally, the selection of SUN editing thresholds was informed by ResMem, and the inclusion of PerceptCLIP was strictly post hoc. Furthermore, the rotating-critic experiment was conducted retrospectively with guidance weights that were not matched by update size.

## 7 Conclusion

We asked whether memorability guidance transfers beyond the predictor that supplies its gradient. Across two prompt sets and real-image editing, held-out scores rise on average, while rotating the guiding critic on DrawBench shows that the transfer is not specific to AMNet. Matched controls further show that the gains follow the critic's gradient direction rather than perturbation magnitude alone. The gains are nevertheless modest and uneven: the held-out predictors agree on only 69 of 200 COCO image pairs, prompt fidelity falls, selecting the best of four unguided samples is competitive and faster, and on SDXL predictor scores can rise despite visibly malformed outputs. A rising critic score is therefore where the evaluation of memorability guidance should begin, not where it should end.

### AI USE STATEMENT

In this work, we used generative AI tools, including OpenAI Codex and Anthropic Claude, to inspect the implementation and the experiment records, to write the statistical re-analysis code, which checks the paired score files and recomputes the reported statistics, and to help interpret the existing results; this included revising claims to distinguish model predictions from human outcomes. Apart from the diffusion models and memorability predictors that are the objects of study, we have not used generative AI tools to produce data or evidence: every reported image and measurement comes from the experiments described in the paper, and no language-model-generated measurements or synthetic human responses are used. Mathematical proofs and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to organize, draft, and rewrite the manuscript text and captions, to search and summarize related literature, to write figure-generation code, and to prepare the LaTeX source and bibliography. We have reviewed all AI-assisted work. The authors checked the AI-assisted analysis, text, and artifacts before submission. The analysis scripts regenerate the reported tables from saved per-image scores after verifying image identifiers, paired membership, split disjointness, and manifest hashes (Appendix A). We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work studies model-predicted memorability using existing image datasets and pretrained models. We collected no participant responses; the SUN human scores describe the source photographs only. Any future behavioral study requires institutional review, informed consent, and a protocol fixed in advance. Memorability control could support communication and learning, but it could also make manipulative or misleading content more salient. Agreement among predictors should not be presented as evidence of a psychological effect. Generated images may inherit biases or unsafe content from the models and their source data.

### REPRODUCIBILITY STATEMENT

Section 3 specifies the method and its controls; Section 4 defines the data, pairing, metrics, and statistical analysis. Appendix A records architecture, preprocessing, sampler settings, statistical definitions, and checkpoint hashes. Appendices B–D give extended results and the SDXL diagnostic, Appendix E the per-image agreement analyses, and Appendix F the latency protocol. The anonymized supplementary material contains the guidance and benchmark code, prompt lists and source-image manifests, per-image score tables, and scripts that regenerate the numerical tables and quantitative figures without loading a diffusion checkpoint. It does not include model weights, so regenerating images requires the external checkpoints identified in Appendix A. Outside the rotating-critic experiment and the SDXL diagnostic, guided sampling is not bitwise repeatable under our settings.Appendix F reports a strict deterministic configuration that restored identical repeats on two diagnostic prompts; among the SD1.4 experiments, only the rotating-critic experiment was run under it.

## REFERENCES

Tahsin Tariq Banna, Swakshar Deb, Sejuti Rahman, and Shafin Rahman. Gemm: A graph embedded model for memorability prediction. In *IJCNN*, pp. 1–8. IEEE, 2023.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 51304–51323, 2024.

Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015.

Shengqi Dang, Yi He, Jiaying Lei, Ziqing Qian, and Nan Cao. Cogblender: Towards continuous cognitive intervention in text-to-image generation. *arXiv preprint arXiv:2603.09286*, 2026.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6363–6372. IEEE, 2018.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5743–5752. IEEE, 2019.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jingyi Huang and Edgar Simo-Serra. Image memorability manipulation with intermediate guided diffusion. In *Proceedings of the 2024 7th Artificial Intelligence and Cloud Computing Conference*, pp. 229–236, 2024.

Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. *Advances in neural information processing systems*, 24, 2011a.

Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*, pp. 145–152. IEEE, 2011b.

Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1469–1482, 2013.

Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pp. 2390–2398, 2015.

Cameron Kyle-Davidson, Adrian G Bors, and Karla K Evans. Modulating human memory for complex scenes with artificially generated images. *Scientific Reports*, 12(1):1583, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Coen D Needell and Wilma A Bainbridge. Embracing new techniques in deep learning for estimating image memorability. *Computational Brain & Behavior*, 5(2):168–184, 2022.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2004.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. How to make an image more memorable? a deep style transfer approach. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ICMR '17, pp. 322–329, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450347013. doi: 10.1145/3078971.3078986. URL https://doi.org/10.1145/3078971.3078986.

Oleksii Sidorov. Changing the image memorability: From basic photo editing to gans. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 790–799, 2019. doi: 10.1109/CVPRW.2019.00107.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Marquant Gwenaelle, and Claire-Helene Demarty. Deep learning for predicting image memorability. In *ICASSP*, pp. 2371–2375, 2018. doi: 10.1109/ICASSP.2018.8462292.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Amit Zalcher, Navve Wasserman, Roman Beliy, Oliver Heinimann, and Michal Irani. Don't judge before you clip: A unified approach for perceptual tasks. *Transactions on Machine Learning Research*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

# A Implementation and reproducibility details

## A.1 Sampler equations and editing procedure

Let $\varnothing$ be the empty prompt; the other symbols are those of Sections 3.1 and 3.2. Classifier-free guidance (CFG) with scale $s$ and the implied clean-latent estimate are

$$\epsilon_{\mathrm{cfg}}(z_t, t, c) = \epsilon_\theta(z_t, t, \varnothing) + s\big(\epsilon_\theta(z_t, t, c) - \epsilon_\theta(z_t, t, \varnothing)\big), \tag{3}$$

$$\widehat{z}_0(z_t) = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_{\mathrm{cfg}}(z_t, t, c)}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

The gradient $g_t$ of Eq. 1 is backpropagated through both CFG passes of the U-Net, so it also accounts for how the noise estimate reacts to a change in $z_t$; treating $\epsilon_{\mathrm{cfg}}$ as a constant would be cheaper but would ignore this. With the guidance shift $u_t = w\sqrt{1-\bar{\alpha}_t}\,g_t$, Eq. 2 reads $\widetilde{\epsilon}_t = \epsilon_{\mathrm{cfg}} - u_t$, and the deterministic DDIM update (Song et al., 2021) to the next, less noisy timestep $t^-$ is

$$z_{t^-} = \sqrt{\bar{\alpha}_{t^-}}\,\frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\widetilde{\epsilon}_t}{\sqrt{\bar{\alpha}_t}} + \sqrt{1-\bar{\alpha}_{t^-}}\,\widetilde{\epsilon}_t. \tag{5}$$

Under the identification $\epsilon \approx -\sqrt{1-\bar{\alpha}_t}\,\nabla_{z_t}\log p_t(z_t)$, Eq. 2 adds $w\,g_t$ to the score implied by $\epsilon_{\mathrm{cfg}}$, as in classifier guidance (Dhariwal & Nichol, 2021), with the critic's score in place of the classifier log-likelihood. Because that score is a regression output, not a log-density, $w$ has no probabilistic calibration, so we chose it empirically (Fig. 5). The SD1.4 runs do not cap $u_t$.

