EMAG: Self-Rectifying Diffusion Sampling with Exponential Moving Average Guidance

Ankit Yadav[1], Ta Duc Huy[1], and Lingqiao Liu[1]

Australian Institute for Machine Learning, Adelaide University, Australia
{ankit.yadav, huy.ta, lingqiao.liu}@adelaide.edu.au

Abstract. In diffusion and flow-matching generative models, guidance techniques are widely used to improve sample quality and consistency. Classifier-free guidance (CFG) is the de facto choice in modern systems and achieves this by contrasting conditional and unconditional samples. Recent work explores contrasting negative samples at inference using a weaker model, via strong/weak model pairs, attention-based masking, stochastic block dropping, or perturbations to the self-attention energy landscape. While these strategies refine the generation quality, they still lack a reliable control over the granularity or difficulty of the negative samples, and target-layer selection is often fixed. We propose *Exponential Moving Average Guidance (EMAG)*, a training-free mechanism that modifies attention at inference time in diffusion transformers, with a statistics-based, adaptive layer-selection rule. Unlike prior methods, EMAG produces harder, semantically faithful negatives (fine-grained degradations), surfacing difficult failure modes, enabling the denoiser to refine subtle artifacts, boosting the quality and human preference score (HPS) by $+\mathbf{0.54}$ over CFG. We further demonstrate that EMAG naturally composes with advanced orthogonal guidance techniques, such as APG and CADS, further improving HPS. Code is available at `https://github.com/drkkgy/EMAG`.

Keywords: Classifier-Free Guidance (CFG) · Diffusion Transformers · Human Preference Score (HPS)

1 Introduction

In recent times, Denoising Diffusion Models (DDMs) [6,11,30,32] and flow-matching models [7, 19] have surpassed GANs [8] at modeling data across modalities (images [26], audio [33], video [3]). Their success stems from an iterative denoising process that starts from Gaussian noise and progressively removes it to sample the target distribution [17]. The advent of conditional diffusion models, where a condition such as a class [6] or text prompt [23] can be used to guide the generation, further propelled their popularity. However, naive guidance signals often result in insufficient quality and weaker prompt adherence [21], leading to the introduction of strategies like classifier guidance [6] and classifier-free guidance [12], where the prior method requires an explicit guidance model while

the latter relies on the implicit Bayesian classifier of the same diffusion model by contrasting the conditional with the unconditional signal. This substantially improved conditional generation quality and prompt adherence, but can often reduce diversity and induce over-saturation [28]. Recent work addresses these effects by providing more semantically aware guidance (S-CFG) [29], decomposing the CFG update into parallel and orthogonal components and down-weights the parallel term (with rescaling/momentum) to mitigate oversaturation (APG) [28], and by perturbing the conditioning signal over an annealed schedule improving diversity (CADS) [27].

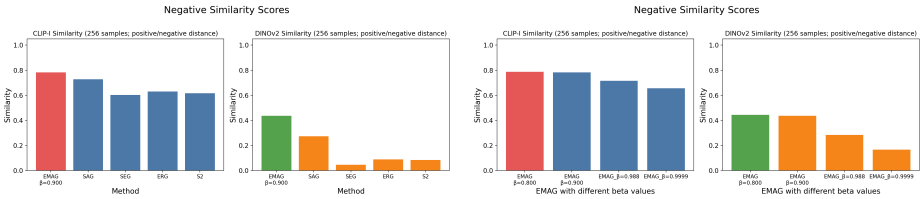
Another line of work exposes failure modes by contrasting strong and weak predictions. Such negative (weak) signals are obtained by perturbing attention maps [1, 14] or weights [13], by guiding with deliberately weaker models (reduced capacity or degraded variants) [16], by sampling sub-networks [4], or via energy-based formulations such as ERG [15] and SEG [13]. These strategies are attractive because, unlike CFG, they apply to both conditional and unconditional settings and often yield *Pareto-favorable* trade-offs. However, most approaches provide a mechanism to obtain an inferior model but do not explicitly construct *hard negatives*. We hypothesize that if degradations are too strong (e.g., heavily distorting an image), they represent failure modes that modern diffusion models are implicitly powerful enough to avoid [7, 25], yielding *easy* negatives.

Fig. 1: CFG (left) vs. EMAG (right) Compared to CFG [12], EMAG produces more *semantically plausible* images, preserving global structure while sharpening fine details and suppressing minor artifacts. When applied to SD3-Medium, EMAG's outputs align more closely with human-preference proxies (HPS [34]).

Motivated by this, we propose Exponential Moving Average Guidance (EMAG), which, unlike other approaches, by design can regulate how much *high-frequency* information (fine-grained details) the weaker model is allowed to destroy by controlling the EMA updates, deliberately injecting subtle, semantically faithful artifacts. These yield *hard negatives* that reveal failure modes that base models are more likely to miss, enabling more effective rectification during sampling, see Fig. 1. Empirically, EMAG improves HPS by +0.54 (Table 2) over CFG and yields additional gains when combined with stronger orthogonal guidance baselines such as APG and CADS across most experimental settings, underscoring the value of controllable, fine-grained negatives in high-quality regimes. We also evaluate EMAG as a standalone guidance method against other baselines (see Supplementary Table S17[1]).

Contributions

[1] Figures and tables prefixed with "S" refer to the supplementary material.



(a) Mean similarity scores between positive and negative samples across different baselines and EMAG. Higher similarity indicates harder negative.

