INTERNAL REFERENCE GUIDANCE VIA LATENT TOKEN MERGING FOR DIFFUSION SAMPLING

Anonymous authors
Paper under double-blind review

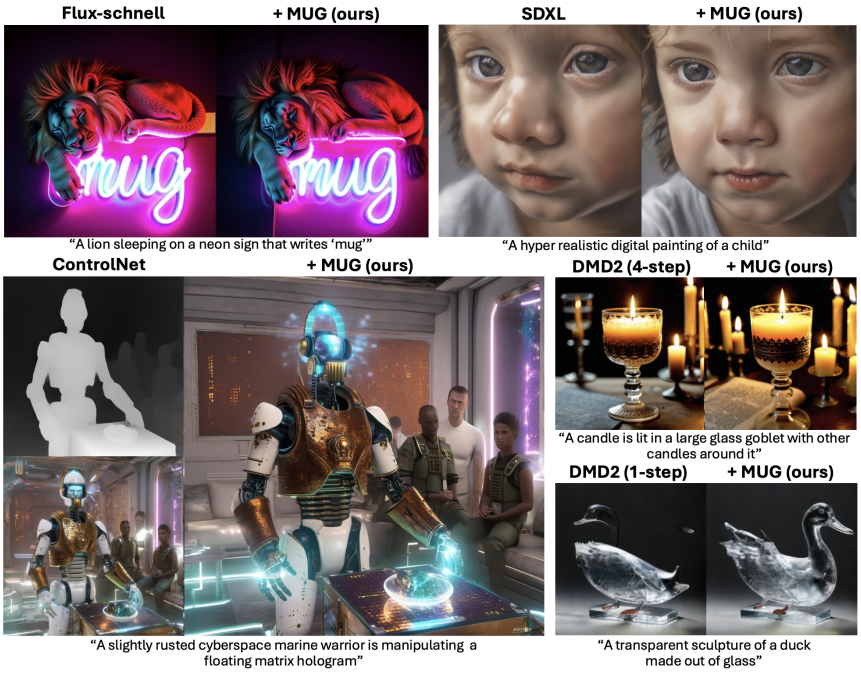


Figure 1: **Merging and Unmerging Guidance (MUG)** substantially enhances image generation quality across a spectrum of backbones, tasks, and regimes.

ABSTRACT

Inference-time diffusion sampling guidance plays a key role in enabling high-performance image synthesis by diffusion models. Existing methods, including Classifier-Free Guidance (CFG), typically maintain two denoising branches throughout the iterative denoising process to guide predictions by extrapolating between the two branches, which amounts to increased overhead in terms of the number of function evaluations (NFE). In this paper, we investigate whether effective guidance signals can be constructed within a single denoising branch. We achieve this by shifting guidance extrapolation from the prediction space to the intermediate latent-token feature space within the diffusion backbone. Specifically, we construct an internal reference within the cross-attention blocks by merging and unmerging latent tokens and perform guidance by extrapolating the original attention features away from this coarsened reference. Extensive experiments on diverse datasets show that the proposed Merging and Unmerging Guidance (MUG) serves as a general-purpose plug-in module that improves both U-Net- (*i.e.*, SDXL) and MMDiT-based (*i.e.*, Flux.1-schnell/dev) diffusion models in image fidelity, prompt alignment, and human preference. Notably, MUG fills an important applicability gap by enabling inference-time guidance for natively single-branch models, particularly guidance-distilled models (*e.g.*, DMD2), for which conventional CFG is not inherently supported. MUG generalizes across different tasks, sampling regimes, and guidance paradigms while complementing existing guidance methods from CFG to recent condition-free techniques.

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.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021; Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2020; Rombach et al., 2022) have been the driving backbone behind state-of-the-art generative applications. They generate high-fidelity images from pure noise through an iterative denoising procedure (Ho et al., 2020; Song et al., 2021; 2020). Early diffusion models (Blattmann et al., 2023; Rombach et al., 2022; Podell et al., 2023) adopt a U-Net-based (Ronneberger et al., 2015) backbone, whereas more recent ones (Black Forest Labs; Esser et al., 2024) employ multi-modal diffusion transformer (MMDiT) (Esser et al., 2024) architecture. They have demonstrated remarkable generation capability in different tasks, including image (Epstein et al., 2023; Podell et al., 2023; Saharia et al., 2022; Guo et al., 2024; Chen et al., 2024; Wu et al., 2025; Black Forest Labs), video (Chen et al., 2023; Singer et al., 2023; Wan et al., 2025; Zheng et al., 2024; Zhou et al., 2022), and 3D content (Poole et al., 2022; Wang et al., 2023; Lin et al., 2023) synthesis.

Behind the success of popular diffusion models, sampling guidance techniques (Dhariwal & Nichol, 2021; Ho & Salimans, 2021; Nichol et al., 2022) play a vital role in steering the sampling toward distributions that are aligned with higher fidelity and better semantics. At each denoising step, the guidance is achieved by guiding the prediction away from a negative reference signal. A prevailing approach is Classifier-Free Guidance (CFG) (Ho & Salimans, 2021), which extrapolates between conditional and unconditional predictions to steer the sampling away from the latter. It has been the standard configuration in many diffusion models (Rombach et al., 2022; Podell et al., 2023; Singer et al., 2023; Blattmann et al., 2023; Saharia et al., 2022; Guo et al., 2024).

While effective, CFG tends to produce oversaturated images or artifacts, especially at higher guidance scale, which has been mitigated by some tailored optimizations (Sadat et al., 2024; Chung et al., 2025; Xia et al., 2025). CFG's reliance on external conditions to construct guidance also limits its applicability in tasks not involving prompts, such as solving inverse problems (Chung et al., 2023; Rout et al., 2024; Zhang et al., 2025). Some condition-free guidance methods (Karras et al., 2024; Hong et al., 2023; Hong, 2024; Ahn et al., 2024) have been developed to fill this gap by manually degrading the prediction to create the negative signal.

A fundamental downside of these guidance methods, including CFG, is that they typically maintain *two separate branches* throughout the iterative denoising process and extrapolate between two predictions at each sampling step, thereby increasing inference overhead in terms of the number of function evaluations (NFE). While recent research (Yin et al., 2024; Lin et al., 2024; Ren et al., 2024) distills the effect of guidance into single-branch models, distillation itself is procedurally complicated and resource-demanding, often accompanied by degraded performance. Conventional CFG is not directly applicable to these models without reintroducing an auxiliary denoising branch, leaving an important applicability gap for inference-time guidance.

