PERTURBING TEXT ATTENTION FOR GUIDING DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free guidance methods steer the sampling process of a diffusion model by extrapolating the prediction of the model away from a weak prediction. The weak prediction is obtained by perturbing intermediate computations of the model at inference time, such as self-attention weights or the arrangement of visual tokens. However, their performance gains over unguided sampling remain marginal for text-conditional generation. To understand why, we present an in-depth analysis of how diffusion models behave under conditions of varying complexity. Our analysis reveals that the primary bottleneck for text-conditional generation lies in aligning generated samples with the given condition, rather than in refining spatial details among visual tokens, which existing methods mainly focus on. Based on this, we introduce text attention guidance (TAG), a simple yet effective training-free guidance method that perturbs the interaction between visual and textual tokens. To this end, we propose to degrade the attention weights between the two modalities, yielding a weak prediction that particularly struggles to align samples with the given condition. This provides a targeted negative signal that mitigates the core bottleneck of text-conditional diffusion models. Extensive experiments on both U-Net and transformer-based models, including SDXL, SD3.5, Flux.1-dev, and Qwen-Image, demonstrate the effectiveness of our approach, which outperforms existing training-free guidance methods significantly and achieves results comparable to or even better than those of classifier-free guidance without any training. We will release the code upon acceptance.

1 INTRODUCTION

Diffusion models[1] (Ho et al., 2020; Song & Ermon, 2019; Liu et al., 2023; Lipman et al., 2023) demonstrate strong performance across various generation tasks, including practical applications conditioned on text prompts, such as text-to-image (Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2024; Esser et al., 2024) and video generation (Ho et al., 2022; Blattmann et al., 2023). A key to their success lies in guidance techniques that steer the sampling process toward high-quality regions of the data distribution. Classifier-free guidance (CFG) (Ho & Salimans, 2022) is a representative example that extrapolates the conditional prediction away from the unconditional one, improving both visual quality and condition alignment.

Alternatively, several works (Karras et al., 2024a; Ahn et al., 2024; Zhang et al., 2026) leverage an intentionally weakened model to guide the sampling process instead of relying on an unconditional prediction. The core idea is that a pair of predictions from two models, sharing the same failure modes but to different degrees, can be used to cancel out shared errors and suppress the deficiencies of the main model. For instance, autoguidance (Karras et al., 2024a) uses a smaller model trained for fewer iterations to capture inherent failure modes, leveraging its predictions as a negative signal to refine the final outputs. While effective, this approach requires specific training procedures to obtain the weak model. To avoid such costs, training-free approaches manipulate intermediate computations in the forward pass of a diffusion model, such as self-attention weights (Ahn et al., 2024; Hong, 2024) or the arrangement of visual tokens (Rajabi et al., 2025; Zhang et al., 2026), to simulate a weak prediction. Although these approaches provide training-free alternatives, their performance

[1]For brevity, we refer to both diffusion and flow-based models as diffusion models throughout this paper.

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.

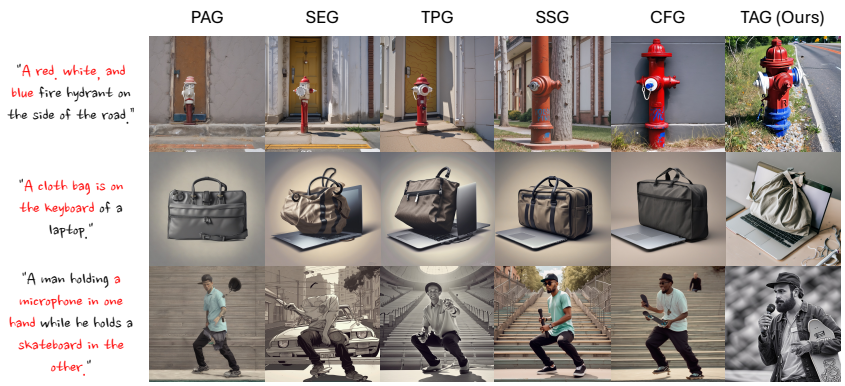


Figure 1: Qualitative comparison of guidance methods on SDXL (Podell et al., 2024) under identical prompts and random seeds. Red words indicate parts of the prompt that baselines fail to reflect. Training-free baselines (Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025; Zhang et al., 2026) struggle to maintain visual quality, and CFG (Ho & Salimans, 2022) requires additional training. Our method generates realistic images that align with the given prompt accurately without any training.

gains in generation quality remain marginal and typically fall short of the efficacy of CFG (Rajabi et al., 2025).

To understand why existing training-free guidance methods struggle to close the performance gap with CFG, we analyze guidance methods under conditions of varying complexity. We conjecture that for practical tasks with complex conditions (*e.g.*, text-conditional generation), the primary performance bottleneck lies in aligning the generated samples with the given condition, rather than merely refining fine-grained visual details. Our analysis reveals that guiding models to properly align samples with the condition becomes substantially more critical as the condition becomes more complex. Consequently, previous training-free methods, which predominantly perturb spatial operations among visual tokens, fail to address this core bottleneck, resulting in suboptimal prompt alignment and visual quality (see the first four columns of Fig. 1).

To address this problem, we propose a simple yet effective method, dubbed text attention guidance (TAG), that perturbs the interaction between visual and textual tokens. Based on our empirical observations, TAG degrades the attention weights between these modalities, yielding a weak prediction that specifically struggles with condition alignment. This targeted perturbation produces a precise negative signal that enhances both prompt alignment and visual quality (see the last column of Fig. 1). Furthermore, we perform extensive experiments across different architectures, spanning from traditional U-Net models (Ronneberger et al., 2015) (*i.e.*, SDXL (Podell et al., 2024)) to recent transformer-based models (Peebles & Xie, 2023) (*i.e.*, SD3.5 (Esser et al., 2024), Flux.1-dev (Black Forest Labs, 2024), and Qwen-Image (Wu et al., 2025)). While existing training-free methods show limited efficacy on U-Net models and struggle even further to improve generation quality on modern multi-modal transformer-based diffusion models (MM-DiTs), TAG provides consistent and substantial improvements across both architectures, particularly for complex prompts. We summarize our contributions as follows:

- We present an in-depth analysis of how diffusion models behave under varying conditional complexities, revealing that the primary performance bottleneck lies in condition alignment rather than spatial refinement.
- Based on this analysis, we propose TAG, a training-free guidance method that intentionally disrupts attention between visual and textual tokens to form a weak prediction, thereby providing a targeted guidance signal that enhances both prompt alignment and visual quality.
- We systematically evaluate our approach across different architectures, demonstrating that TAG consistently outperforms existing training-free methods on both U-Net and MM-DiT models across various benchmarks (Lin et al., 2014; Li et al., 2024).