(b) Mean similarity scores across different β values for EMAG. Higher similarity indicates harder negative.

Fig. 2: Negative sample hardness quantified via cosine similarity (CLIP-I and DINOv2). (a) EMAG produces harder negatives than prior guidance methods across both similarity metrics. **(b)** Within EMAG, the decay factor β directly controls negative hardness.

- We propose **EMAG**, a training-free attention-map EMA guidance that produces *controllable* weak signals without extra training. We also study the impact of negative sample hardness on guidance effectiveness, use cosine-similarity to quantify it, and show that EMAG's EMA β controls it.
- We propose a statistics-based *adaptive layer selection* that chooses target layers per timestep to create controllable degradations in negative samples.
- We conduct extensive experiments on DiT [25] and MMDiT (SD3 [7]) backbones, covering both class-conditional and text-to-image generation, and report competitive performance against strong guidance baselines, including notable improvement in Human Preference Score (HPS) [34].
- We demonstrate the complementary nature of our approach with advanced orthogonal guidance methods such as APG [28] and CADS [27], further improving HPS.

2 Related work

Classifier-free guidance (CFG) [12] combines conditional and unconditional predictions to improve prompt alignment and perceptual quality, but suffers from reduced generation diversity [2,15,27] and visual over-saturation [12]. To alleviate these issues, recent studies have revisited CFG and developed strategies that better balance semantic alignment and visual fidelity. Adaptive Projected Guidance (APG) [28] decomposes the CFG update into orthogonal/parallel components and down-weights saturation-prone components, enabling higher guidance scales with fewer artifacts. Condition-Annealed Diffusion Sampler (CADS) [27] boosts diversity via annealing the conditioning strength during inference, especially at high guidance scales.

A useful unifying view is *auto-guidance* [16], where a strong model is contrasted with a deliberately weakened (reduced parameters or training steps) variant to steer sampling. This separation helps stabilize updates and expose failure modes, highlighting the artifacts typical of high guidance scales. *Karras*

et al. report that auto-guidance disentangles control over image quality while preserving variation, yielding more favorable quality–diversity trade-offs than CFG [16].

Fig. 3: CFG(left) vs EMAG hard negatives(right) EMAG samples at varying β; larger β produces stronger degradations.

However, explicitly maintaining strong/weak model pairs is not always practical; several strategies have been introduced to achieve a similar effect with a single model. Self-Attention Guidance (SAG) degrades the *input* by blurring regions indicated by self-attention, yielding a weaker score that guides the original model [14]. Smoothed Energy Guidance (SEG) weakens the prediction by smoothing the attention energy (Gaussian blurring of attention weights), with an efficient *query blurring* equivalent; in practice, one can approximate this by *replacing image queries with their channel-wise means* [13]. Entropy Rectifying Guidance (ERG) perturbs the attention mechanism (including conditioning pathways) at inference to improve quality, diversity, and prompt consistency [15]. Perturbed-Attention Guidance (PAG) forms weak intermediates by replacing selected self-attention maps with the identity and then guiding away from them [1]. S^2-Guidance (Stochastic Self-Guidance) constructs weak sub-networks by stochastically dropping layers [4]. While guiding with weaker or negative guidance signals can outperform CFG, most methods offer limited control over the *quality* (hardness and granularity) of these negatives. We introduce *Exponential Moving Average Guidance (EMAG)*, a training-free approach that generates *controllable, semantically faithful hard negatives* by regulating fine-grained (high-frequency) degradations through EMA-based attention smoothing, where the decay factor β directly controls the degradation strength, together with adaptive layer selection. This design exposes hard failure modes that coarse degradations often miss and complements advanced guidance methods such as APG and CADS.

3 Method

3.1 Diffusion preliminary

Diffusion preliminaries. In diffusion-based generation [6, 7, 19], a clean image x_0 is progressively corrupted with Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ over timesteps $t \in [0, T]$, yielding noisy samples x_t and this is known as the forward process. This can be represented as a stochastic differential equation (SDE)

$$dx \;=\; f(x,t)\,dt \;+\; g(t)\,dw, \tag{1}$$

where $f(x, t)$ and $g(t)$ control the corruption process and w is a standard wiener process. To recover x_0 from x_t, one integrates the reverse-time SDE [31]

$$dx \;=\; \big[f(x,t) \;-\; g(t)^2 \nabla_x \log p_t(x)\big]\,dt \;+\; g(t)\,d\bar{w}, \tag{2}$$

where $\nabla_x \log p_t(x)$ is the score of the perturbed distribution and $\bar{w}$ is a standard Wiener process in reverse time. A neural network (e.g., a Transformer) is trained to approximate the score, $e_\theta(x,t) \approx \nabla_x \log p_t(x)$, enabling generation from random Gaussian noise. With external conditioning c (e.g., class labels or text), the conditional score is learned, $e_\theta(x,t,c) \approx \nabla_x \log p_t(x \mid c)$ [6].