In this paper, we investigate whether effective guidance signals can instead be constructed within a single denoising branch. Concretely, we shift guidance extrapolation from the model-output space to the intermediate latent-token feature space within the diffusion backbone. Within cross-attention blocks, we propose to innovatively use token merging (Bolya et al., 2023) to degrade the raw latent-token features, followed by unmerging to restore token count to maintain generation resolution, thereby constructing an internal reference aligned with the original attention features. We then perform guidance by extrapolating the original attention features away from this coarsened reference. This formulation avoids the need for an auxiliary denoising branch to provide the negative reference signal. We dub our method "Merging and Unmerging Guidance" (MUG).

Concurrent to our work, PLADIS (Kim & Sim, 2025) also explores diffusion sampling guidance without an auxiliary denoising branch. However, we find that its performance gains are limited, even negative under certain setups, and the method is sensitive to the guidance scale. MUG operates with a fundamentally different and essentially opposite guidance mechanism. While PLADIS seeks a supposedly better prediction hinged on sparse attention weights and *guides toward* it, we oppositely craft a certainly worse prediction using token merging and *steer away* from it. We present a head-to-head comparison between MUG and PLADIS, demonstrating that MUG achieves higher generation performance, greater stability, and better efficiency.

We conduct extensive experiments to validate the effectiveness of MUG. Across multiple *datasets* (*i.e.*, MS-COCO (Lin et al., 2014), DrawBench (Saharia et al., 2022), HPDv2 (Wu et al., 2023),

Pick-a-Pic (Kirstain et al., 2023)) and *evaluation metrics* (CLIP Score (Radford et al., 2021), ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), Human Preference Study), MUG substantially improves the baselines and consistently outperforms PLADIS. It maintains its advantage across U-Net- (*i.e.*, SDXL (Podell et al., 2023)) and MMDiT-based (*i.e.*, Flux.1-schnell/dev (Black Forest Labs)) *architectures*, a wide spectrum of *sampling regimes* (from 50 steps to 1 step), and additional *downstream tasks* such as depth-guided generation with ControlNet (Zhang et al., 2023). More importantly, MUG fills a critical practical gap by extending inference-time guidance to natively single-branch models (Yin et al., 2024), for which conventional CFG is not directly applicable We expect that MUG could inspire future research on single-branch sampling guidance as an important baseline in this direction.

2 Related Work

Diffusion models are advanced generative models capable of producing high-fidelity images and videos. They build on an iterative denoising process that gradually produces an image from pure noise. Diffusion models are first developed with a U-Net (Ronneberger et al., 2015) backbone. They comprise stacked transformer blocks that process tokens through self-attention (SA), cross-attention (CA), and the feedforward (FF) network. Examples are Stable Diffusion 1.5 (SD1.5) (Rombach et al., 2022), Stable Diffusion XL (SDXL) (Podell et al., 2023), and other variants (Saharia et al., 2022; Guo et al., 2024; Blattmann et al., 2023; Chen et al., 2023; Singer et al., 2023; Zhou et al., 2022). More recent designs embrace Multi-modal diffusion transformer (MMDiT) (Esser et al., 2024) consisting of DiT (Peebles & Xie, 2023) blocks as the diffusion backbone. The DiT blocks adopt a flat and hybrid architecture that jointly processes image and text tokens via pure attention. MMDiT-based models have delivered state-of-the-art performance in many tasks, including image (Chen et al., 2024; Esser et al., 2024; Black Forest Labs; Wu et al., 2025) and video synthesis (Wan et al., 2025; Zheng et al., 2024). Some popular open-source MMDiT-based models include the Flux series (Black Forest Labs), Qwen-Image (Wu et al., 2025), and Wan (Wan et al., 2025).

Diffusion sampling guidance is a family of training-free techniques that steers the sampling of the diffusion model toward distributions of better semantics and higher fidelity. A widely adopted example is Classifier-Free Guidance (CFG) (Ho & Salimans, 2021), which extrapolates between conditional and unconditional predictions to steer the sampling away from the latter:

$$\epsilon_{\theta,\text{CFG}}(x_t, c) = \epsilon_\theta(x_t, c) + \eta\big(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t)\big), \tag{1}$$

where η is the guidance scale and c is the prompt condition. CFG has been a de facto configuration in a bunch of established diffusion models (Rombach et al., 2022; Podell et al., 2023; Singer et al., 2023; Blattmann et al., 2023; Saharia et al., 2022; Guo et al., 2024). Nonetheless, CFG is prone to oversaturation and degradation, especially for higher guidance scale. Efforts have been made to tackle this, including APG (Sadat et al., 2024), ReCFG (Xia et al., 2025), and CFG++ (Chung et al., 2025). The other limitation of CFG is its reliance on prompts to construct guidance. AutoGuidance (Karras et al., 2024) bypasses this reliance by introducing an auxiliary model. Other methods (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025; Ifriqi et al., 2025; Sadat et al., 2025; Zhang et al., 2026) apply manual perturbation to one of the forward paths to mimic the effect of CFG's unconditional branch, whereas NAG (Chen et al., 2025) gets away from the attention features of specified negative prompts. Yet, all these methods including CFG demand two forward branches, which implies doubled function evaluations. Unlike them, we propose a guidance method devoid of a second branch.

Token merging (Bolya et al., 2023) is a technique to speed up vision transformers by reducing the number of tokens to be processed. Specifically, it partitions all tokens into a source (*src*) set and a destination (*dst*) set. Then, for tokens in *dst*, the k most similar tokens in *src* are selected and merged with their corresponding *dst* tokens. This reduces token count by k, thereby making subsequent token processing faster. Token merging has proven successful in different applications such as video understanding (Li et al., 2024; Wu et al., 2024; Lee et al., 2024; Shen et al., 2025) and image generation (Bolya & Hoffman, 2023; Kim et al., 2024; Lu et al., 2025; Wang et al., 2024; Fang et al., 2025) for acceleration and compression. Nonetheless, its implication for training-free diffusion sampling guidance remains unexplored.