**Editing.** A source photo $x$ is center-cropped and resized to $512^2$, encoded once with the mode of the VAE posterior, and noised as in SDEdit (Meng et al., 2022):

$$z_{t_s} = \sqrt{\bar{\alpha}_{t_s}}\,E_{\mathrm{mode}}(x) + \sqrt{1-\bar{\alpha}_{t_s}}\,\xi, \qquad \xi \sim \mathcal{N}(0, I). \tag{6}$$

For editing strength $\gamma$, $t_s$ is the $k$-th least noisy of the 20 schedule timesteps, with $k = \mathrm{round}(20\gamma)$, and sampling runs only these last $k$ steps. The prompt is empty, so Eq. 3 reduces to $\epsilon_\theta(z_t, t, \varnothing)$ and only the critic steers. Reconstruction, guidance, and both controls start from the same noised latent $z_{t_s}$.

**Norm-matched random control.** At every step, this control replaces $g_t$ by a random direction of the same length,

$$g_t^{\mathrm{rnd}} = \frac{\|g_t\|_2}{\max(\|r_t\|_2, 10^{-12})}\,r_t, \qquad r_t \sim \mathcal{N}(0, I), \tag{7}$$

where $r_t$ is redrawn at every step and $\|g_t\|_2$ is the trained critic's gradient norm at the control's own current latent.

## A.2 Critics and the rotating-critic run

**Preprocessing.** During guidance, the critic's preprocessing $P$ clamps the decoded image to $[-1, 1]$ and maps it to $[0, 1]$. For AMNet, $P$ then resizes to $224^2$ bilinearly without antialiasing and applies the ImageNet normalization, whereas offline AMNet scoring resizes with antialiasing. For ResMem and ViTMem, $P$ reimplements the released transforms on tensors: an antialiased resize to $256^2$, then a center crop to $227^2$ for ResMem and $224^2$ for ViTMem. On one synthetic test image, these tensor transforms change the scores by less than $2 \times 10^{-4}$. Every reported score is computed offline from the saved final image with the predictor's released preprocessing. PerceptCLIP is a LoRA-adapted CLIP ViT-L/14 encoder with an MLP head (Zalcher et al., 2025).

**Rotating-critic run.** This run uses strict deterministic execution (Appendix F): the runner generates the first prompt of every arm twice and stops unless the two final latents are identical. It reuses the prompts and starting latents of the four-seed DrawBench experiment, and its unguided images reproduce that experiment's unguided scores exactly. The update ratio of Table 2 is $\|u_t\|_2 / \|\epsilon_{\mathrm{cfg}}\|_2$, averaged over steps and images.

### A.3 STATISTICS

A unit is a prompt or a source photo, and its paired difference $d_i$ (Section 4) is averaged over the unit's seeds. Tests are two-sided paired $t$ tests over units, and $d_z = \bar{d}/s_d$ is the paired effect size, where $\bar{d}$ and $s_d$ are the mean and SD of the $d_i$. A multi-seed entry $a \pm b$ gives the mean and SD over seeds of the per-seed result. Holm correction covers the planned held-out tests (ResMem and ViTMem) within each prompt set, the held-out tests and, separately, the CLIP tests of the rotating-critic experiment, and the single primary editing test, ResMem at strength 0.35. PerceptCLIP was added post hoc and is not corrected. The non-inferiority test $p_{\mathrm{NI}}$ asks whether the CLIP loss is smaller than 5% of the unguided mean.

## B ADDITIONAL TEXT-TO-IMAGE RESULTS

Table 5: **Memorability gains grow with $w$ and reverse under $w = -100$; the untrained critic has no effect.** Mean change (%) from the unguided image on DrawBench (200 prompts, seed base 42, AMNet critic; no tests). Untrained critic: AMNet without its trained weights at $w = 100$, a weak control because the size of its push is not matched. Prompt suffix: one fixed phrase appended to each prompt, without guidance.