Flow matching Flow matching trains a *deterministic* neural ODE to transport a simple prior to the data by regressing a time-dependent velocity field:

$$\dot{x} \;=\; v_\theta(x,t), \tag{3}$$

with the standard regression objective

$$\min_\theta \; \mathbb{E}_{t\sim\mathcal{U}(0,1),\,(x_0,x_1)} \; \big\| \, v_\theta\big((1-t)x_0 + t x_1,\; t\big) \;-\; (x_1 - x_0) \big\|_2^2, \tag{4}$$

where using the linear path yields the popular *rectified flow* instance. In contrast, diffusion employs a *stochastic* reverse SDE or its deterministic probability-flow ODE; the latter evolves the same marginals as the reverse SDE:

$$\frac{dx}{dt} \;=\; f(x,t) \;-\; \tfrac{1}{2}\, g(t)^2\, \nabla_x \log p_t(x). \tag{5}$$

3.2 Existing guidance approaches

Conditioned Diffusion models have recently achieved the capability to generate high-quality multimedia. Classifier Free Guidance (CFG) [12] is one of the major breakthroughs behind this dramatic improvement in the generation quality, and prompt alignment, where the unconditional and conditional score estimates are combined to create the final estimate with the help of guidance scale as depicted in Eq. 6 where ϵ_θ is the denoiser neural network, x_t is the input latent, c is the conditioning input and $\varnothing$ is null input. Hence $\epsilon_\theta(x_t, \varnothing)$ and $\epsilon_\theta(x_t, c)$ are the denoiser network with null-conditioning (unconditional) and conditioning and guidance scale $w > 1$.

Table 1: Comparison of SOTA CFG-complementary guidance methods with and without EMAG on SD3 (COCO 2014, identical setup). **CFG+EMAG ≡ EMAG by design*

Guidance	**HPS ↑**	**HPS ↑**
		+EMAG
CFG	29.22	**29.76***
ERG	29.35	**29.56**
S2	29.15	**29.51**
SAG + CFG	29.28	-
SEG + CFG	29.25	**29.29**
EMAG	**29.76**	-

$$\tilde{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, \varnothing) + w \cdot (\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \varnothing)) \tag{6}$$

However, CFG suffers from inherent drawbacks such as reduced generation diversity and visual over-saturation . To alleviate these issues, recent studies have revisited CFG and developed strategies that better balance semantic alignment and visual fidelity. One such direction is Auto-guidance , which employs two

versions of the denoiser that share the same network structure but differ in capacity. Specifically as shown in Eq. 7,

$$\tilde{\epsilon}(x_t, c) = \epsilon_0(x_t, c) + w \cdot (\epsilon_1(x_t, c) - \epsilon_0(x_t, c)) \tag{7}$$

where the weaker model is denoted as ϵ_0 and the stronger model as ϵ_1. Here the weaker model's score prediction guides the stronger one in Eq. 7.

This Auto-guidance framework in Eq. 7 can be achieved by using different ways to construct the weaker model through controlled degradation. For example, SEG forms the weaker model by directly perturbing the denoiser's attention weights, such as by adding Gaussian noise or replacing image queries with their channel-wise means. Let the perturbed denoiser be ϵ_1'. With this weaker version defined, replacing $\epsilon_0(x_t, c)$ with $\epsilon_1'(x_t, \varnothing)$ in Eq. 7 produces the SEG guidance formulation (only for unconditional). More recent approaches, such as ERG, extend this perturbation to both the network and the conditioning encoder, resulting in a perturbed denoiser ϵ_1' and perturbed conditioning c'. Substituting these into Eq. 7 leads to the ERG objective, i.e., replacing $\epsilon_0(x_t, c)$ with $\epsilon_1'(x_t, c')$.

$$\tilde{\epsilon}(x_t) = \epsilon_1(x_t) + w \cdot (\epsilon_1(x_t) - \epsilon_1'(x_t)) \tag{8}$$

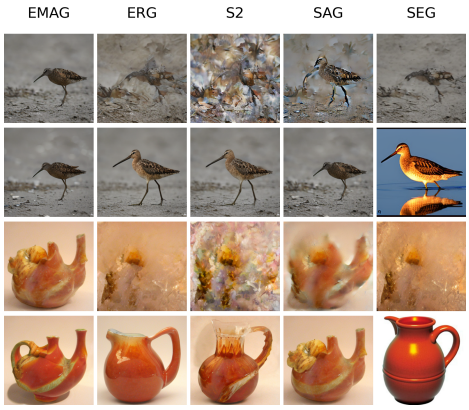


Fig. 4: Negative-sample comparison (top: negative samples; bottom: positive samples (clean image). Prior guidance method (SAG [14], auto guidance [16], ERG [15], S^2-Guidance [4]) often yield obvious "easy" degradations. In contrast, **EMAG** produces subtle, semantically near-miss negatives that reveal difficult failure modes yet retain global structure, enabling finer refinement. (Samples from DiT)

PAG perturbs self-attention by replacing selected attention maps with the identity, and then guides sampling away from the resulting degraded intermediates; its guidance rule departs from Eq. 7 and is given in Eq. 8, where ϵ_1' denotes the network with perturbed attention. Similarly, SAG uses self-attention maps from the denoiser ϵ_θ (the score prediction network) to detect high-frequency regions in the latent input. These regions are then selectively blurred to obtain a degraded latent x_t', effectively producing a weaker model $\epsilon_1(x_t')$. Substituting this degraded score into Eq. 8 in place of perturbed noise $\epsilon_1'(x_t)$ yields the SAG objective. Alternatively, *S^2-Guidance* constructs the weaker surrogate via stochastic layer dropping and, instead of Eq. 7, improves the CFG update by subtracting the perturbed model's noise prediction scaled by s; equivalently, Eq. 6 is augmented with the term $-s \cdot \epsilon_1'(x_t, c, m)$, where m controls which layers are dropped at each timestep, ϵ_1' is the perturbed sub-network, and s is the S^2 scale.