2 Related Work

2.1 Diffusion Models and Architectures

Diffusion models (Ho et al., 2020; Song & Ermon, 2019; Lipman et al., 2023; Liu et al., 2023) demonstrate strong performance in generation tasks by learning to reverse a process that gradually adds Gaussian noise to the data. For text-to-image generation, these models require a mechanism to incorporate textual conditions. Early models (Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2024) rely on the U-Net (Ronneberger et al., 2015) architecture, where they utilize self-attention for spatial context and separate cross-attention layers to inject text embeddings from external text encoders (Radford et al., 2021; Raffel et al., 2020). As generation tasks demand higher visual fidelity, the architectural trend shifts towards transformer-based models (Peebles & Xie, 2023). Building upon this foundation, multi-modal designs such as MM-DiT improve generation performance by enhancing how the network processes the text modality. Specifically, recent models (Esser et al., 2024; Black Forest Labs, 2024; Wu et al., 2025) allocate separate sets of learnable parameters for textual and visual tokens. Instead of using cross-attention to merely inject fixed text embeddings, these models enable both modalities to continuously exchange information and update their respective representations through joint attention at every layer. Despite this shift in aggregating multi-modal information, existing training-free guidance methods (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025; Zhang et al., 2026) focus primarily on operations among visual tokens in U-Net architectures. In this work, we systematically evaluate guidance methods across different architectures, covering both U-Net and MM-DiT models.

2.2 Sampling Guidance

Guidance methods steer the sampling trajectory of diffusion models toward high-quality regions of the target distribution. For example, classifier guidance (Dhariwal & Nichol, 2021) adjusts the denoising direction using gradients from an auxiliary classifier, but it incurs inference overhead and requires training the classifier on noisy inputs. As an alternative, CFG (Ho & Salimans, 2022) bypasses the classifier by extrapolating predictions away from an unconditional prediction, which requires dropping the textual condition with a certain probability during training. Autoguidance (Karras et al., 2024a) instead replaces the unconditional prediction with the output of a smaller model trained for fewer iterations, showing that a negative signal sharing the failure modes of the original model provides effective guidance. However, all the aforementioned approaches require specialized training procedures, limiting their plug-and-play applicability to large-scale pretrained models.

To eliminate such training costs, several works extract a weak prediction directly from the diffusion model itself at inference time by perturbing the forward pass of the model. These training-free approaches can be broadly categorized into two groups. The first group perturbs the visual tokens directly. For instance, self-attention guidance (SAG) (Hong et al., 2023) uses self-attention maps to identify salient regions and blurs the corresponding visual tokens, feeding them back into the network to provide a weak prediction. More recently, token perturbation guidance (TPG) (Rajabi et al., 2025) shuffles visual tokens to introduce local spatial corruption, and self-swap guidance (SSG) (Zhang et al., 2026) further applies channel-wise swapping operations. The second group perturbs the attention mechanisms within diffusion models. Perturbed-attention guidance (PAG) (Ahn et al., 2024) replaces self-attention maps with an identity matrix to restrict spatial context, and smoothed energy guidance (SEG) (Hong, 2024) blurs the attention weights, preventing tokens from attending to specific visual features. However, the performance gains of existing training-free approaches in terms of visual fidelity remain marginal, especially for complex text prompts. Differently, TAG substantially improves the generation quality by directly targeting alignment between textual and visual tokens, which we identify as the primary bottleneck for modern text-conditional diffusion models.

3 Method

In this section, we first describe network architectures and existing guidance methods briefly (Sec. 3.1). We then analyze the behavior of diffusion models under conditions of varying complexity (Sec. 3.2). Based on this analysis, we introduce our TAG method (Sec. 3.3).

3.1 PRELIMINARIES

Network architecture. We denote by v_θ a diffusion model, parameterized by θ, that predicts the velocity for iteratively denoising noisy samples.[2] For text-conditional generation, diffusion models (Podell et al., 2024; Esser et al., 2024) exploit attention layers (Vaswani et al., 2017) to process noisy samples and textual conditions. We denote by $X \in \mathbb{R}^{N \times d}$ and $C \in \mathbb{R}^{M \times d}$ the visual tokens and the textual tokens, respectively, where N and M are the numbers of tokens and d is the channel dimension. Given query, key, and value matrices Q, K, and V, linearly projected from X or C, an attention layer computes attention logits S and the output, denoted by $\mathrm{Attn}(Q, K, V)$, as follows:

$$S = \frac{QK^\top}{\sqrt{d}}, \quad \mathrm{Attn}(Q, K, V) = \mathrm{softmax}(S)V, \tag{1}$$

where the softmax function is applied to each row of S, converting the logits into attention weights. MM-DiT models (Esser et al., 2024; Black Forest Labs, 2024; Wu et al., 2025) perform joint attention over the concatenated tokens $[X; C]$, such that Q, K, and V contain both visual and textual tokens. Following the token order, we partition the attention logits $S \in \mathbb{R}^{(N+M)\times(N+M)}$ into four blocks as follows:

$$S = \begin{bmatrix} S_{XX} & S_{XC} \\ S_{CX} & S_{CC} \end{bmatrix}, \tag{2}$$

where $S_{XX} \in \mathbb{R}^{N \times N}$ and $S_{CC} \in \mathbb{R}^{M \times M}$ denote the attention logits within visual and textual tokens, respectively. We denote by $S_{XC} \in \mathbb{R}^{N \times M}$ and $S_{CX} \in \mathbb{R}^{M \times N}$ the attention logits from visual queries to textual keys and those in the opposite direction, respectively. For U-Net models, which employ separate self- and cross-attention layers instead of joint attention, we use the same notation S_{XX} and S_{XC} for their respective attention logits, since they are computed between the same pairs of tokens.

Guidance method. Guidance methods steer the sampling process by extrapolating the prediction of the model away from a weak prediction $\hat{v}$ as follows:

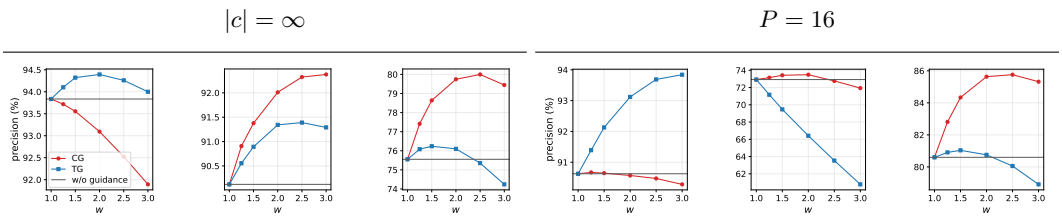
$$\tilde{v}_\theta = v_\theta + (w - 1)\,(v_\theta - \hat{v})\,, \tag{3}$$

where $w \geq 1$ is a guidance weight that controls the strength of guidance. At inference time, the guided prediction $\tilde{v}_\theta$ is used to update the noisy sample at each sampling step. CFG (Ho & Salimans, 2022) uses an unconditional prediction as the weak prediction $\hat{v}$, obtained by feeding an empty prompt into the model. Training-free approaches (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025; Zhang et al., 2026) instead obtain the weak prediction from the main model itself by perturbing operations in the forward pass of v_θ. Existing methods, however, perturb operations among visual tokens X only, and the effect of perturbing operations involving textual tokens C remains less explored.