| Arm | AMNet | ResMem | ViTMem | CLIP |
|---|---|---|---|---|
| $w = 50$ | $+3.13$ | $+1.34$ | $+1.65$ | $-0.30$ |
| $w = 100$ | $+5.64$ | $+3.42$ | $+2.72$ | $-0.77$ |
| $w = 200$ | $+8.71$ | $+4.80$ | $+4.74$ | $-4.23$ |
| $w = -100$ (reverse) | $-7.12$ | $-4.12$ | $-3.10$ | $-0.82$ |
| Untrained critic | $+0.02$ | $-0.03$ | $-0.08$ | $+0.03$ |
| Prompt suffix | $-0.97$ | $-0.73$ | $-0.83$ | $-1.43$ |
| Suffix + guidance | $+4.84$ | $+2.23$ | $+1.84$ | $-2.71$ |

**Controls.** In Table 5, the gains grow with $w$ and change sign under reverse guidance, while an untrained critic moves every score by less than $0.1\%$. The prompt suffix ("highly memorable, vivid colors, dramatic lighting, striking composition"), with CLIP still measured against the original prompt, lowers all four scores on its own and weakens guidance when combined with it.

**Best-of-$K$ selection.** Best-of-$K$ keeps, among the unguided images at the first $K$ seed bases, the one with the highest offline AMNet score. Best-of-three is weaker than best-of-four: over one unguided sample, it raises ResMem by $1.63\%$ on DrawBench and $3.36\%$ on COCO, against $2.22\%$ and $3.80\%$ for best-of-four (Table 3).

### B.1 DEVELOPMENT DOSE RESPONSE

Before the main runs, we swept eight guidance weights from 2 to 500 on the first 20 DrawBench prompts (Fig. 5). AMNet rises steadily with $w$, while CLIP drops sharply beyond $w = 100$; we took $w = 100$ as the largest weight before this drop. With only 20 prompts, the held-out predictors are noisy, and none of them rises steadily with $w$.

## C SUN EDITING: SETTINGS, PRESERVATION, AND EXAMPLES

**Metrics and settings.** SSIM (Wang et al., 2004), LPIPS with an AlexNet backbone (Zhang et al., 2018), and CLIP-I, the cosine similarity of CLIP ViT-B/32 image embeddings (Radford et al., 2021), compare each output with its source photo. We chose the editing setting on the 20 development photos, among strengths 0.20, 0.35, and 0.50 and weights 50, 100, and 200, keeping the pairs whose mean metrics met SSIM $\geq 0.60$, LPIPS $\leq 0.40$, and CLIP-I $\geq 0.75$. The largest ResMem gain was at strength 0.50 with $w = 200$, but only 10 of the 20 photos met all three thresholds there, against 17 at strength 0.35. Before scoring any test photo, we therefore made strength 0.35 the primary setting and kept 0.50 as the secondary one.

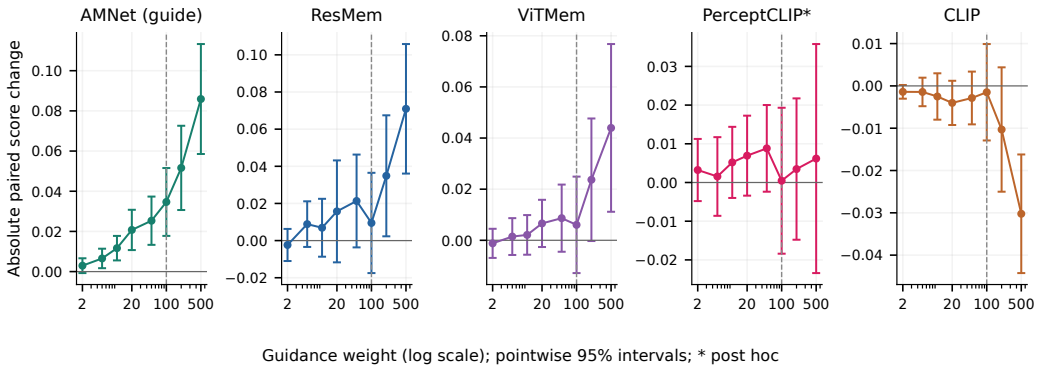


Figure 5: **On the development prompts, AMNet keeps rising with $w$ while CLIP drops beyond $w = 100$.** Mean change from the unguided image on the first 20 DrawBench prompts (AMNet critic), with pointwise 95% $t$ intervals; the $w$ axis is logarithmic and the dashed line marks $w = 100$. *Added post hoc.