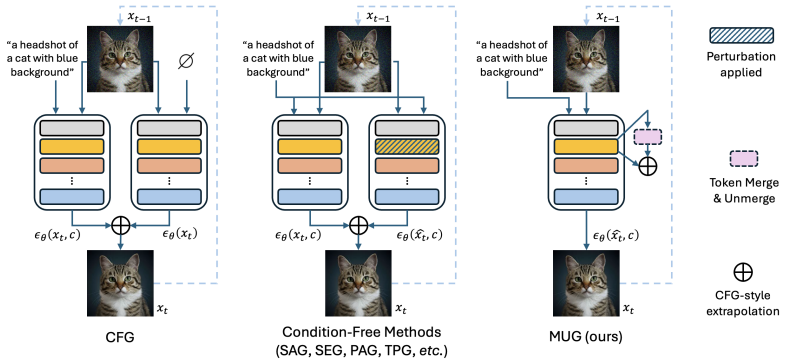


Figure 2: **A schematic comparison of Merging and Unmerging Guidance (MUG) and existing diffusion sampling guidance methods.**

3 METHOD

In this section, we provide a detailed description of our proposed method, Merging and Unmerging Guidance (MUG), for training-free diffusion sampling guidance with a single forward pass.

3.1 OVERVIEW

Existing guidance methods, including CFG (Ho & Salimans, 2021) and condition-free methods (Hong et al., 2023; Hong, 2024; Ahn et al., 2024; Rajabi et al., 2025), compute guidance by extrapolating between the predictions of two separate forward prediction paths. A schematic is provided in Figure 2. The core idea of MUG is to shift this guidance operation from noise prediction space to latent space within the backbone and obtain a "bad latent" via manual degradation. We introduce an ad-hoc token merging and unmerging block to extract a degraded version of the corresponding original latent. We then immediately fuse the two latents just as other guidance methods extrapolate in the noise space. Hence, the fused latent is a "guided" version of the original latent, and then proceeds to the subsequent layers in the usual way. This paradigm is illustrated to the right of Figure 2. In this way, we do not need to keep a separate forward path that would have to traverse all layers. Next, we provide more details on how MUG operates.

3.2 IMAGE LATENT MERGING

The token merging and unmerging module is inserted in parallel to each cross-attention block of the backbone. It involves token merging and unmerging operations to degrade the latent while keeping its dimensionality. To begin with, the unaltered latent produced by cross-attention is computed by:

$$\mathbf{Z} = \text{Softmax}\left(\mathbf{Q}_I(\mathbf{K}_T)^\top/\sqrt{d}\right)\mathbf{V}_T, \tag{2}$$

where query $\mathbf{Q}_I$ comes from the current backbone features, while key $\mathbf{K}_T$ and value $\mathbf{V}_T$ are external conditioning (text) features. Now, we are interested in getting a "degraded" version of $\mathbf{Z}$ in which token features are dispersed and blended, using token merging.

Since our goal is not to remove redundant tokens for model acceleration, we use a simplified version of the original token merging (Bolya et al., 2023). Specifically, instead of merging the most similar tokens, we simply merge random pairs of tokens. This means inter-token similarity and bipartite soft matching (Bolya et al., 2023; Bolya & Hoffman, 2023) are not needed, which makes our implementation much faster.

Inspired by Bolya & Hoffman (2023), we design a window-based src–dst split rule. Specifically, we randomly sample one token from each 2×2 spatial window as dst. This helps make dst tokens more evenly located and representative. We use average merging: For a pair of src token x_i and dst token x_j, x_i is merged into x_j with $h_j^m = (h_i + h_j)/2$ as the updated h_j, and h_i eliminated. The token merging process is expressed as $\mathbf{H}^m = \text{Merge}(\mathbf{H}, r)$, where $\mathbf{H}$ and $\mathbf{H}^m$ are the original latent before cross-attention and its merged version, respectively, and r is the merge ratio that controls how many src tokens are eliminated. The merged image query $\mathbf{Q}_I^m$ is then obtained via pre-trained linear

Algorithm 1: **Pseudo-code for Merging and Unmerging Guidance (MUG).**

Latent Rectification with MUG

Require: Latent before CA $\mathbf{H}$, original latent after CA $\mathbf{Z}$ and original key $\mathbf{K}_T$, value $\mathbf{V}_T$, MUG scale η, merge ratio r, set of all cross-attention. $\mathcal{S}_{\text{CA}}$

if Attn $\in \mathcal{S}_{\text{CA}}$ **then**
  $\mathbf{H}^m \leftarrow \text{Merge}(\mathbf{H}, r)$
  $\mathbf{Q}_I^m \leftarrow \text{ProjToQ}(\mathbf{H}^m)$
  $\mathbf{Z}^m \leftarrow \text{Attn}(\mathbf{Q}_I^m, \mathbf{K}_T, \mathbf{V}_T)$
  $\mathbf{Z}^{'} \leftarrow \text{Unmerge}(\mathbf{Z}^m)$
  $\mathbf{Z}^{\text{MUG}} \leftarrow \mathbf{Z} + \eta \cdot (\mathbf{Z} - \mathbf{Z}^{'})$
  $\mathbf{Z} \leftarrow \text{Normalize}(\mathbf{Z}^{\text{MUG}})$

Diffusion Sampling with MUG

Require: Diffusion model Model$(\cdot)$ consisting of L cross-attention layers Attn$(\cdot)$, total number of timesteps T.

Sample $x_T \sim \mathcal{N}(\mathbf{0}, I)$
for $l = 1, \ldots, L$ **do**
  Replace Attn$(\cdot)$ with Attn$_{\text{MUG}}(\cdot)$
for $t = T-1, \ldots, 1$ **do**
  $\epsilon_t \leftarrow \text{Model}(x_t)$
  $x_{t-1} \leftarrow \text{SolverStep}(x_t, \epsilon_t)$
Return x_0

projection $\mathbf{W}_Q$, *i.e.*, $\mathbf{Q}_I^m = \mathbf{H}^m \mathbf{W}_Q$, (denoted by ProjToQ$(\cdot)$ in Algorithm 1). With $\mathbf{Q}_I^m$, we obtain the shrunk latent as:

$$\mathbf{Z}^m = \text{Softmax}\left(\mathbf{Q}_I^m (\mathbf{K}_T)^\top \sqrt{d}\right) \mathbf{V}_T. \tag{3}$$

3.3 Cross-Modal Latent Unmerging

Latent $\mathbf{Z}^m$ has a reduced dimensionality. For generation tasks, the number of tokens must be maintained to finally produce a high-fidelity image. We employ a simple unmerging operation to restore the token count. Specifically, if token h_i has been merged into h_j as $h_j^m = (h_i + h_j)/2$, we simply unmerge them by putting their mean value back at both their original positions. Note that through our proposed token merging and unmerging, we achieve the effect of token feature mixing and spatial shuffle. With token unmerging, the features are restored to their original dimensionality. This process is expressed as $\mathbf{Z}^{'} = \text{Unmerge}(\mathbf{Z}^m)$. It is noteworthy that all parameters involved are weights of the pre-trained diffusion model and MUG does not introduce any trainable parameters.