3.2 EMPIRICAL ANALYSIS

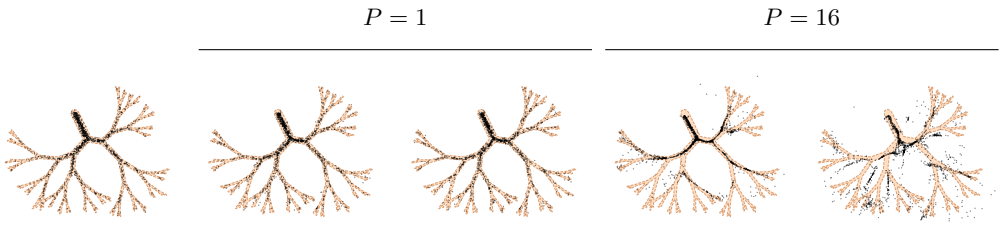
Here we present an analysis on how the complexity of a condition affects the effectiveness of guidance. The guidance method in Eq. (3) uses the prediction of a weak model as a negative signal, and extrapolates the prediction of a main model away from it. This suppresses the errors that the weak model makes more severely than the main model, and the errors to suppress are determined by how the weak model is degraded. We categorize guidance methods into condition guidance (CG) and target guidance (TG) based on the type of errors that the weak model makes. The weak model for CG fails to align samples with a given condition, whereas the one for TG fails to capture the target distribution even with the condition processed properly. For text-conditional generation, a diffusion model has to learn a large number of conditions that are difficult to interpret, and we conjecture that aligning samples with such complex conditions is the main bottleneck of the model. Existing training-free methods (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025; Zhang et al., 2026), however, perturb operations among visual tokens only, leaving the condition processed exactly as in the main model. The weak model makes errors only in capturing the target distribution, which corresponds to TG. They do not address the bottleneck of condition alignment, providing marginal gains in both visual quality and condition alignment (Fig. 1).

[2] We provide a detailed explanation of diffusion models in Appendix A for completeness.



(a) $P = 1$. (b) $P = 4$. (c) $P = 16$. (d) $|c| = 64$. (e) $|c| = 1024$. (f) $|c| = 4096$.

Figure 2: Comparison of CG and TG in our empirical analysis with conditions of varying complexity. We plot precision (%) with respect to the guidance weight w for (a-c) different numbers of chunks P with $|c| = \infty$, and for (d-f) different numbers of conditions $|c|$ with $P = 16$. TG provides better precision when the condition is easy to interpret (*i.e.*, small P) or the number of conditions is small (*i.e.*, small $|c|$), while CG becomes more effective as the condition becomes harder to interpret or the number of conditions increases.

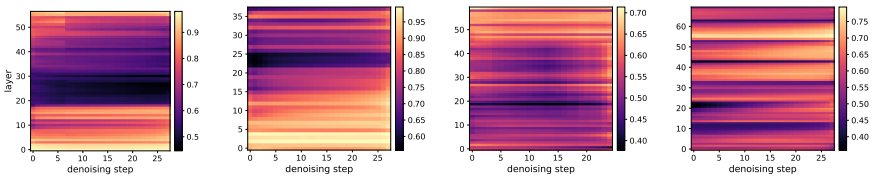


(a) Ground truth. (b) CG (c) TG (d) CG (e) TG

Figure 3: Qualitative comparison of CG and TG in our empirical analysis. The orange region indicates the support of the ground-truth distribution, and approximately 99% of the probability mass lies inside the region. We show in (a) ground-truth samples (black) drawn from the distribution, and in (b-e) generated samples (black) with CG and TG for (b-c) $P = 1$ and (d-e) $P = 16$. Both types of guidance produce samples along the branches faithfully for the simple condition. However, for the complex condition, TG generates many samples off the support, while CG keeps most of them on the support.

To compare the effectiveness of CG and TG, we conduct a synthetic experiment that simulates two guidance methods and conditions of varying complexity. We use 2D tree-shaped distributions (Karras et al., 2024a) as target distributions, each specified by a continuous condition that determines the position and orientation of the target distribution. We control the complexity of the condition in terms of the difficulty in interpreting each condition and the number of distinct conditions to learn. Specifically, we split the continuous condition space into P chunks and permute the chunks randomly before mapping each condition to its target distribution, making the mapping discontinuous across chunk boundaries. We also restrict the conditions for both training and evaluation to a fixed set of $|c|$ values, where $|c| = \infty$ indicates that conditions are drawn from the entire continuous space. By increasing P and $|c|$, we can make the condition harder to interpret and increase the number of conditions to learn, similar to the challenges of text-conditional generation. We then train a diffusion model that consists of an encoder processing the condition and a denoiser predicting the velocity from the processed condition and a noisy 2D point. We train two variants of the weak model, one with a narrower encoder and the other with a narrower denoiser, to simulate CG and TG, respectively. To quantitatively compare CG and TG, we measure precision (*i.e.*, the fraction of generated samples within the support of the ground-truth distribution) of each guidance method. More details can be found in Appendix B.

We show in Fig. 2(a-c) precision curves for $P \in \{1, 4, 16\}$ with $|c| = \infty$. TG outperforms CG for the condition that is easy to interpret ($P = 1$), while CG provides better precision as P increases, with the gap between the two widening. This suggests that for complex conditions, guiding diffusion models to align samples with the condition is more effective than guiding them to refine how they model the target distribution. This is also evident in Fig. 3, where we show qualitative results for $P = 1$ and $P = 16$. TG produces many samples off the support for the complex condition, while CG keeps most of them on the support. We also show in Fig. 2(d-f) precision curves with $P = 16$ and $|c| \in \{64, 1024, 4096\}$. CG outperforms TG for $|c| \geq 1024$, while TG provides better precision for $|c| = 64$. Our analysis shows that CG becomes increasingly effective relative to TG as the condition becomes more complex. Since text-conditional generation involves a large and complicated condition space, this observation suggests that condition alignment may become particularly impor-



(a) Flux.1-dev. (b) SD3.5. (c) Qwen-Image. (d) SDXL.

Figure 4: Visualization of the normalized entropy of attention weights from visual to textual tokens for each layer across denoising steps. We show the normalized entropy in Eq. (5) for (a) Flux.1-dev, (b) SD3.5, (c) Qwen-Image, and (d) SDXL, averaged over 8 prompts. A small number of layers show distinctly low entropy regardless of the denoising step, indicating that they attend to a few textual tokens sparsely.

tant in this setting. However, existing training-free methods mainly perturb operations among visual tokens, and are therefore limited in their ability to directly address errors arising from condition alignment. Based on this observation, we instead propose to perturb the interaction between visual and textual tokens to obtain a weak model.

3.3 Text attention guidance

As discussed in Sec. 3.2, existing training-free methods perturb operations among visual tokens, such as attention with S_{XX} in Eq. (2). We instead perturb the attention logits S_{XC} from visual queries to textual keys, where information flows from the condition to the visual tokens. The weak model thus fails to align generated samples with the text prompt properly, which corresponds to CG. In the following, we describe the details of our TAG method.

Text attention perturbation. Visual tokens selectively attend to a few textual tokens relevant to their contents, aligning generated samples with the text prompt (Hertz et al., 2023). We prevent this by flattening the attention logits S_{XC} with a temperature $\tau > 1$ as follows:

$$\hat{S} = \begin{bmatrix} S_{XX} & S_{XC}/\tau \\ S_{CX} & S_{CC} \end{bmatrix}, \tag{4}$$

where $\hat{S}$ replaces S in Eq. (1). A large temperature makes the attention weights from each visual token to textual tokens close to uniform, such that visual tokens aggregate the textual tokens without distinguishing relevant ones from the others. We obtain the weak prediction $\hat{v}$ in Eq. (3) from a forward pass of v_θ with the perturbed logits in Eq. (4). For U-Net models, we apply the same perturbation to the logits of cross-attention layers.