Table 6: **Most drift from the source is already present in the reconstruction.** Mean SSIM, LPIPS, and CLIP-I between outputs and their source photos (200 SUN test photos, three seeds, AMNet critic). Balanced and Aggressive are strengths 0.35 and 0.50; Forward and Reverse use $w = \pm 200$. Pass counts the photos whose seed-averaged metrics meet all three thresholds.

| Setting | Arm | SSIM ↑ | LPIPS ↓ | CLIP-I ↑ | Pass |
|---|---|---|---|---|---|
| Balanced | Reconstruction | 0.691 | 0.232 | 0.913 | 154/200 |
| Balanced | Forward | 0.690 | 0.235 | 0.909 | 154/200 |
| Balanced | Reverse | 0.690 | 0.235 | 0.913 | 154/200 |
| Balanced | Matched random | 0.691 | 0.232 | 0.914 | 154/200 |
| Aggressive | Reconstruction | 0.615 | 0.322 | 0.863 | 107/200 |
| Aggressive | Forward | 0.614 | 0.333 | 0.843 | 100/200 |
| Aggressive | Reverse | 0.612 | 0.332 | 0.857 | 104/200 |
| Aggressive | Matched random | 0.615 | 0.322 | 0.863 | 107/200 |

**Stronger editing.** Table 7 extends Table 4 at strength 0.50, on the test photos and, with one seed, on all 2,222 photos. Guidance and reverse guidance move the held-out scores further than at strength 0.35 and in the same directions, while the random direction again changes nothing significantly. The random control matches the size of each step, not its effect: it leaves the output almost unchanged (Table 6; Fig. 6).

Table 7: **At strength 0.50, guidance and reverse guidance move the held-out scores further, in the same directions.** Change from matched reconstruction (AMNet critic, $w = \pm 200$) in score units ($\Delta$) and in percent of the reconstruction mean; W/L counts the photos whose score rose/fell. Test set: mean $\pm$ SD over 3 seeds. All SUN: one seed on all 2,222 photos, development and test included, so it is exploratory; every change has $p < 10^{-200}$. †Added post hoc.

| Direction | Predictor | $\Delta$ | Relative (%) | W/L |
|---|---|---|---|---|
| *Test set, 200 photos, 3 seeds* | | | | |
| Forward | ResMem | $+0.05699 \pm 0.00224$ | $+8.48 \pm 0.37$ | 194/6 |
| Forward | ViTMem | $+0.03773 \pm 0.00126$ | $+5.66 \pm 0.20$ | 198/2 |
| Forward | PerceptCLIP† | $+0.02712 \pm 0.00072$ | $+4.32 \pm 0.12$ | 186/14 |
| Reverse | ResMem | $-0.04205 \pm 0.00243$ | $-6.26 \pm 0.35$ | 12/188 |
| Reverse | ViTMem | $-0.03182 \pm 0.00190$ | $-4.77 \pm 0.30$ | 4/196 |
| Reverse | PerceptCLIP† | $-0.02387 \pm 0.00146$ | $-3.80 \pm 0.22$ | 16/184 |
| Matched random | ResMem | $+0.00053 \pm 0.00076$ | $+0.08 \pm 0.11$ | 99/101 |
| Matched random | ViTMem | $+0.00006 \pm 0.00019$ | $+0.01 \pm 0.03$ | 98/102 |
| Matched random | PerceptCLIP† | $-0.00007 \pm 0.00028$ | $-0.01 \pm 0.05$ | 93/107 |
| *All SUN, 2,222 photos, 1 seed* | | | | |
| Forward | ResMem | $+0.0504$ | $+7.54$ | 1925/297 |
| Forward | ViTMem | $+0.0311$ | $+4.68$ | 2009/213 |
| Forward | PerceptCLIP† | $+0.0245$ | $+3.88$ | 1895/327 |
| Reverse | ResMem | $-0.0373$ | $-5.58$ | 382/1840 |
| Reverse | ViTMem | $-0.0285$ | $-4.29$ | 196/2026 |
| Reverse | PerceptCLIP† | $-0.0223$ | $-3.53$ | 288/1934 |