In the auto-guidance framework, a weaker model helps steer sampling toward higher-quality regions [16]. When derived from the same backbone, the weaker model tends to replicate the strong model's errors but with greater magnitude. This discrepancy highlights failure modes that the strong model can correct. [16]. Following this intuition, we target subtle, fine-grained errors (*hard negatives*) by constructing lightly degraded surrogates that act as guidance, improving fidelity and HPS [34]. Our approach follows the attention-perturbation family, e.g., ERG, PAG, and SEG, which instantiate such surrogates by modifying self-attention during sampling.

3.3 Hard Negatives

Why quality of negative samples matters? In prior approaches, including auto-guidance [16], Entropy Rectifying Guidance (ERG) [15], and S^2-Guidance [4] instantiate a weaker surrogate to surface failure modes complementary to a strong model, thereby guiding sampling toward higher quality. However, these methods primarily specify how to obtain a weaker surrogate (e.g., capacity reduction, attention perturbation, or stochastic layer dropping) rather than offering direct, fine-grained control over which failure modes to elicit and to what extent. In this work, we introduce a strategy that explicitly controls the degradation level while preserving core semantics, yielding "hard negatives", subtle, fine-grained degradations that expose nuanced errors without altering global structure and enable the model to rectify them see Fig. 4. In Fig. 3, we observe that while the global structure of the mobile phone remains intact, subtle local deformations (e.g., distorted buttons) emerge; such minimally degraded yet semantically consistent samples constitute the hard negatives considered in this work. As evidenced in Fig. 1 and Table 1, this controllable hard-negative guidance improves the Human Preference Score (HPS) [34] over both the baseline and several strong methods. We further quantify negative hardness in Fig. 2a, which reports the similarity between negative samples and their corresponding generated outputs; higher similarity indicates harder negatives.

Measuring negative "hardness" via cosine similarity. We quantify how "hard" a negative is by measuring how *close* it is to the final guided output under a semantic image representation. Let x^+ denote the final guided image (or decoded final sample), and let x^- denote the corresponding negative image produced by the EMA-perturbed branch (using the same condition/seed). We compute the hardness score as cosine similarity in a frozen embedding space:

$$\mathcal{H}(x^-, x^+) \;=\; \cos\big(\phi(x^-), \phi(x^+)\big) \;=\; \frac{\langle \phi(x^-), \phi(x^+) \rangle}{\|\phi(x^-)\|_2 \, \|\phi(x^+)\|_2}, \tag{9}$$

where $\phi(\cdot)$ denotes a pretrained image encoder. We compute cosine similarity using both CLIP image embeddings, which capture semantic similarity, and DINOv2 embeddings, which better reflect structural and perceptual similarity.

Algorithm 1 EMAG (Conditional)

Require: ϵ_θ: denoiser, ϵ'_θ: perturbed denoiser, τ_s, τ_e: start/end timesteps, w_e: EMAG scale, c: condition, w_{cfg}: CFG scale

$x_T \sim \mathcal{N}(0, I)$
for $t = T, T-1, \ldots, 1$ **do**
  $z_t^c \leftarrow \epsilon_\theta(x_t, c)$
  $z_t \leftarrow \epsilon_\theta(x_t)$
  if $\tau_e < t < \tau_s$ **then**
    $\hat{z}_t^c \leftarrow \epsilon'_\theta(x_t, c)$ {Eq. (11), (12), (13)}
    $\bar{z}_t^c \leftarrow \hat{z}_t^c + w_e \cdot (z_t^c - \hat{z}_t^c)$ {Eq. (14)}
    $\bar{z}_t \leftarrow z_t + w_{cfg} \cdot (\bar{z}_t^c - z_t)$ {Eq. (15)}
  else
    $\bar{z}_t \leftarrow z_t + w_{cfg} \cdot (z_t^c - z_t)$ {Eq. (6)}
  end if
end for
return $\bar{z}_t$

Intuitively, larger values of $\mathcal{H}$ indicate *harder* negatives: the negative sample remains semantically close to the final output while differing primarily in fine-grained details. We compare mean similarity scores across baselines in Fig. 2a, where EMAG consistently produces the hardest negatives under both metrics, with SAG being the closest competitor. Within EMAG, smaller values of β yield harder negatives (Fig. 2b), as evidenced by higher similarity scores, particularly under DINOv2 [24], which is sensitive to structural and perceptual similarity. Qualitative examples are provided in Fig. S10

3.4 Exponential Moving Average Guidance

We adopt a strategy akin to prior works [13, 14, 16] where we contrast denoising outputs from a strong model versus a deliberately degraded weaker prediction via Eq. 7. Following [1, 13, 15], we perturb self-attention by replacing the attention map at a selected layer and timestep t, A_t, with its exponential moving average $\mathrm{EMA}(A_{<t})$, using our adaptive layer-selection strategy (Sec 3.4). The detailed steps are provided in Alg. 1 and Alg. 2 (supplementary). Intuitively, since self-attention refines semantics from coarse to fine, swapping in the EMA suppresses high-frequency refinements while preserving global structure, yielding a controllable, semantically faithful "hard negative." We then contrast this negative with the original prediction during sampling. Fig. 4 shows that our EMA-guided approach degrades selectively without losing core semantic details, and Fig. 3 and 2b demonstrate how adjusting the decay factor β controls the degradation strength. Now we will discuss our approach and its components in detail.