Layer selection. Existing methods perturb a subset of layers, since perturbing all layers degrades the weak model too severely to provide a useful negative signal. The subset is chosen empirically for each model (*e.g.*, the first several layers in the mid-block of U-Net), which requires an extensive search over combinations of layers. Moreover, such heuristic selection strategies are not directly applicable to transformer-based models since all the layers are structurally identical, providing no natural candidates for the layers to perturb. To address this, we design a principled layer selection method tailored to TAG that generalizes across different network architectures. Specifically, we measure how sparsely the visual tokens of the l-th layer attend to textual tokens with the normalized entropy defined as follows:

Table 1: Spearman correlations ρ of the layer-wise entropy in Eq. (5), averaged over all 4,950 pairs of 100 prompts.

	Flux.1-dev	SD3.5	Qwen-Image	SDXL
ρ	0.9953	0.9920	0.8523	0.9273

$$H_l = -\frac{1}{N \log M} \sum_{i=1}^{N} \sum_{j=1}^{M} \bar{A}^l_{ij} \log \bar{A}^l_{ij}, \tag{5}$$

where $\bar{A}^l \in \mathbb{R}^{N \times M}$ is obtained by applying the softmax function to each row of S_{XC} in the l-th layer. We average the entropy over attention heads. The normalization term bounds H_l within $[0, 1]$, where $H_l = 1$ corresponds to uniform attention over textual tokens. A low entropy indicates that visual tokens attend to a few textual tokens selectively, suggesting that the layer is crucial for

Table 2: Quantitative comparison of guidance methods on the validation split of COCO (Lin et al., 2014) for Flux.1-dev (Black Forest Labs, 2024) and SDXL (Podell et al., 2024). **Bold** denotes the best and underline the second-best result in each column. †Requires training.

Model	Method	IR ↑	PickScore ↑	HPSv2 ↑	Aesthetic ↑	FID ↓	CLIP ↑
Flux.1-dev	w/o guidance	0.269	21.36	0.242	5.50	30.07	29.53
	PAG (Ahn et al., 2024)	0.643	22.08	0.275	5.82	30.27	29.99
	SEG (Hong, 2024)	0.752	22.25	0.288	5.78	34.55	30.03
	TPG (Rajabi et al., 2025)	0.083	21.21	0.240	5.72	44.98	29.68
	SSG (Zhang et al., 2026)	-2.080	18.48	0.104	3.90	163.43	22.41
	CFG† (Ho & Salimans, 2022)	0.949	22.62	0.278	5.51	**28.73**	**31.44**
	Ours	**1.062**	**23.09**	**0.290**	**5.85**	30.61	31.04
SDXL	w/o guidance	-0.853	20.18	0.188	5.67	55.09	27.01
	PAG (Ahn et al., 2024)	-0.516	20.69	0.210	5.59	39.43	28.44
	SEG (Hong, 2024)	-0.018	21.43	0.253	5.86	37.50	28.90
	TPG (Rajabi et al., 2025)	0.226	21.84	0.264	5.82	32.90	29.76
	SSG (Zhang et al., 2026)	0.290	22.17	0.274	**5.89**	31.81	29.76
	CFG† (Ho & Salimans, 2022)	0.820	**22.56**	**0.284**	5.59	**25.33**	31.83
	Ours	**0.880**	22.18	0.280	5.69	32.16	**31.86**

Table 3: Quantitative comparison of guidance methods on GenAI-Bench (Li et al., 2024) in terms of VQAScore. We report the average score for each skill and over all prompts. **Bold** denotes the best and underline the second-best result in each column. †Requires training.

		Basic					Advanced					
Model	Method	Attr.	Scene	Spatial	Action	Part	Count.	Comp.	Diff.	Neg.	Univ.	Overall
Flux.1-dev	w/o guidance	0.693	0.703	0.702	0.705	0.665	0.684	0.671	0.676	**0.475**	0.645	0.680
	PAG (Ahn et al., 2024)	0.725	0.729	0.749	0.744	0.700	0.718	0.705	0.728	0.462	0.677	0.713
	SEG (Hong, 2024)	0.730	0.734	0.741	0.741	0.706	0.725	0.712	0.719	0.463	0.664	0.712
	TPG (Rajabi et al., 2025)	0.662	0.669	0.673	0.667	0.634	0.649	0.647	0.641	0.466	0.640	0.650
	SSG (Zhang et al., 2026)	0.334	0.328	0.336	0.302	0.327	0.382	0.398	0.368	0.409	0.320	0.332
	CFG† (Ho & Salimans, 2022)	0.765	0.760	0.780	0.784	0.752	0.759	0.729	0.763	0.447	0.706	0.750
	Ours	**0.780**	**0.774**	**0.795**	**0.801**	**0.766**	**0.770**	**0.742**	**0.774**	0.465	**0.712**	**0.763**
SDXL	w/o guidance	0.585	0.603	0.590	0.580	0.568	0.580	0.595	0.596	0.512	0.560	0.582
	PAG (Ahn et al., 2024)	0.598	0.613	0.601	0.609	0.593	0.592	0.569	0.597	0.505	0.564	0.595
	SEG (Hong, 2024)	0.627	0.636	0.627	0.628	0.609	0.615	0.592	0.599	0.500	0.578	0.623
	TPG (Rajabi et al., 2025)	0.636	0.640	0.649	0.639	0.604	0.637	0.613	0.615	0.506	0.579	0.633
	SSG (Zhang et al., 2026)	0.627	0.638	0.626	0.630	0.593	0.622	0.608	0.592	0.490	0.587	0.622
	CFG† (Ho & Salimans, 2022)	0.733	0.738	0.733	0.742	0.713	0.723	0.682	0.686	0.523	0.663	0.726
	Ours	**0.763**	**0.765**	**0.766**	**0.782**	**0.751**	**0.762**	**0.735**	**0.743**	**0.534**	**0.696**	**0.756**

condition alignment. We thus select the k layers with the lowest entropy to apply the perturbation in Eq. (4). Note that the entropy varies with the denoising step and the prompt, and the layers to perturb might differ across inputs. However, we have found that the layers with low entropy are largely consistent across denoising steps and prompts. A similar observation is reported in (Kim et al., 2025), where visual tokens in a specific layer of MM-DiT attend to the textual tokens describing the objects they belong to, regardless of the prompt and the denoising step, such that the attention weights of the layer serve as segmentation masks. To verify this, we show in Fig. 4 the entropy of each layer across denoising steps. A small number of layers show distinctly low entropy regardless of the denoising step, while the entropy of the other layers is much higher. We also compare in Table 1 the ranking of layers across prompts. Specifically, we compute Spearman correlations of the entropy, averaged over denoising steps, for all 4,950 pairs of 100 prompts, and report the mean value. The correlation is above 0.85 for all models. This suggests that the selective attention to textual tokens is a property of each layer rather than of the input. Based on this, we average the entropy over denoising steps with a few prompts, and fix the selection once for each model.

4 Experiments

In this section, we describe our experimental settings (Sec. 4.1), compare TAG with state-of-the-art guidance methods (Sec. 4.2), and present a detailed analysis of our approach (Sec. 4.3).