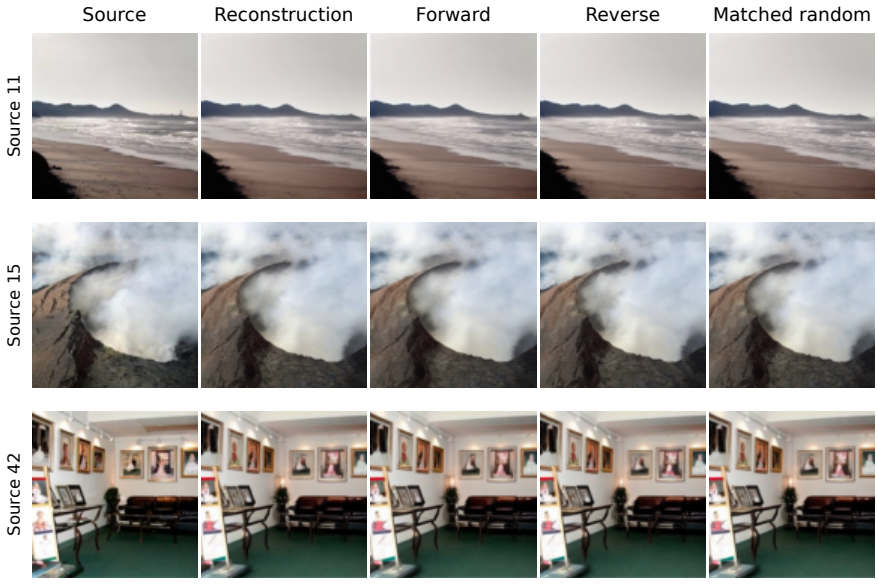


Figure 6: **At strength 0.35, the four outputs of a photo differ only subtly.** Columns: source photo; reconstruction ($w = 0$); forward ($w = 200$); reverse ($w = -200$); matched random. All outputs in a row start from the same noised source (AMNet critic, seed base 42). Rows are the first three SUN test photos by ID, not chosen by score.

## D SDXL DIAGNOSTIC

To see whether the SD1.4 guidance rule carries over to a larger generator, we applied it to SDXL (Podell et al., 2024) on five DrawBench development prompts with one seed. The outcome is a failure case, not a benchmark: memorability scores rise while the images break (Fig. 7).

**Setup.** We use SDXL base 1.0 (Hugging Face revision `462165984030`) at its native $1024^2$ resolution, with 20 DDIM steps, CFG 5.0, $w = 100$, AMNet as the critic, and strict deterministic execution. Unlike the SD1.4 runs, the guidance shift is capped at 5% of the size of the CFG noise estimate:

$$u_t^{\text{cap}} = u_t \min\left(1, \frac{\kappa \|\epsilon_{\text{cfg}}\|_2}{\max(\|u_t\|_2, 10^{-12})}\right), \qquad \kappa = 0.05. \tag{8}$$

**Arms.** All arms share the prompt and the starting latent and are compared with the unguided image. *Transferred* applies the SD1.4 rule at every step, including AMNet's non-antialiased resize. *Antialias* (AA) antialiases the critic input instead. *AA + delay* also skips guidance on the first three steps. *Matched random* is AA + delay with the norm-matched random direction of Eq. 7.