3.4 MUG: Merging and Unmerging Guidance

With both $\mathbf{Z}$ and $\mathbf{Z}^{'}$ computed, we apply a CFG-style extrapolation between them to obtain the rectified latent. Mathematically,

$$\mathbf{Z}^{\text{MUG}} = \mathbf{Z} + \eta \cdot (\mathbf{Z} - \mathbf{Z}^{'}), \tag{4}$$

where η is a scalar guidance scale that controls the strength of guidance applied at a step. The resultant latent $\mathbf{Z}^{\text{MUG}}$ has the same dimensionality as the original $\mathbf{Z}$ and proceeds through the subsequent modules as usual. Since we are directly manipulating the latent representation of the single forward branch, we apply representation renormalization to $\mathbf{Z}^{\text{MUG}}$ to mitigate accumulated distribution shift throughout the iterative sampling process. In practice, we adopt the simple technique introduced in Chen et al. (2025), which involves norm ratio rescaling and linear blending. The final latent $\mathbf{Z}^{\text{MUG}}$ is forwarded to the subsequent modules. More details are provided in the appendix.

The above MUG guidance process, described by the pseudo-code presented to the left of Algorithm 1, is executed at every cross-attention module and at each timestep, as the inference proceeds. A pseudo-code for guided diffusion sampling with MUG is provided to the right of Algorithm 1.

4 Experiments

4.1 Experimental Details

Models. We deploy the proposed method on different diffusion architectures and backbones to validate its effectiveness and universality. Specifically, for U-Net (Ronneberger et al., 2015)-based backbones, we choose Stable Diffusion XL (SDXL) (Podell et al., 2023); for more advanced MM-DiT (Esser et al., 2024) backbones, we choose Flux.1-schnell and Flux.1-dev (Black Forest Labs). Meanwhile, we also experiment with a guidance-distilled model, DMD2 (Yin et al., 2024), using both its official 1-step and 4-step weights. We further evaluate our method on the structure-guided image synthesis task using ControlNet (Zhang et al., 2023).

Table 1: **Quantitative comparison on SDXL.**

Method	#Steps	NFE	CLIP↑	ImageReward↑	PickScore↑
Vanilla	50	50	0.280	-0.870	20.17
Vanilla+PLADIS	50	50	0.285	-0.756	20.19
Vanilla+MUG (ours)	50	50	**0.288**	**-0.573**	**20.43**
CFG	50	100	0.327	0.635	22.29
CFG+PLADIS	50	100	0.332	0.796	22.24
CFG+MUG (ours)	50	100	**0.333**	**0.890**	**22.47**

Table 2: **Quantitative comparison on Flux-schnell.**

Method	#Steps	NFE	CLIP↑	ImageReward↑	PickScore↑
Vanilla	4	4	0.329	1.036	22.66
Vanilla+PLADIS	4	4	0.331	1.040	22.65
Vanilla+MUG (ours)	4	4	**0.333**	**1.108**	**22.85**

Datasets and metrics. We used the validation sets of MS-COCO 2014/MS-COCO 2017 (Lin et al., 2014) as our evaluation data. MS-COCO 2014 *val* contains 40,504 images and MS-COCO 2017 *val* contains 5,000 higher-quality images sampled from MS-COCO 2014 *val*. Both datasets come with text annotations. We also use a group of human preference datasets, including DrawBench (Saharia et al., 2022), HPDv2 (Wu et al., 2023), and Pick-a-Pic (Kirstain et al., 2023). We employ CLIP Score (Radford et al., 2021) for prompt alignment evaluation, and ImageReward (Xu et al., 2023) and PickScore (Kirstain et al., 2023) as human preference assessment.

Implementation. Our method is implemented based on the `diffusers` library. We use the default configurations for all diffusion models and sampling guidance methods used, unless otherwise stated. For SDXL, we run 50 inference steps by default. For Flux-schnell, Flux-dev, and DMD2, we run 4, 50, and 4 steps, respectively, with CFG disabled all the time. Experiments are conducted on four NVIDIA Tesla V100 GPUs.

4.2 Main Results

Generation with SDXL. We evaluate our method on SDXL under the standard 50-step sampling configurations and report the results in Table 1. MUG substantially improves the quality of images generated by the pre-trained SDXL across CLIP Score, ImageReward, and PickScore, while also consistently outperforming PLADIS. Notably, while PLADIS struggles to improve PickScore and even degrades it, MUG achieves significant gains of about 0.2 points. With CFG, PLADIS yields a $+0.16$ gain on ImageReward, whereas MUG achieves a much larger gain of $+0.26$. These results validate the effectiveness of MUG in boosting diffusion sampling at inference. Qualitative results are presented in Figures 1 and the appendix, further corroborating the quantitative evidence.

Generation with Flux. We present the results of applying MUG on Flux-schnell under its default 4-step setup in Table 2. Guiding Flux using a single branch is challenging, further exacerbated by the extremely low-step regime and its already decent quality. Nonetheless, we find that MUG leads to improvements across all three metrics and beats PLADIS overall, whereas the latter slightly degrades in terms of PickScore. Visually, generated images in Figure 1 show that MUG produces more natural and realistic details and geometries that are instructed by the prompt (*e.g.*, the lion's tail and foot). Additional results on Flux-dev are provided in Appendix B.

Results on human preference datasets. Quantitative evaluations on the DrawBench, HPDv2, and Pick-a-Pic datasets are presented in Table 3. The results show that MUG always brings substantial gains when applied to a CFG-powered baseline in terms of image quality, aesthetics, and prompt adherence. MUG also consistently outperforms the pioneering method of PLADIS. Results show that MUG's performance edge persists under this commonly adopted few-step scheme.

Structure-guided generation with ControlNet. To demonstrate the applicability of MUG on further downstream tasks, we apply it to structure-guided generation with ControlNet and use a depth map as input to specify the layout. Examples in Figures 1 and 3 show that MUG significantly enhances prompt alignment and image fidelity. In contrast, PLADIS tends to produce unnatural details (*i.e.*, see the Mercedes logo, the windshield, and the ground in the upper images).