4.1 Implementation details

Datasets and models. We mainly evaluate TAG on the validation split of COCO (Lin et al., 2014), using the first caption of each of the 5,000 images as a prompt. To evaluate prompt alignment explicitly, we also use GenAI-Bench (Li et al., 2024), which consists of 1,600 compositional prompts annotated with the skills they require. We evaluate all methods on SDXL (Podell et al., 2024) for

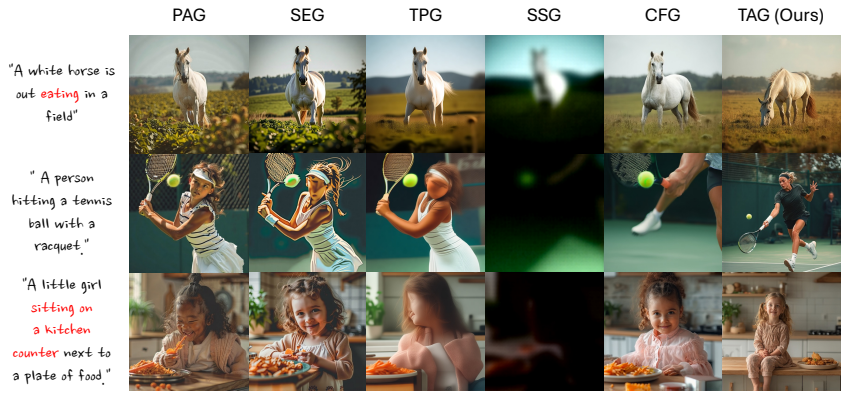


Figure 5: Qualitative comparison of guidance methods on Flux.1-dev (Black Forest Labs, 2024) with the same prompts and random seeds. Baselines fail to reflect the parts of the prompt in red.

Table 4: Quantitative results of TAG on 5,000 prompts of the validation split of COCO (Lin et al., 2014) with Flux.1-dev for different temperatures τ in Eq. (4). The highlighted row corresponds to our default setting, selected on 200 prompts from the training split.

Setting	IR ↑	Pick ↑	HPSv2 ↑	Aes. ↑	FID ↓	CLIP ↑
$\tau = 1.3$	1.028	22.97	0.284	5.83	30.30	30.93
$\tau = 1.5$	1.062	23.09	0.290	5.85	30.61	31.04
$\tau = 1.7$	1.067	23.11	0.292	5.85	30.90	31.04
$\tau = 2$	1.068	23.10	0.292	5.84	31.31	31.04
$\tau = 5$	1.022	23.00	0.290	5.80	32.04	31.00
$\tau = 10$	1.016	22.99	0.290	5.79	32.23	30.99

Table 5: Quantitative results of TAG on 5,000 prompts of the validation split of COCO (Lin et al., 2014) with Flux.1-dev for different numbers of perturbed layers k. "Reverse" perturbs the 16 layers of highest entropy in Eq. (5) instead of the lowest. The highlighted row corresponds to our default setting.

Setting	IR ↑	Pick ↑	HPSv2 ↑	Aes. ↑	FID ↓	CLIP ↑
$k = 12$	1.017	22.87	0.286	5.84	28.93	30.90
$k = 14$	1.030	22.93	0.287	5.79	28.32	31.09
$k = 16$	1.062	23.09	0.290	5.85	30.61	31.04
$k = 18$	1.063	23.11	0.291	5.84	31.16	30.97
$k = 20$	1.049	23.10	0.289	5.84	32.03	30.91
Reverse	0.850	22.85	0.284	5.76	30.50	30.46

U-Net, and SD3.5 (Esser et al., 2024), Flux.1-dev (Black Forest Labs, 2024), and Qwen-Image (Wu et al., 2025) for MM-DiT. We generate images of size 1024×1024 with 50 sampling steps for SDXL and 28 steps for SD3.5, and images of size 512×512 with 28 steps for Flux.1-dev and 25 steps for Qwen-Image. We tune hyperparameters of all methods on 200 prompts from the training split of COCO, and fix them for all experiments. Due to the space limit, we report results for SD3.5 and Qwen-Image in Appendix C, and more details on experimental settings in Appendix B.

Evaluation metrics. For COCO, we mainly use ImageReward (IR) (Xu et al., 2023), PickScore (Kirstain et al., 2023), and HPSv2 (Wu et al., 2023), which are trained on human preference data and thus reflect both visual quality and prompt alignment. We also report the aesthetic score (Schuhmann et al., 2022), the CLIP score (Hessel et al., 2021), and FID (Heusel et al., 2017) for reference. Note that the CLIP score is often unreliable (Lin et al., 2024), and FID penalizes any deviation from the style of COCO images regardless of visual quality. For GenAI-Bench, we use VQAScore (Lin et al., 2024), which measures prompt alignment with a visual question answering model, and report the average score for each skill and over all prompts.

4.2 Results

Quantitative results. We show in Table 2 quantitative comparisons of guidance methods on COCO for Flux.1-dev and SDXL. We summarize our findings as follows: (1) CFG and TAG (correspond to CG) outperform existing training-free methods (correspond to TG) in terms of human preference metrics for both models. This confirms our analysis in Sec. 3.2 that guiding diffusion models to align samples with the prompt is more important than refining spatial details in text-conditional generation. (2) TAG provides results comparable to or even better than those of CFG without any training. This suggests that perturbing text attention is as effective as the unconditional prediction of CFG, which requires addition training. (3) TPG and SSG degrade the performance of Flux.1-dev even below that of unguided sampling, although they are effective for SDXL. This indicates that perturbations designed for U-Net do not transfer to MM-DiT, whereas TAG generalizes to both.

We also show in Table 3 quantitative comparisons of guidance methods on GenAI-Bench for Flux.1-dev and SDXL. TAG outperforms existing training-free methods in terms of VQAScore, by large margins of up to 0.123 for both models. This demonstrates that perturbing text attention improves prompt alignment. Please refer to Tables 7 and 8 in Appendix C for more quantitative results on SD3.5 and Qwen-Image.

Qualitative results. We show in Figs. 1 and 5 qualitative comparisons of guidance methods on SDXL and Flux.1-dev, respectively. Training-free baselines degrade visual quality severely, producing oversaturated colors, blurry textures, and distorted objects, and fail to reflect parts of the prompt, highlighted in red. TAG generates realistic images that reflect the whole prompt without any training. This suggests that perturbing text attention improves visual quality and prompt alignment jointly. Qualitative comparisons on SD3.5 and Qwen-Image can be found in Figs. 7 and 8 of Appendix C.

4.3 Discussion

Analysis of temperature. We show in Table 4 quantitative results of TAG on COCO with Flux.1-dev for different temperatures τ in Eq. (4). TAG outperforms CFG in Table 2 in terms of IR for all values of τ, where IR varies by less than 0.06. This indicates that TAG is robust to the temperature, and does not require extensive hyperparameter tuning. We argue that perturbing text attention itself is the key to improving the performance, while the temperature only controls the strength of the perturbation.