**Results.** In all three guided arms, AMNet and ViTMem rise, but the other held-out predictors disagree: ResMem falls in the two arms without delay and PerceptCLIP falls in the arm with delay (Table 8). The red car stays malformed in every guided arm. Antialiasing and delay make the updates smaller, but they do not repair the images.

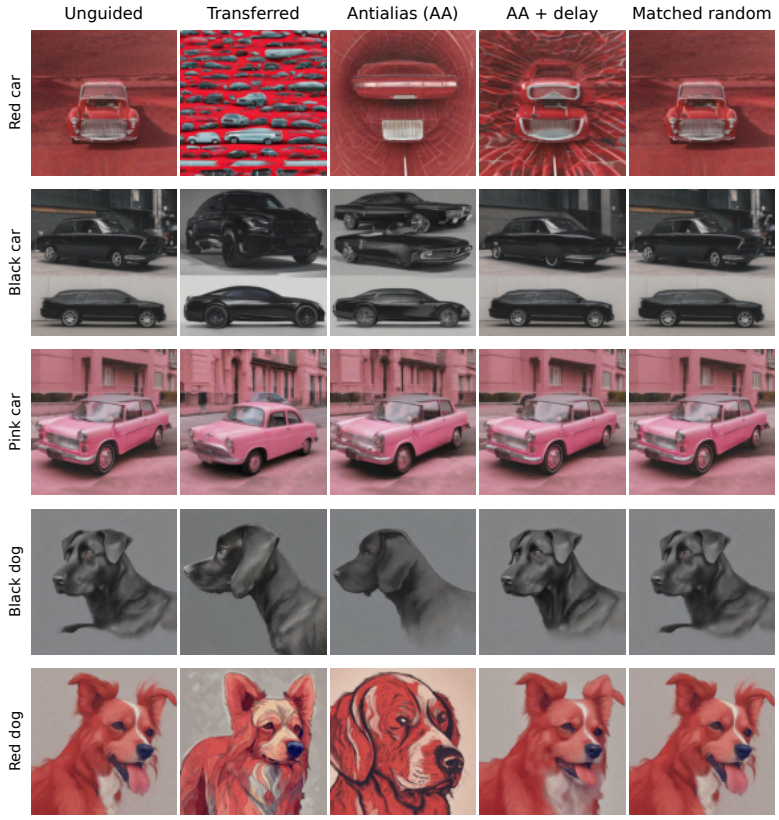


Figure 7: **The red car stays malformed in all three guided arms.** Every image of the SDXL diagnostic: rows are the five prompts, columns the unguided image and the four arms, all from the same starting latent. AA is Antialias.

Table 8: **SDXL: the memorability scores disagree in sign within every guided arm.** Mean change from the unguided image in score units (five prompts, one seed, $w = 100$, AMNet critic; no tests). †P-CLIP is PerceptCLIP, added post hoc; CLIP is text–image similarity.

| Arm | Δ AMNet | Δ ResMem | Δ ViTMem | Δ P-CLIP† | Δ CLIP |
|---|---|---|---|---|---|
| Transferred | +0.04089 | −0.00164 | +0.05350 | +0.00952 | −0.00103 |
| Antialias | +0.04376 | −0.01833 | +0.05097 | +0.00408 | −0.00986 |
| Antialias + delay | +0.02679 | +0.01771 | +0.03284 | −0.01660 | −0.00205 |
| Matched random + delay | +0.00031 | −0.00069 | −0.00024 | −0.00124 | −0.00005 |

## E PER-IMAGE AGREEMENT AND HUMAN RANK AGREEMENT

All analyses in this appendix use AMNet as the critic and are post hoc.

**Joint direction of the held-out predictors.** Table 9 counts the units on which all three held-out predictors rise, all fall, or disagree. The all-positive share is highest in editing and lowest on the one-seed confirmation set, partly because averaging seeds reduces noise: at single seeds, the all-positive counts on DrawBench are 68 to 78 of 200, against 107 after averaging.