Compatibility with guidance-distilled models. Existing guidance methods (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025; Karras et al., 2024) are known to be incompatible with

Table 3: **Quantitative comparison on SDXL using DrawBench, HPDv2, and Pick-a-Pic datasets.**

Dataset	Method	#Steps	CLIP↑	ImageReward↑	PickScore↑
DrawBench (Saharia et al., 2022)	CFG	50	0.325	0.472	22.16
	CFG + PLADIS	50	**0.334**	0.608	22.20
	CFG + MUG	50	**0.334**	**0.689**	**22.39**
HPDv2 (Wu et al., 2023)	CFG	50	0.344	0.764	22.40
	CFG + PLADIS	50	0.353	0.885	22.42
	CFG + MUG	50	**0.354**	**1.000**	**22.48**
Pick-a-Pic (Kirstain et al., 2023)	CFG	50	0.337	0.709	22.41
	CFG + PLADIS	50	0.346	0.913	22.50
	CFG + MUG	50	**0.347**	**1.061**	**22.58**



Figure 3: **Qualitative effect of MUG on ControlNet.**

Table 4: **Quantitative comparison on DMD2.**

Method	#Steps	CLIP↑	ImageReward↑	PickScore↑
Vanilla	4	0.332	0.911	22.76
PLADIS	4	**0.334**	0.962	22.78
MUG (ours)	4	**0.334**	**1.084**	**22.86**
Vanilla	1	0.333	0.822	22.27
PLADIS	1	0.335	0.872	22.22
MUG (ours)	1	**0.336**	**1.004**	**22.31**

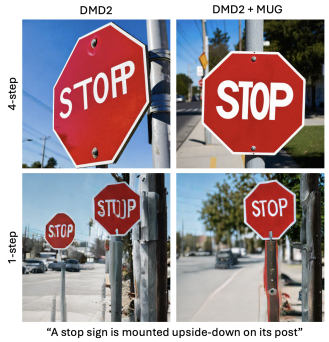


Figure 4: **Qualitative effect of MUG on DMD2.**

guidance-distilled models that maintain a single prediction branch with CFG already baked in. In Table 4 and Figures 1 and 4, we demonstrate using DMD2 that our single-branch MUG guidance can be seamlessly integrated with a guidance-distilled model without introducing a second branch. MUG consistently improves the DMD2 baseline under both 4-step and 1-step schemes. Notably, MUG maintains its effectiveness under the challenging 1-step generation setup. Whereas PLADIS gives no gain in PickScore and a limited gain of $+0.05$ in ImageReward, MUG improves both metrics, with a notable $+0.18$ in ImageReward. Therefore, MUG is not only compatible with guidance-distilled models but also excels under extremely low-step generation settings.

Compatibility with existing guidance methods. The proposed MUG guidance can be conveniently integrated with any type of existing guidance methods, including condition-based CFG (Ho & Salimans, 2021) and condition-free guidance such as PAG (Ahn et al., 2024) and SEG (Hong, 2024). As demonstrated by Table 5, MUG dramatically improves upon existing guidance and surpasses PLADIS by significant margins while not introducing extra NFE. Notably, MUG also integrates well with multiple guidance methods used altogether (*e.g.*, MUG+CFG+PAG and MUG+CFG+SEG) to yield further improvements.

Human preference study. To complement metric-based evaluations, we conduct a human preference study asking participants to judge whether aspects such as image quality and prompt alignment have been improved by our method. Details are provided in the appendix. From the visualized response in Figure 5, images generated using the MUG guidance are more frequently favored in terms of both image quality and prompt alignment, which further validates its effectiveness.

4.3 Further Analysis

Sensitivity to hyperparameters. We begin by understanding how sensitive MUG is to its two hyperparameters, namely guidance scale η and merge ratio r. As evident from the SDXL examples in Figure 6, thanks to its design, MUG produces stable and realistic image quality over a wide range of hyperparameters. We notice that even a very small η or r yields tangibly improved visual quality, which tends to saturate at higher values. In the appendix, we demonstrate the effect of applying renormalization in stabilizing generation performance over a broad range of guidance scale values.

Ablations on guidance policy. When combined with existing dual-branch guidance methods (Ho & Salimans, 2021; Hong et al., 2023; Hong, 2024; Ahn et al., 2024), the default design is to apply MUG on each of the two branches separately. For example, when applying MUG on top of CFG, the

Table 5: **Compatibility with existing guidance methods.**

CFG	Method	#Steps	NFE	CLIP↑	ImageReward↑	PickScore↑
✗	PAG (Ahn et al., 2024)	50	100	0.306	-0.0145	21.54
	PAG+PLADIS	50	100	0.314	0.291	**21.78**
	PAG+MUG (ours)	50	100	**0.315**	**0.344**	21.72
	SEG (Hong, 2024)	50	100	0.301	-0.0575	21.35
	SEG+PLADIS	50	100	0.309	0.233	21.53
	SEG+MUG (ours)	50	100	**0.310**	**0.318**	**21.64**
✓	PAG (Ahn et al., 2024)	50	150	0.324	0.585	22.39
	PAG+PLADIS	50	150	**0.330**	0.828	22.51
	PAG+MUG (ours)	50	150	**0.330**	**0.881**	**22.59**
	SEG (Hong, 2024)	50	150	0.324	0.597	22.20
	SEG+PLADIS	50	150	**0.331**	0.815	22.29
	SEG+MUG (ours)	50	150	**0.331**	**0.889**	**22.40**

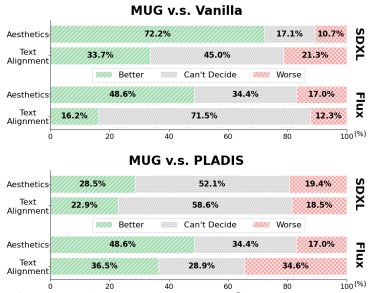


Figure 5: **Responses from the human preference study.**



Figure 6: **Effect of varying hyperparameter values on SDXL.**

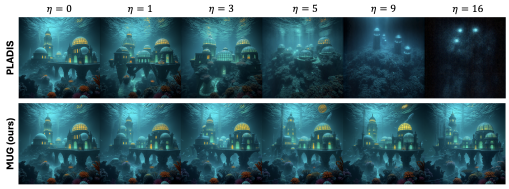


Figure 7: **Robustness to varying guidance scale values on Flux-schnell.**

conditional and unconditional predictions of CFG will each get rectified by MUG at each timestep. We are interested in determining whether this default design is optimal, and whether an asymmetric application of MUG could potentially lead to better performance. To this end, we experiment with applying MUG to only either one of CFG's two branches but not together. Results in Table 6 indicate that applying MUG to both branches yields the best performance and applying it only to the conditional branch slightly degrades performance but only marginally. Not guiding the conditional branch leads to considerable degradation.