Analysis of layer selection. We show in Table 5 quantitative results for different numbers of perturbed layers k. As with the temperature, TAG outperforms CFG in terms of IR for all values of k from 12 to 20, where IR varies by less than 0.05, indicating that TAG is robust to the number of perturbed layers. To verify that the entropy in Eq. (5) selects the layers to perturb properly, we also perturb the 16 layers of highest entropy instead, denoted by "Reverse". This degrades IR by 0.21 and lowers all the other metrics except FID with the same k and τ, a much larger change than varying k or τ. This demonstrates that the layers attending to a few textual tokens selectively are crucial for condition alignment, and that our entropy-based criterion identifies them effectively without an extensive search. However, even with the reverse selection, TAG still outperforms all existing training-free methods in Table 2, suggesting that CG is substantially more effective than TG for text-conditional generation.

Visualization of predictions. We visualize in Fig. 6 the clean samples predicted at the fourth sampling step of Flux.1-dev from the weak prediction $\hat{v}$, the main prediction v_θ, and the guided prediction $\tilde{v}_\theta$ of TAG in Eq. (3) for the first prompt in Fig. 5. The weak prediction fails to reflect the prompt, generating the horse with its head up, while the main prediction overlays the horse with its head raised and lowered. The guided prediction, extrapolated away from the weak one, emphasizes the lowered head of the horse (appears blue), while suppressing the horse with its head up (appears white). This suggests that the weak prediction of TAG provides a negative signal targeted at condition alignment, making TAG the only method that accurately reflects the "eating" part of the prompt in Fig. 5.



Figure 6: Visualization of clean samples predicted at the fourth step of Flux.1-dev from the weak prediction $\hat{v}$ (left), the main prediction v_θ (middle), and the guided prediction $\tilde{v}_\theta$ (right) of TAG in Eq. (3).

5 Conclusion

We have shown through an empirical analysis that guiding samples to align with the condition is crucial for tasks with complex conditions. Based on this, we have introduced TAG, a training-free guidance method that flattens the attention from visual to textual tokens. We have also presented an entropy-based criterion for selecting the layers to perturb, which generalizes across U-Net and MM-DiT architectures. We have demonstrated that TAG outperforms existing training-free methods by large margins and matches or even surpasses CFG without any training.

6 AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing of this paper, including correcting grammar and improving the readability of sentences. We have not used generative AI tools for any of the tasks that require disclosure, including implementing the method, designing experiments, etc. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *ECCV*, 2024.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *NeurIPS*, 2022.

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In *NeurIPS*, 2024.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *NeurIPS*, 2024a.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *CVPR*, 2024b.

Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4Diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers. In *NeurIPS*, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. GenAI-Bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

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 *ECCV*, 2014.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

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 *ICLR*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token perturbation guidance for diffusion models. In *NeurIPS*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.

Weijia Zhang, Yuehao Liu, Shanyan Guan, Wu Ran, Yanhao Ge, Wei Li, and Chao Ma. Guiding a diffusion model by swapping its tokens. In *CVPR*, 2026.

In the appendix, we provide a detailed description of diffusion models (Appendix A), implementation details (Appendix B), and more results (Appendix C).

A Diffusion models

Let $x \sim p_{\text{data}}(x)$ and $\epsilon \sim \mathcal{N}(0, I)$ denote a data sample and Gaussian noise, respectively. Diffusion models (Ho et al., 2020; Song & Ermon, 2019; Lipman et al., 2023; Liu et al., 2023) define a noisy sample z_t at timestep $t \in [0, T]$ as follows:

$$z_t = \alpha_t x + \sigma_t \epsilon, \tag{6}$$

where α_t and σ_t are predefined noise scheduling coefficients that decrease and increase monotonically with t, respectively, satisfying $\alpha_0 = \sigma_T = 1$ and $\alpha_T = \sigma_0 = 0$ (*i.e.*, $z_0 = x$ and $z_T = \epsilon$). We denote by v_θ a diffusion model, parameterized by θ, that takes z_t, t, and a condition c (*e.g.*, a text prompt) as inputs. It is trained to predict the velocity $\mathrm{d}z_t/\mathrm{d}t$ with the following objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x,\epsilon}\left[\left\| v_\theta(z_t, t, c) - (\dot{\alpha}_t x + \dot{\sigma}_t \epsilon)\right\|_2^2\right], \tag{7}$$

where $\dot{\alpha}_t$ and $\dot{\sigma}_t$ denote the time derivatives of α_t and σ_t, respectively. At inference time, a sample is generated by solving the ODE $\mathrm{d}z_t/\mathrm{d}t = v_\theta(z_t, t, c)$ backward in time, starting from $z_T \sim \mathcal{N}(0, I)$ (Song et al., 2021):

$$z_0 = z_T + \int_T^0 v_\theta(z_t, t, c)\,\mathrm{d}t. \tag{8}$$

In practice, the integral is approximated using numerical ODE solvers (*e.g.*, the Euler method) over discretized timesteps.

B Implementation details

B.1 Empirical analysis

Target distribution and condition. We use the two-dimensional tree-shaped distribution of Karras et al. (2024a), a Gaussian mixture with 1,016 components placed along the branches of a tree, and denote by p_{gt} its density. The condition c specifies a rigid motion of the tree, namely the position r of the root within $[-0.2, 0.2]^2$ and the direction ψ of the trunk within $[0, 2\pi)$. We denote by R_ψ the rotation matrix of angle ψ and by o the root of the original tree, and obtain the conditional density and its score in closed form as follows:

$$p(x \mid c) = p_{\text{gt}}(y), \qquad \nabla_x \log p(x \mid c) = R_\psi\, \nabla_y \log p_{\text{gt}}(y), \tag{9}$$

where $y = R_\psi^\top (x - r) + o$ is the sample mapped back to the original tree. To control the complexity of the condition, we draw $c \sim \mathcal{U}(0, 1)^3$ and map it to the rigid motion by shuffling each coordinate c_j as follows:

$$b_j = \lfloor c_j P \rfloor, \qquad u_j = \frac{\pi_j(b_j) + c_j P - b_j}{P}, \qquad j \in \{1, 2, 3\}, \tag{10}$$

where P is the number of chunks, b_j is the index of the chunk that c_j falls into, and π_j is a fixed random permutation of the chunks. We rescale $u \in [0, 1]^3$ linearly to the ranges of r and ψ. This map is a piecewise-linear bijection of the unit cube, so that the family of trees is the same for every P, while P controls how discontinuous the map from c to the tree is. To vary the number of conditions $|c|$, we draw c from a fixed set of $|c|$ conditions during training instead of the continuous prior.

Network. The network consists of an encoder E and a denoiser D that meet at a single vector, so that narrowing one of them leaves the other unchanged. E is an MLP with four hidden layers that takes the condition c only and outputs a 256-dimensional vector h. D consists of three residual MLP blocks that take a noisy sample z_t, where σ_t and h modulate each block through AdaLN (Peebles & Xie, 2023). Following EDM (Karras et al., 2022), we set $\alpha_t = 1$ in Eq. (6) and index the noise level by σ_t. D outputs an energy $\log p_\theta(z_t \mid c, \sigma_t)$, and we obtain the score by differentiating it

Table 6: Hyperparameters of all methods. We report the guidance weight w, the indices of perturbed layers, and the method-specific hyperparameter λ for each model, where λ denotes σ for SEG, the swap ratio for SSG, and τ for TAG. For TAG, top-k denotes the k layers of lowest entropy. A dash indicates that the method does not have the corresponding hyperparameter. We also report the resolution and the number of sampling steps for each model.