Attention EMA calculation Let $A_t \in [0,1]^{B \times H \times Q \times K}$ denote the softmax-normalized attention at timestep t (rows sum to 1 over K). We maintain an

(a) Unconditional generation. **Top:** EMAG. **Bottom:** unguided (baseline).

(b) Class-conditional generation. **Top:** EMAG. **Bottom:** CFG (baseline).

Fig. 5: Qualitative comparison for (a) Unconditional and (b) Text-conditional settings.

exponential moving average (EMA) E_t of the same shape as A_t, initialized on the first call as $E_1 := A_1$, and updated as mentioned in Eq.10.

$$E_t \;=\; \beta\, E_{t-1} \;+\; (1-\beta)\, A_t, \qquad \beta = 0.988. \tag{10}$$

We set β by its EMA half-life H via $\beta = e^{-\ln(2)/H}$: guidance becomes net-positive once the half-life exceeds $\approx 0.7T$ of the sampling horizon, and our default β=0.988 (half-life $\approx 2T$ for 28 steps) lies in this saturated regime. The full β ablation is provided in Supp. Tab. S21. We also introduce parameters δ_t, τ_s, τ_e where δ_t indicates the warmup period for the EMA to stabilize before we can start swapping the attention maps with the EMA version, τ_s indicates the timesteps when we start accumulating EMA and apply the perturbation, and τ_e indicates the timestep when we don't perturb the attention. Since the denoising happens in reverse, we have $t_0 < \tau_e < \tau_s < t_{max}$.

Replacing the attention map with its EMA effectively suppresses the incremental (high-frequency) refinements the model applies over timesteps while preserving the global structure. This induces inferior but semantically faithful predictions that guide sampling toward higher quality by introducing precise failure modes. In practice, the exponential smoothing prevents drift: even with $\delta_t = 1$, the guidance remains faithful to the original structure, unlike strong perturbation approaches [4, 15] (see Fig. 4).

Layer Selection As prior work shows [1, 14, 15], middle layers are especially effective for attention perturbations. For instance, [15] demonstrates that middle layers exhibit lower maximum attention probabilities than first or last layers. Methods like [1, 14] also extract masks or apply perturbations in these middle layers of the U-Net. Inspired by this, we apply EMAG specifically to middle layers: for DiT-XL/2 [25] we set $(l_{\min}, l_{\max}) = (12, 15)$, and for SD3-Medium [7] we set $(l_{\min}, l_{\max}) = (6, 8)$.

At each timestep t and layers (l_{min}, l_{max}) we select a layer l_n where $n \in (l_{min}, l_{max})$ using the Eq.11 and Eq. 12 where l_n^t is the selected layer at time

step t and $|A_t|$ is the total number of elements in the attention map. For per-layer selection, we compute $\Delta_t^{(\ell)}$ on the attention tensor of layer ℓ and choose layers with the largest $\Delta_t^{(\ell)}$.

$$l_n^t = \arg \max_{n \in \{l_{\min}, \ldots, l_{\max}\}} \Delta_t^{(n)} \tag{11}$$

$$\Delta_t \;=\; \mathrm{MAE}(E_t, A_t) = \frac{1}{|A_t|} \sum_i \left| E_t^{(i)} - A_t^{(i)} \right| \tag{12}$$

Finally, we replace the EMA for the attention layer using a convex blend, see Eq.13 with $\lambda = 1$ (hard replace) as our default setting. $\lambda \in [0, 1]$ allows for a soft update, and $\widetilde{A}_t$ is the perturbed attention map.

$$\widetilde{A}_t \;=\; (1 - \lambda)\, A_t \;+\; \lambda\, E_t, \tag{13}$$

This enables us to select the layer with the largest Δ_t at each timestep t, ensuring maximal perturbation. We maintain EMA buffers for all layers in $(l_{\min}, l_{\max})$ and at every timestep replace the attention map of the selected layer l_n^t. Since different layers contribute unequally across timesteps [20], our statistic-guided selection picks the layer actively inserting high-frequency content, yielding stronger negative samples. This approach also opens the possibility for providing more controlled negative samples by carefully designing optimization for the $\Delta_t^{(\ell)}$ as against the greedy approach Eq. 11. We defer this to future work.

EMAG-I In MMDiT's joint attention over text and image tokens, our default EMAG perturbs both Image→Image and Image→Text sub-blocks. While EMAG-I perturbs only the Image→Image block. See appendix Section E

EMAG-Q Instead of maintaining EMA over post-softmax attention maps, EMAG-Q tracks EMA over query embeddings and perturbs only queries, enabling fused SDPA and substantially reducing overhead. We defer the formulation and approximation analysis to the appendix section G

3.5 Guidance update

Once we obtain the perturbed noise estimates, we perform the guidance update in two steps first, we leverage Eq. 7 for the guidance update and rewrite it as Eq.14 presenting the first step of Guidance update for EMAG, where $\epsilon'_\theta(x_t, c)$ is the degraded score estimate and w_e is the EMAG scale. Now using this perturbed signal, we perform a CFG-style update but with our perturbed conditional noise, see eq.15 where w is CFG scale.