Generation stability on Flux. Flux is known to be susceptible to collapse when its feature distributions are altered (Chen et al., 2025). We find that our guidance method is surprisingly robust on Flux, even though we have aggressively modified the latent. According to Figure 7, MUG demonstrates stable generation on Flux-schnell even under very high guidance scales, whereas PLADIS begins to collapse. This observation suggests that steering the generation away from an intentionally degraded latent representation is more effective than using sparse attention weights as a supposedly "salient" direction and guiding the model toward it. We hypothesize that the latter approach is more susceptible to out-of-distribution behavior at the layer level, leading to instability during generation.

MUG on text embeddings. We ablate the key choice of obtaining the query from merged image features. Specifically, we merge the contextualized text tokens and use them to obtain the degraded keys and values. This variant does not require token unmerging because the cross-attention output retains the number of the original image queries. However, as noticed in Table 7, merging text tokens leads to substantially inferior performance, with samples visualized in Figure 12. This is because merging text keys and values perturbs the semantic basis of cross-attention, as the compact 77-token representation already encodes distinct concepts. The resulting residual reflects semantic redistribution rather than spatial coarsening, entangling the guidance direction with shifts in the conditioning signal. The empirical results underscore the importance of constructing the degraded representation and applying guidance in the image feature space.

Efficiency analysis. We evaluate the end-to-end latency T_{E2E} and incremental GPU memory cost Δ_{Mem} of MUG. In Table 8, we report the measurements on Flux-schnell, a guidance-distilled model whose native sampling process does not require a separate classifier-free guidance branch. The better results between PLADIS and MUG are highlighted in bold. As shown, MUG consistently improves all reported quality metrics over the vanilla model and PLADIS. It is also 20% faster and requires significantly less incremental memory cost than PLADIS, while incurring only a modest latency increase over the vanilla model. These results demonstrate that MUG provides a favorable performance–efficiency trade-off for native single-branch models without introducing additional NFEs, and is a more effective and efficient alternative to PLADIS. A component-level runtime analysis and further efficiency results on SDXL are provided in Appendix C.

Table 6: **Effect of MUG applied to different CFG branches.**

Guided Branch	CLIP↑	ImageReward↑	PickScore↑
cond	**0.333**	0.887	**22.47**
uncond	0.327	0.635	22.29
cond + uncond	**0.333**	**0.890**	**22.47**

Table 7: **Effect of merging different embeddings.**

Modality	Config.	Ratio	CLIP↑	ImageReward↑	PickScore↑
Image	$[Q_m, K, V]$	0.75	**0.333**	**0.890**	**22.47**
Text	$[Q, K_m, V_m]$	0.75	0.330	0.771	22.34
		0.50	0.330	0.727	22.68
		0.25	0.328	0.631	22.15

Table 8: **Quantitative comparison of performance and efficiency on Flux-schnell.**

Method	NFE	CLIP↑	ImageReward↑	PickScore↑	$T_{\mathbf{E2E}}$ (s/img)↓	$\Delta_{\mathbf{Mem}}$ (MB)↓
Vanilla	4	0.329	1.036	22.66	**2.204**	**265.21**
PLADIS	4	0.331	1.040	22.65	3.041	382.99
MUG	4	**0.333**	**1.108**	**22.85**	2.440	362.84
CFG	N/A	N/A	N/A	N/A	N/A	N/A

Further discussion. We emphasize that MUG is not positioned as a replacement for dual-branch CFG, even though MUG is faster and more memory-efficient than it. Importantly, **MUG addresses an important applicability gap for native single-branch models**, including guidance-distilled models like Flux-schnell, Flux-dev, and DMD2, for which conventional dual-branch guidance is not naturally supported. Our results show that MUG consistently improves these native baselines without introducing an additional denoising branch. On models where CFG is readily available, as in SDXL, MUG instead serves as a complementary module and can further improve generation quality when combined with CFG.

Among existing approaches, PLADIS is the most directly comparable one because it shares the same training-free and single-branch setting. Whereas PLADIS constructs a "better" sampling direction through computationally expensive sparse attention, MUG uses simple token merging and unmerging to get a "worse" prediction. **MUG outperforms PLADIS while offering a more favorable latency and memory trade-off.** Together, these properties establish MUG as an effective and efficient single-branch guidance method that also remains compatible with existing dual-branch guidance.

Additional results. We kindly refer readers to the appendix for additional experimental details, results and discussion.

Limitations and future work. The computations of MUG are entirely parasitic on the original forward path of a pre-trained diffusion model. Consequently, the guidance it produces is heavily influenced by the quality of intermediate latents during inference. When they are suboptimal, the resulting guidance correspondingly deteriorates. Moreover, since diffusion-based generation is inherently iterative, any deficiency in the early guidance may be progressively amplified throughout the denoising process. As a result, imperfect guidance can accumulate over steps, ultimately constraining the quality of generated images. For example, the quality gain brought about by MUG is limited when applied to SDXL without CFG, though not using MUG gives worse results. This is a common limitation for single-branch guidance methods and represents a meaningful topic for future research.

5 Conclusion

In this paper, we have introduced Merging and Unmerging Guidance (MUG), a training-free method for improving pre-trained diffusion models at inference time within a single denoising branch. At the core of its design, MUG shifts CFG-style extrapolation from the model-output space to the intermediate latent-token feature space within the diffusion backbone. Specifically, MUG constructs a coarsened internal reference within cross-attention blocks by merging latent tokens and unmerging the resulting attention features back to the original token resolution, and guides the original attention features away from this reference. This design eliminates the need for an auxiliary full-denoiser branch to provide a negative reference signal. Extensive experiments demonstrate the effectiveness and generality of MUG, with consistent improvements in image fidelity, human preference, and prompt alignment across different datasets, tasks, architectures, and sampling regimes. MUG fills a practical applicability gap by enabling inference-time guidance for natively single-branch models for which conventional CFG is not directly applicable. We believe that MUG provides a practical baseline for future research on single-branch diffusion sampling guidance.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar, clarity, and readability of portions of the manuscript, and to assist with code completion, debugging, and refactoring. We did not use generative AI tools to develop the main research ideas, formulate hypotheses or mathematical claims, design the methodology or experiments, implement the proposed method, analyze data, or interpret experimental results. Tasks involving proofs and qualitative analysis are not applicable to this work.