Method	Flux.1-dev			SD3.5			Qwen-Image			SDXL		
	w	Layers	λ	w	Layers	λ	w	Layers	λ	w	Layers	λ
Resolution	512×512			1024×1024			512×512			1024×1024		
Sampling steps	28			28			25			50		
CFG	3.5	–	–	3.5	–	–	4	–	–	7.5	–	–
PAG	8	7	–	3	13	–	3	40	–	3	24	–
SEG	3	0–18	10^7	2	16–20	10^7	4	25–31	10^7	4	24–33	10^7
TPG	1.25	2–6	–	1.25	3–12	–	1.25	22–36	–	4	–	–
SSG	2	13–18	0.1	2	32–37	0.1	1.1	27–32	0.1	4	34–39	0.1
TAG (ours)	5.5	top-16	1.5	5.5	top-5	3	6	top-30	1.5	18.5	top-70	3

with respect to z_t. Note that the score is a linear reparameterization of the velocity at a given noise level (Song et al., 2021), and the guidance in Eq. (3) thus applies to the score in the same way. For the main model, E and D have widths of 560 and 232, respectively, with 1.1M parameters each. The weak model for CG narrows E to a width of 280, and the one for TG narrows D to a width of 100, removing 35.5% and 35.7% of the parameters, respectively, so that CG and TG differ only in which component is weakened.

Training and sampling. We train all models with a denoising score matching loss for 32,768 iterations with a batch size of 4,096. The target is the score of the noisy conditional density, which we compute in closed form from Eq. (9), since a Gaussian mixture corrupted by Gaussian noise remains a Gaussian mixture. Following EDM2 (Karras et al., 2024b), we sample the noise level as $\ln \sigma_t \sim \mathcal{N}(-2.3, 1.5^2)$ and weight the squared error by σ_t^2. The weak models are trained with the same schedule and differ from the main model in width only. We sample with the Heun sampler of Karras et al. (2022) for 32 steps, and compute the guided score with Eq. (3) using the same initial noise and conditions for all models and guidance weights.

Evaluation. We define the support of the ground-truth distribution as the super-level set $\{x : p(x \mid c) \geq \eta \max_x p(x \mid c)\}$ with $\eta = e^{-4.49}$, which corresponds to the outline of the tree in Fig. 3 following Karras et al. (2024a). We evaluate on the same 64 conditions from the training distribution for every P and $|c|$. For each guidance weight $w \in \{1, 1.25, 1.5, 2, 2.5, 3\}$, we compute precision with 256 samples for each condition, and average over the conditions and four training seeds. We show in Fig. 2 the results for $P \in \{1, 4, 16\}$ with $|c| = \infty$ and for $|c| \in \{64, 1024, 4096\}$ with $P = 16$.

B.2 Main experiments

Datasets and models. The validation split of COCO (Lin et al., 2014) contains 5,000 images with five captions each, and we generate a single image for the first caption of each image. GenAI-Bench (Li et al., 2024) annotates each prompt with basic skills (attribute, scene, and spatial, action, and part relations) and advanced skills (counting, comparison, differentiation, negation, and universality). We fix the random seed for each prompt, so that every method generates the image from the same initial noise. Flux.1-dev (Black Forest Labs, 2024) is a guidance-distilled model taking the guidance scale as an additional input and thus performs guidance without computing an unconditional prediction. For a fair comparison, we disable this embedded guidance by fixing its scale to 1 and apply all methods on top of the same base model.

Hyperparameters. We summarize in Table 6 the hyperparameters of all methods, together with the resolution and the number of sampling steps for each model. We report the guidance weight w in Eq. (3) throughout, so that $w = 1$ corresponds to unguided sampling for every method. Layer

Table 7: Quantitative comparison of guidance methods on the validation split of COCO (Lin et al., 2014) for SD3.5 (Esser et al., 2024) and Qwen-Image (Wu et al., 2025). **Bold** denotes the best and underline the second-best result in each column. †CFG is the only method that is not training-free.

Model	Method	IR ↑	PickScore ↑	HPSv2 ↑	Aesthetic ↑	FID ↓	CLIP ↑
SD3.5	w/o guidance	0.441	21.77	0.244	5.38	34.14	30.77
	PAG (Ahn et al., 2024)	0.704	22.33	0.278	5.62	**24.38**	30.81
	SEG (Hong, 2024)	0.772	22.41	0.287	**5.72**	30.17	31.06
	TPG (Rajabi et al., 2025)	-0.114	20.89	0.223	5.31	53.65	30.15
	SSG (Zhang et al., 2026)	-2.186	17.87	0.097	3.85	154.58	21.81
	CFG† (Ho & Salimans, 2022)	1.047	22.80	0.293	5.49	26.14	**31.87**
	Ours	**1.054**	**22.88**	**0.306**	5.64	28.85	31.13
Qwen-Image	w/o guidance	0.954	22.44	0.275	5.48	**28.33**	31.48
	PAG (Ahn et al., 2024)	1.081	22.71	0.290	5.57	34.53	31.65
	SEG (Hong, 2024)	1.146	22.81	0.306	5.64	40.46	31.31
	TPG (Rajabi et al., 2025)	0.704	21.91	0.272	**5.68**	52.70	30.99
	SSG (Zhang et al., 2026)	0.980	22.43	0.282	5.58	34.30	31.56
	CFG† (Ho & Salimans, 2022)	1.238	23.04	0.302	5.52	38.91	**32.12**
	Ours	**1.240**	**23.10**	**0.317**	5.52	41.46	31.54

Table 8: Quantitative comparison of guidance methods on GenAI-Bench (Li et al., 2024) in terms of VQAScore for SD3.5 (Esser et al., 2024) and Qwen-Image (Wu et al., 2025). We report the average score for each skill and over all prompts. **Bold** denotes the best and underline the second-best result in each column. †CFG is the only method that is not training-free.

		Basic					Advanced					
Model	Method	Attr.	Scene	Spatial	Action	Part	Count.	Comp.	Diff.	Neg.	Univ.	Overall
SD3.5	w/o guidance	0.726	0.715	0.739	0.733	0.712	0.706	0.695	0.706	**0.500**	0.647	0.709
	PAG (Ahn et al., 2024)	0.738	0.729	0.739	0.744	0.739	0.706	0.707	0.725	0.480	0.670	0.722
	SEG (Hong, 2024)	0.736	0.719	0.751	0.748	0.726	0.712	0.712	0.702	0.466	0.659	0.720
	TPG (Rajabi et al., 2025)	0.658	0.653	0.661	0.658	0.642	0.644	0.637	0.626	0.493	0.607	0.645
	SSG (Zhang et al., 2026)	0.253	0.253	0.243	0.230	0.247	0.260	0.296	0.265	0.333	0.249	0.253
	CFG† (Ho & Salimans, 2022)	**0.779**	**0.769**	**0.794**	**0.797**	**0.767**	**0.760**	**0.743**	**0.754**	0.471	**0.718**	**0.763**
	Ours	0.758	0.748	0.771	0.773	0.754	0.736	0.727	0.743	0.456	0.667	0.741
Qwen-Image	w/o guidance	0.784	0.775	0.799	0.798	0.776	0.778	0.769	0.793	0.484	0.727	0.769
	PAG (Ahn et al., 2024)	0.802	0.790	0.815	0.815	0.797	0.787	0.778	0.801	0.484	0.745	0.784
	SEG (Hong, 2024)	0.806	0.788	0.822	0.819	0.800	0.796	0.798	0.818	0.470	0.745	0.788
	TPG (Rajabi et al., 2025)	0.751	0.737	0.773	0.763	0.736	0.751	0.746	0.765	**0.495**	0.710	0.739
	SSG (Zhang et al., 2026)	0.786	0.771	0.801	0.799	0.777	0.781	0.777	0.797	0.481	0.729	0.771
	CFG† (Ho & Salimans, 2022)	**0.819**	**0.805**	**0.833**	**0.835**	**0.806**	**0.812**	**0.806**	**0.835**	0.483	**0.762**	**0.800**
	Ours	0.807	0.792	0.825	0.821	0.796	0.795	0.796	0.826	0.469	0.754	0.790