$$\tilde{\epsilon}(x_t, c)' = \epsilon'_\theta(x_t, c) + w_e \cdot (\epsilon_\theta(x_t, c) - \epsilon'_\theta(x_t, c)) \tag{14}$$

$$\tilde{\epsilon}(x_t, c) = \epsilon_\theta(x_t, \varnothing) + w \cdot (\tilde{\epsilon}(x_t, c)' - \epsilon_\theta(x_t, \varnothing)) \tag{15}$$

We can also apply EMAG for unconditional generation as presented in Eq. 16.

$$\tilde{\epsilon}(x_t) = \epsilon'_\theta(x_t, \varnothing) + w_e \cdot (\epsilon_\theta(x_t) - \epsilon'_\theta(x_t, \varnothing)) \tag{16}$$

Combining EMAG with APG and CADS EMAG exhibits complementary behavior when combined with more advanced guidance techniques like APG [28], and CADS [27], where for the APG we follow the algorithm provided by the author, specifically the eq:14 remains the same while APG is applied to the Eq: 15. In essence, it's similar to applying APG to CFG, but in this case, we use our perturbed conditioning.

For CADS, we simply apply the interpolation between the positive tokens and the Gaussian noise to the encoder hidden states of the network. We apply this interpolation to the conditioning of the perturbed network as well.

4 Experimentation and results

4.1 Experimental setup

Models We study Transformer-based diffusion backbones: (i) **DiT-XL/2** in the *class-conditional* setting at **256×256** and **512×512** resolutions [25], and (ii) **MMDiT** via the **Stable Diffusion 3 Medium** checkpoint for **1024×1024** *text-to-image* generation [7]. See additional details in Supplementary Sec. A.

Datasets & tasks We evaluate both *conditional* and *unconditional* generation. For *class-conditional* experiments with DiT-XL/2 we use ImageNet-1K labels [5] and evaluate against the *official ADM reference batches/statistics* from the Guided Diffusion repository [5, 6]. For *text-conditioned* experiments with SD3 we use captions from the **COCO-2014 40k** validation split [18]. Unless stated otherwise, we generate **50K** samples for class-conditional ImageNet experiments and **40K** samples for text-to-image experiments (one per caption from the COCO-2014 validation split) with the seed set to 8.

Table 2: Quantitative SD3 text-to-image results on the COCO 2014 validation set (identical data, sampling, and evaluation). Settings are selected from Pareto optimal scales (Supp Sec. D). Bottom block highlights EMAG combos that surpass the corresponding baseline. For ± std over 3 seeds; see Table S22. † Incompatible with SD3, see supplementary section (E.2)

Guidance	FID ↓	HPS ↑
(No CFG)	23.859	21.61 ± 0.11
CFG	22.877	29.22 ± 0.05
SAG + CFG	23.039	29.28 ± 0.04
PAG + CFG †	24.066	28.73 ± 0.04
SEG + CFG	23.453	29.25 ± 0.04
APG	21.816	29.45 ± 0.02
CADS	**18.320**	28.36
S^2	20.821	29.15 ± 0.00
ERG	23.989	29.35 ± 0.03
EMAG-I (ours)	22.154	29.56
EMAG (ours)	22.890	**29.76** ± 0.03
EMAG-Q (ours)	23.530	29.64
EMAG + APG	21.819	**29.79** ± 0.04
EMAG + CADS	19.950	**28.86** ± 0.02
EMAG-Q + APG	22.480	29.77
EMAG-Q + CADS	20.030	28.96

Baselines and compared guidance methods We compare against *standard* conditional generation (no guidance), *classifier-free guidance* (CFG), and recent *attention-based* guidance methods: **PAG** [1], **SAG** [14], **SEG** [13], **ERG** [15], and stochastic block dropping method $\mathbf{S}^2$ (Stochastic Self-Guidance) [4]. We

Table 3: Quantitative results for unconditional generation on the COCO,2014 validation set (SD3) and ImageNet-256 (DiT-XL/2). All methods share identical data, sampling steps, and evaluation code; settings follow author recommendations or are selected via hyperparameter sweep (Supp Sec. D). **Baseline methods are reported where compatible with the corresponding backbone and sampling formulation.**† Incompatible with SD3, see supplementary section (E.2)

Guidance	FID ↓	Precision ↑	Recall ↑	Density ↑	Coverage ↑
SD3 (MMDiT) [7]					
(No-Guidance)	101.94	0.283	**0.298**	0.199	0.095
SAG	95.38	0.339	0.098	0.248	0.046
PAG †	111.24	0.339	0.000	0.238	0.024
SEG	90.21	**0.609**	0.181	**0.674**	0.090
I-ERG	<u>82.11</u>	0.557	0.173	0.608	<u>0.090</u>
EMAG (ours)	**74.98**	<u>0.562</u>	<u>0.186</u>	<u>0.642</u>	**0.112**
DiT-XL/2 [25] 256x256					
(No-Guidance)	52.76	0.431	<u>0.664</u>	0.392	0.554
PAG	39.51	<u>0.626</u>	0.357	<u>0.811</u>	<u>0.703</u>
SAG	40.87	0.513	**0.655**	0.521	0.672
SEG	<u>39.17</u>	**0.638**	0.340	**0.844**	0.701
EMAG (ours)	**29.85**	0.621	0.561	0.805	**0.768**

also evaluate two orthogonal guidance strategies, **APG** (Adaptive Projected Guidance) [28] and **CADS** (Condition-Annealed Diffusion Sampling) [27], and further demonstrate that *combining* APG/CADS with our approach yields additional gains (see Table 2). For methods originally proposed for U-Net backbones, we adapt them to Transformer backbones (DiT/MMDiT) by operating within the self-attention stack (e.g., perturbing queries or attention maps, or inserting attention-space degradations) while preserving the residual/MLP pathways; implementations are based on the strategies adopted in [15]. Specifics are provided in the Supplementary Sec. E.