Table 9: **All three held-out predictors rise together least often on the one-seed confirmation set and most often in SUN editing.** Each row sorts 200 units (prompts or photos) by the signs of their seed-averaged changes in ResMem, ViTMem, and PerceptCLIP. Changes are from the unguided image (text-to-image) or the matched reconstruction (SUN; balanced: strength 0.35, aggressive: 0.50).

| Study | All + | Rate | Mixed | All − |
|---|---|---|---|---|
| COCO confirmation | 69/200 | 34.5% | 101 | 30 |
| DrawBench, 4 seeds | 107/200 | 53.5% | 88 | 5 |
| COCO, 4 seeds | 112/200 | 56.0% | 81 | 7 |
| SUN, balanced | 171/200 | 85.5% | 29 | 0 |
| SUN, aggressive | 179/200 | 89.5% | 21 | 0 |

**Rank agreement with human scores.** SUN is the only set with human memorability scores, and these describe the original photos. Table 10 asks whether each predictor still ranks the photos as people did after editing. The critic, AMNet, loses the most agreement, and part of every loss already comes from reconstruction alone. On the 200 test photos at strength 0.35, only AMNet's agreement falls significantly (from 0.634 to 0.385).

Table 10: **The critic, AMNet, loses about three times more rank agreement with human scores than any held-out predictor.** Spearman $\rho$ between predictor scores and SUN human memorability scores for all 2,222 photos, before and after guided editing (strength 0.50, $w = 200$, one seed; exploratory). People rated only the originals. All four changes differ from zero (paired bootstrap, 10,000 resamples).

| Predictor | Original photos | Edited photos | Change |
|---|---|---|---|
| AMNet (critic) | 0.626 | 0.264 | −0.361 |
| ResMem | 0.615 | 0.511 | −0.104 |
| ViTMem | 0.811 | 0.701 | −0.110 |
| PerceptCLIP | 0.738 | 0.623 | −0.116 |

**Fidelity on the same units.** Agreement among the held-out predictors does not protect fidelity: CLIP falls on 62 of the 107 all-positive DrawBench prompts and on 68 of the 112 all-positive COCO prompts. In editing at strength 0.50, seven photos that meet all three preservation thresholds as reconstructions fail them after guidance, and none move the other way.

## F LATENCY, MEMORY, AND SAME-SEED REPEATABILITY

We time four request types with SD1.4 in float32 at batch size one on one RTX 5090: one unguided image, one guided image ($w = 100$), and best-of-three and best-of-four selection. Twelve prompts (six from DrawBench and six from COCO) each receive two repetitions of every request type, in randomized order after one warm-up per type. Timing covers prompt encoding through final decoding and, for best-of-$K$, AMNet scoring and selection; model loading, disk writes, and held-out scoring are excluded. Peak memory is the maximum allocated by PyTorch (Table 11).

Table 11: **A guided request takes the longest and needs about twice the memory of the others.** Mean wall-clock time per request after warm-up, with the SD over the 12 prompts, and peak GPU memory allocated by PyTorch.

| Request | Seconds | SD | Peak GiB |
|---|---|---|---|
| One unguided | 1.16 | 0.004 | 6.33 |
| Guided $w = 100$ | 6.14 | 0.011 | 12.37 |
| Best-of-three | 3.50 | 0.013 | 6.33 |
| Best-of-four | 4.66 | 0.014 | 6.33 |

**Same-seed repeatability.** Unguided and best-of-$K$ requests reproduce exactly from a seed, but guided requests do not: two guided runs from one seed differ by a mean absolute pixel difference of about 20 to 23 on the 0–255 scale, because the guided backward pass is nondeterministic on the GPU. Strict deterministic execution (deterministic cuBLAS, cuDNN, and PyTorch algorithms) made guided repeats identical on the two prompts we tested; among the SD1.4 experiments, only the rotating-critic run used it.