indices enumerate the attention layers in execution order, namely the joint attention layers for MM-DiT models and the transformer blocks of the U-Net for SDXL. We tune all methods on 200 captions from the training split of COCO using ImageReward as the criterion, except for the baselines on SDXL, for which we use the settings reported in their papers, and CFG, which uses the default guidance weight of each model. For TAG, we perturb the k layers of lowest entropy in Eq. (5), measured once per model over 8 prompts from the same split and averaged over attention heads, visual tokens, and denoising steps.

Evaluation. Following SSG (Zhang et al., 2026), we compute FID after center-cropping images along the longer edge and resizing them to 256×256, using the validation split of COCO as the reference, and apply the same preprocessing to the aesthetic score.

C More results

We show in Tables 7 and 8 quantitative comparisons of guidance methods on COCO (Lin et al., 2014) and GenAI-Bench (Li et al., 2024), respectively, for SD3.5 (Esser et al., 2024) and Qwen-Image (Wu et al., 2025). We can see that the results are consistent with those for Flux.1-dev and SDXL in Tables 2 and 3. TAG provides the best IR, PickScore, and HPSv2 for both models, outperforming CFG without any training, whereas TPG degrades the performance below that of unguided sampling for both models, and SSG for SD3.5. TAG also outperforms all existing training-free

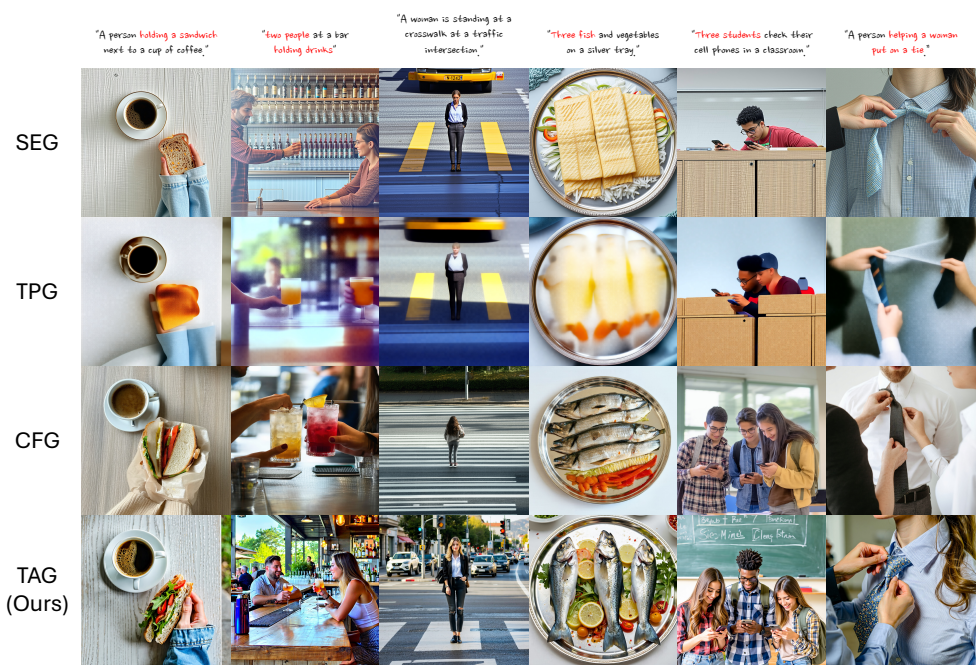


Figure 7: Qualitative comparison of guidance methods on SD3.5 (Esser et al., 2024) under identical prompts and random seeds. Red words indicate parts of the prompt that baselines fail to reflect.

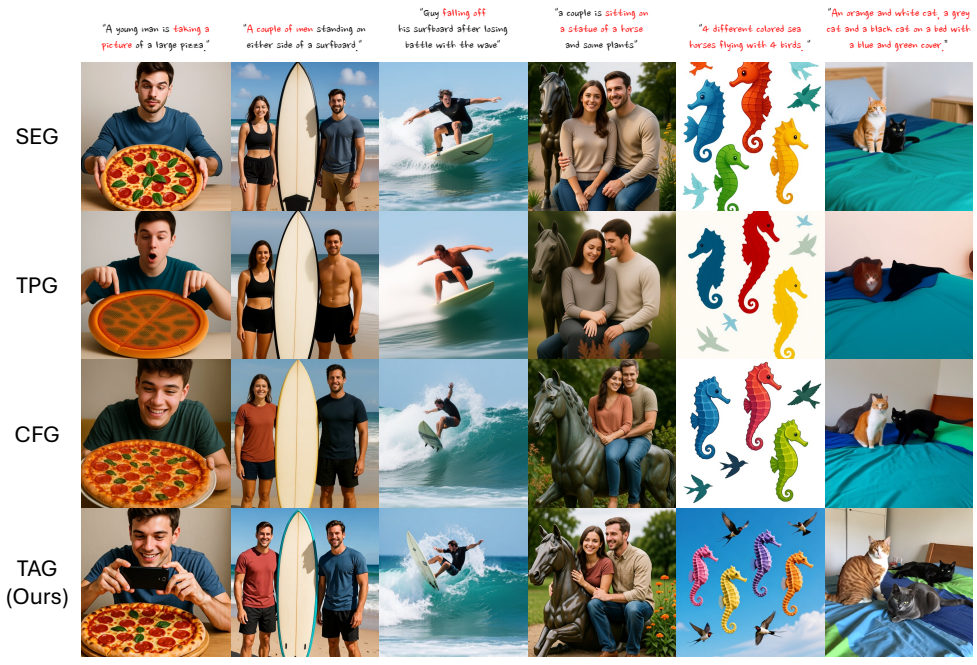


Figure 8: Qualitative comparison of guidance methods on Qwen-Image (Wu et al., 2025) under identical prompts and random seeds. Red words indicate parts of the prompt that baselines fail to reflect.

methods in terms of VQAScore by large margins, second only to CFG. This indicates that perturbing text attention generalizes across MM-DiT models, while the perturbations designed for U-Net do not. We also show in Figs. 7 and 8 qualitative comparisons of guidance methods on SD3.5 and Qwen-Image, respectively. As in Figs. 1 and 5, training-free baselines produce blurry or oversat-

urated images and miss the parts of the prompt highlighted in red, while TAG reflects the objects, attributes, and their relations in the prompt with realistic details.