Metrics We assess *quality* with **HPS v2** [34], **FID** [10] and **Precision**/**Density** [22]; and *diversity* with **Recall**/**Coverage** [22]. We scale HPS by 100 when reporting. For text–image alignment, we report **CLIPScore** [9]. In text-to-image (conditional) results, we primarily report **HPS** and **FID**. For brevity, we refer to HPS v2 as HPS throughout the paper.

Hyperparameter sweeps For each baseline (CFG, SAG, PAG, SEG, ERG, S^2, APG, CADS), we perform a grid search over combinations of the CFG scale and the method-specific control scale(s). For SD3, we construct the FID–HPS Pareto frontier for each method. We report configurations chosen from the Pareto-optimal scales of each method's FID–HPS frontier. The full sweep results are

Table 4: Wall-clock inference timing on a single NVIDIA A100-SXM4-40GB. All methods use SD3 Medium at 1024×1024 with 28 steps, batch size 1. Timings are averaged over 45 images after 5 warmup iterations. Overhead is relative to vanilla CFG. FID / HPS are on the COCO 2014 validation set (same setting as Table 2).

Method	s/image	± std	Overhead	Peak Mem.	FID/HPS
CFG (baseline)	4.04	0.001	–	16.89 GB	22.88 / 29.22
SAG [14] + CFG	7.22	0.005	+78.5%	24.85 GB	23.04 / 29.28
SEG [13] + CFG	6.05	0.001	+49.6%	18.76 GB	23.45 / 29.25
PAG [1] + CFG	9.41	0.002	+132.7%	18.33 GB	24.07 / 28.73
EMAG(Ours)	12.66	0.002	+213.2%	29.58 GB	22.89 / **29.76**
EMAG-Q(Ours)	7.86	0.005	+94.5%	16.96 GB	23.53 / 29.64

reported in Table S11, and the corresponding Pareto curves in Figs. 6 and S8. We apply the same selection protocol to DiT, with details provided in Supplementary Sec. D.

4.2 Results

Conditional generation We compare EMAG against conditional baselines, including both unguided and CFG-based sampling, as well as recent training-free guidance methods for text-to-image generation, including S^2-Guidance and ERG. Quantitative results in Table 2 show that EMAG achieves the highest HPS while remaining competitive on FID. We further evaluate compatibility with advanced guidance strategies such as APG and CADS, and find that it consistently yields additional gains in HPS without sacrificing their benefits on FID. Additional PRDC results in Supplementary Table S1.

Table 5: Ablation for adaptive layer selection (5K samples, SD3, COCO 2014 val). All runs share identical sampling steps and evaluation code.

Guidance	FID ↓	HPS ↑
EMAG (L 6)	**27.47**	29.54
EMAG (L 7)	28.94	29.51
EMAG (L 8)	28.23	**29.66**
EMAG (L all)	35.60	28.80
EMAG (Adaptive)	28.52	29.60

We also compare EMAG with guidance methods that are complementary to CFG and designed to boost its performance (Table 1). EMAG delivers the largest HPS gain over CFG alone, attributable to EMA-induced hard negatives that encourage subtle refinements. It also serves as a complementary plug-in to strong guidance methods such as S^2 and ERG, further improving HPS. Qualitative examples are presented in Fig. 5 and Supplementary Sec. F. Tables S8 and S9 present the results for class-conditioned DiT experiments.

We additionally compare standalone EMAG against other baselines in Table S17 and evaluate its behaviour under reduced inference budgets in Table S20.

We further note that, across methods, negative hardness alone does not determine final quality, and the perturbation mechanism also matters We further note

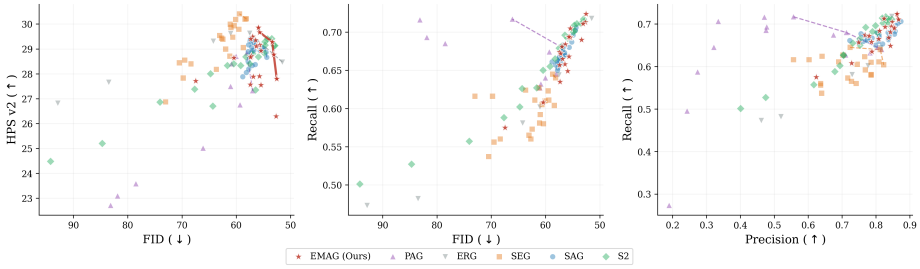


Fig. 6: Pairwise Pareto frontiers across guidance methods on MS-COCO 2014 (1K samples, SD3). Each point represents a specific guidance configuration; axes are oriented so that the *top-right corner is optimal* in all three panels. **(Left)** FID vs. HPS: the primary quality–preference trade-off. EMAG (★) consistently occupies the top-right region, achieving a high HPS -FID tradeoff. **(Middle)** FID vs. Recall: fidelity–diversity trade-off. EMAG maintains competitive recall across the full FID range. **(Right)** Precision vs. Recall: quality–diversity trade-off. EMAG achieves a balanced operating point, avoiding the precision–recall collapse. Both method-specific and CFG scales are swept; full numerical results in Tab. S11.

that, across methods, negative hardness alone does not determine final quality—the perturbation mechanism also matters (e.g., S^2 and ERG show comparable hardness in Fig. 2a yet differ on the Pareto frontier in Fig. 6).