All AI-assisted text was reviewed and revised by the authors. All AI-assisted code was inspected and tested by the authors, and the reported experimental results were independently reproduced using the final codebase. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce the results of this work throughout the paper and the supplementary material. The proposed method and its implementation details are described in Sections 3-4 and Appendix A. These include the algorithm, inference procedures, hyperparameter settings, datasets, models, evaluation protocols, and computational resources. Pseudo-code for the implementation and integration of the proposed method is provided in Algorithm 1. The source code, datasets, and scripts for reproducing the experiments will be open-sourced.

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, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *CVPR*, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ICLR*, 2023.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized attention guidance: Universal negative guidance for diffusion model. In *NeurIPS*, 2025.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ICLR*, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *ICLR*, 2025.

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

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Haipeng Fang, Sheng Tang, Juan Cao, Enshuo Zhang, Fan Tang, and Tong-Yee Lee. Attend to not attended: Structure-then-detail token merging for post-training dit acceleration. In *CVPR*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

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

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.

Tariq Berrada Ifriqi, Adriana Romero-Soriano, Michal Drozdzal, Jakob Verbeek, and Karteek Alahari. Entropy rectifying guidance for diffusion and flow models. In *NeurIPS*, 2025.

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*, 2024.

Kwanyoung Kim and Byeongsu Sim. Pladis: Pushing the limits of attention in diffusion models at inference time by leveraging sparsity. In *ICCV*, 2025.

Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *WACV*, 2024.

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.

Seon-Ho Lee, Jue Wang, Zhikang Zhang, David Fan, and Xinyu Li. Video token merging for long-form video understanding. In *NeurIPS*, 2024.

Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *CVPR*, 2024.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 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.

Wenbo Lu, Shaoyi Zheng, Yuxuan Xia, and Shengjie Wang. Toma: Token merge with attention for diffusion models. In *ICML*, 2025.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.

Geon Yeong Park, Jeongsol Kim, Beomsu Kim, Sang Wan Lee, and Jong Chul Ye. Energy-based cross attention for bayesian context update in text-to-image diffusion models. In *NeurIPS*, 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. *arXiv preprint arXiv:2307.01952*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

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

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

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. In *NeurIPS*, 2024.

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.

Litu Rout, Yujia Chen, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Beyond first-order tweedie: Solving inverse problems using latent diffusion. In *CVPR*, 2024.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *ICLR*, 2024.

Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *ICLR*, 2025.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. Tempme: Video temporal token merging for efficient text-video retrieval. In *ICLR*, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *ICLR*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

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*, 2020.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Hongjie Wang, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K Jha, and Yuchen Liu. Attention-driven training-free efficiency enhancement of diffusion models. In *CVPR*, 2024.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023.

Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *CVPR*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 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.

Mengfei Xia, Nan Xue, Yujun Shen, Ran Yi, Tieliang Gong, and Yong-Jin Liu. Rectified diffusion guidance for conditional generation. In *CVPR*, 2025.

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.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *CVPR*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 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.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

APPENDIX

A ADDITIONAL DETAILS

A.1 DETAILS ON RENORMALIZATION

Inspired by Chen et al. (2025), we apply simple feature renormalization to the MUG-guided feature $\mathbf{Z}^{\text{MUG}}$ in Equation 4. Specifically, we perform channel-wise rescaling of $\mathbf{Z}^{\text{MUG}}$ as:

$$\hat{\mathbf{Z}}_i^{\text{MUG}} = \mathbf{Z}_i^{\text{MUG}} \cdot \hat{s}_i / s_i, \quad i \in \{1, \dots, D\}, \tag{5}$$

where D is the number of token channels; s_i is a channel-wise ratio computed by $s_i = ||\mathbf{Z}_i^{\text{MUG}}||/||\mathbf{Z}_i||$, and $\hat{s}_i$ is the clipped value of s_i. The final latent is simply the average of the rescaled latent $\hat{\mathbf{Z}}^{\text{MUG}}$ and the original latent $\mathbf{Z}$:

$$\mathbf{Z}^{\text{MUG}} = (\hat{\mathbf{Z}}^{\text{MUG}} + \mathbf{Z})/2. \tag{6}$$

A.2 DETAILS ON MAIN EXPERIMENTS

Datasets. For MS-COCO 2017 (Lin et al., 2014), we use all 5,000 images and associated prompts for experiments. For MS-COCO 2014 (Lin et al., 2014), we randomly sample 3,000 prompts. For human preference datasets, we use 200 prompts from DrawBench (Saharia et al., 2022), 800 from HPDv2 (Wu et al., 2023), and 200 from Pick-a-Pic (Kirstain et al., 2023). We generate 5 images per prompt for DrawBench, 2 images per prompt for HPDv2, and 4 images per prompt for Pick-a-Pic.

Implementation. In all experiments, we apply MUG to the cross-attention of all layers within the diffusion model. For SDXL (Podell et al., 2023) experiments, we use `stable-diffusion-xl-base-1.0` and 50 inference steps. CFG is not used by default. When used, the CFG guidance scale is set to 4 by default. For Flux (Black Forest Labs), we use `flux.1-schnell` and `flux.1-dev`. We use 4 and 50 inference steps, respectively, and do not use CFG. For DMD2 (Yin et al., 2024) experiments, we use `dmd2_sdxl_4step_lora_fp16` for the 4-step model, and `dmd2_sdxl_1step_unet_fp16` for the 1-step model. CFG is disabled. For ControlNet (Zhang et al., 2023), we use `controlnet-depth-sdxl-1.0` and 25 inference steps. CFG scale is set to 4 by default. For MUG's hyperparameters, we set $\eta = 9$ and $r = 0.75$ on SDXL and $\eta = 1.5$ and $r = 0.25$ on Flux models.

A.3 DETAILS ON DUAL-BRANCH IMPLEMENTATION OF MUG

For the dual-branch variant of MUG reported in Table 13, we apply our token merging and unmerging to the self-attention blocks, aligning with the designs in previous works (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025). Token merging is applied before the self-attention block (and its corresponding residual shortcut), followed by unmerging after the block to restore latent dimension. MUG is applied to the first 6 upsampling layers within the U-Net backbone of SDXL. The CFG-style extrapolation happens between the noise predictions of the two branches as in dual-branch methods. We apply a guidance scale of 3 and a merge ratio of 0.5. Renormalization is not used. All compared methods in Table 13 are implemented using their official code and reproduced using recommended parameters.