Unconditional generation We evaluate EMAG against a no-guidance baseline and representative guidance methods applicable in this setting, such as SAG, SEG (adapted to SD3), and I-ERG, and additionally report results for DiT-XL/2. Quantitative results in Table 3 show that EMAG attains the lowest FID and the highest Coverage across both settings, while remaining competitive on the remaining metrics. Qualitative comparisons in Fig. 5(a) show that, relative to the no-guidance baseline, EMAG preserves the core structure while refining fine details through EMA-induced hard negatives.

A practical limitation of standard EMAG is its computational overhead: maintaining EMA over post-softmax attention maps prevents the use of fused SDPA kernels and increases both runtime and memory (Table 4). To address this, we introduce **EMAG-Q**. It reduces inference overhead from +213% to +95% and peak memory from 29.58 GB to 16.96 GB, while preserving generation quality with only a +0.63 increase in FID and a −0.12 drop in HPS relative to EMAG (Table S16). Further discussion is provided in Appendix Sec. G.

4.3 Ablation results

Adaptive layer selection Table 5 ablates the adaptive layer-selection mechanism against fixed single-layer and all-layers alternatives. Fixed single-layer variants may improve one metric but not both, while the all-layers strategy increases overhead and degrades performance. Our adaptive selection achieves the best FID–HPS balance.

Guidance scale selection Fig. S3 shows how FID and HPS vary with CFG and EMAG scales. Unless noted otherwise (e.g., in ablations), we use a CFG scale of **7** and an EMAG scale of **1.5** for conditional and **5.125** for unconditional. A detailed Pareto frontier analysis is provided in Supplementary Section D.5.

5 Conclusion

We introduce *Exponential Moving Average Guidance (EMAG)*, a training-free method that uses temporal attention smoothing to construct semantically faithful hard negatives for fine-grained refinement during sampling. EMAG improves HPS across conditional settings while remaining competitive on fidelity and diversity. It composes effectively with orthogonal techniques such as APG and CADS, as well as negative-signal guidance methods such as S^2.

References

1. Ahn, D., Cho, H., Min, J., Jang, W., Kim, J., Kim, S., Park, H.H., Jin, K.H., Kim, S.: Self-rectifying diffusion sampling with perturbed-attention guidance. In: European Conference on Computer Vision. pp. 1–17. Springer (2024)
2. Astolfi, P., Careil, M., Hall, M., Mañas, O., Muckley, M., Verbeek, J., Soriano, A.R., Drozdzal, M.: Consistency-diversity-realism pareto fronts of conditional image generative models. arXiv preprint arXiv:2406.10429 (2024)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Chen, C., Zhu, J., Feng, X., Huang, N., Wu, M., Mao, F., Wu, J., Chu, X., Li, X.: s^2-guidance: Stochastic self guidance for training-free enhancement of diffusion models. arXiv preprint arXiv:2508.12880 (2025)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
9. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

12. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
13. Hong, S.: Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. Advances in Neural Information Processing Systems **37**, 66743–66772 (2024)
14. Hong, S., Lee, G., Jang, W., Kim, S.: Improving sample quality of diffusion models using self-attention guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7462–7471 (2023)
15. Ifriqi, T.B., Romero-Soriano, A., Drozdzal, M., Verbeek, J., Alahari, K.: Entropy rectifying guidance for diffusion and flow models. arXiv preprint arXiv:2504.13987 (2025)
16. Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S.: Guiding a diffusion model with a bad version of itself. Advances in Neural Information Processing Systems **37**, 52996–53021 (2024)
17. Kingma, D., Gao, R.: Understanding diffusion objectives as the elbo with simple data augmentation. Advances in Neural Information Processing Systems **36**, 65484–65516 (2023)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
19. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
20. Ma, X., Fang, G., Bi Mi, M., Wang, X.: Learning-to-cache: Accelerating diffusion transformer via layer caching. Advances in Neural Information Processing Systems **37**, 133282–133304 (2024)
21. Mukhopadhyay, S., Gwilliam, M., Agarwal, V., Padmanabhan, N., Swaminathan, A., Hegde, S., Zhou, T., Shrivastava, A.: Diffusion models beat gans on image classification. arXiv preprint arXiv:2307.08702 (2023)
22. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International conference on machine learning. pp. 7176–7185. PMLR (2020)
23. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
27. Sadat, S., Buhmann, J., Bradley, D., Hilliges, O., Weber, R.M.: Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. arXiv preprint arXiv:2310.17347 (2023)
28. Sadat, S., Hilliges, O., Weber, R.M.: Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In: The Thirteenth International Conference on Learning Representations (2024)
29. Shen, D., Song, G., Xue, Z., Wang, F.Y., Liu, Y.: Rethinking the spatial inconsistency in classifier-free diffusion guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9370–9379 (2024)

30. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
31. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
32. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
33. Wang, Y., Ju, Z., Tan, X., He, L., Wu, Z., Bian, J., et al.: Audit: Audio editing by following instructions with latent diffusion models. Advances in Neural Information Processing Systems **36**, 71340–71357 (2023)
34. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)