A.4 DETAILS ON HUMAN PREFERENCE STUDY

The study involves 14 human participants aged between 18 and 65. We used 15 different prompts to generate 15 images under each of the three generation configurations: 1) using MUG, 2) using PLADIS, and 3) using neither MUG nor PLADIS (*i.e.*, "Vanilla"), resulting in a total of 45 test images.

For each prompt, participants were asked to make two pairwise comparisons: the MUG-generated image versus the PLADIS image, and the MUG-generated image versus the vanilla image. For each comparison, participants were asked to determine which image was better in terms of (1) image quality and aesthetics and (2) prompt alignment, and choose their response from (A) the left image is better, (B) the right image is better, and (C) can't decide.

Table 9: **Quantitative comparison on Flux-dev.**

Method	#Steps	NFE	CLIP↑	ImageReward↑	PickScore↑
Vanilla	50	50	0.319	0.857	22.69
Vanilla+PLADIS	50	50	0.320	0.894	22.74
Vanilla+MUG (ours)	50	50	**0.322**	**0.920**	**22.84**

Table 10: **Quantitative results of 25-step generation on SDXL using COCO 2017.**

Method	#Steps	CLIP↑	ImageReward↑	PickScore↑
CFG	25	0.325	0.565	22.11
CFG+PLADIS	25	0.331	0.735	22.09
CFG+MUG (ours)	25	**0.332**	**0.853**	**22.32**

Finally, the responses were aggregated across all participants over all images. Results are reported in Figure 5 as the average percentage of preference for each generation configuration under each evaluation criterion.

B Additional Results

Results on Flux-dev. We conduct additional experiments on the higher-performance Flux.1-dev (Black Forest Labs) model. We sample 200 prompts from the COCO 2017 validation prompt set and run the model for 50 inference steps for evaluation. CFG is not used by default. The results presented in Table 9 show that MUG consistently outperforms both the vanilla baseline and PLADIS across all three metrics. These empirical results further corroborate MUG's generalization to more advanced and capable diffusion models.

Few-step performance. Using the 25-step setting in Kim & Sim (2025), we examine few-step generation performance using COCO 2017 prompts and human preference datasets and report the quantitative results in Tables 10 and 11, respectively. From Table 10, MUG consistently achieves the best scores under each metric. Its significant lead on ImageReward and PickScore implies its tendency to produce more natural and aesthetic images even with a halved sampling schedule. Further results on other human preference datasets in Table 11 corroborate this observation, whereby MUG consistently leads other methods by even slightly larger margins. Therefore, MUG discovers more ideal guidance directions and hence produces better images more quickly in a limited inference budget.

Generalization over CFG scales and variants. The CFG scale of 4 was selected as a generic moderate setting rather than as a carefully tuned optimum. This choice is also consistent with the setting in PLADIS. To empirically verify the generalization property of MUG across different CFG scales, we conduct evaluation using 1K prompts from the COCO 2017 validation set. We use the same default MUG scale in all cases without tuning it. In addition, we also evaluate MUG in combination with CFG++ (Chung et al., 2025), a stronger variant of CFG, at an empirically selected CFG++ scale of 0.6. As shown by the results in Table 12, MUG consistently improves performance across different CFG strengths and variants. Applying MUG often yields larger gains than simply increasing the CFG scale, which may meanwhile introduce artifacts and oversaturation. Its gains persist with the stronger CFG++ baseline. These results demonstrate MUG's generalizability across base guidance of varying strength.

MUG as dual-branch guidance. We postulate that the spirit of token merging as perturbation generalizes beyond a single-branch inference scheme. To demonstrate its broader implication, we integrate our token merging and unmerging guidance into the standard dual-branch configurations on SDXL as in previous works (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025). We then perform a head-to-head comparison of our dual-branch variant of MUG against these methods in Table 13, under the same setup on 3,000 captions sampled from COCO 2014. We also report Aesthetic Score (AES) (Schuhmann et al., 2022) following standard protocols. Surprisingly, MUG is able to outperform most existing methods specifically designed for the dual-branch guidance scheme. These results highlight the wider potential of the proposed token merging and unmerging guidance.

Table 11: **Quantitative results of 25-step generation on SDXL using DrawBench, HPDv2, and Pick-a-Pic datasets.**

Dataset	Method	#Steps	CLIP↑	ImageReward↑	PickScore↑
DrawBench (Saharia et al., 2022)	CFG	25	0.325	0.371	22.03
	CFG + PLADIS	25	**0.333**	0.522	22.05
	CFG + MUG	25	**0.333**	**0.668**	**22.25**
HPDv2 (Wu et al., 2023)	CFG	25	0.344	0.672	22.25
	CFG + PLADIS	25	0.352	0.809	22.30
	CFG + MUG	25	**0.354**	**0.932**	**22.36**
Pick-a-Pic (Kirstain et al., 2023)	CFG	25	0.337	0.638	22.28
	CFG + PLADIS	25	0.345	0.834	22.40
	CFG + MUG	25	**0.348**	**0.987**	**22.47**

Table 12: **Quantitative comparison of MUG under different CFG scales.**

CFG Scale	Method	CLIP↑	ImageReward↑	PickScore↑
3	CFG	0.323	0.442	22.03
	CFG + MUG	**0.330**	**0.765**	**22.23**
4	CFG	0.327	0.583	22.25
	CFG + MUG	**0.333**	**0.880**	**22.42**
6	CFG	0.330	0.714	22.44
	CFG + MUG	**0.336**	**0.965**	**22.57**
0.6	CFG++	0.333	0.775	22.43
	CFG++ + MUG	**0.339**	**1.005**	**22.53**

Note that in the single-branch scheme, we apply MUG to cross-attention because query-only merging-unmerging perturbs how image tokens retrieve from fixed text keys and values, while preserving the original text condition. In self-attention, it instead changes the attention weights between image tokens and therefore directly alters how each token aggregates spatial and contextual information from other image tokens. Under single-branch formulation, the resulting guided self-attention output is propagated through subsequent layers and denoising steps, which can amplify structural distortions. This differs from dual-branch self-attention guidance, where the clean prediction is computed along an independent, unmodified path and the perturbed branch is used only to construct an output-level guidance residual.

C Further Analysis

Sensitivity to guidance scale. We study the sensitivity of our method to its hyperparameters, using a set of ablation experiments. Figure 8 plots how the evaluation metrics vary with the guidance scale η. Our method produces the highest results across all evaluation metrics within the typical range of 1.5 to 3 – the same range in which PLADIS achieves its best performance. Further increasing the guidance scale leads to performance degradation, likely because the guided features drift outside the distribution expected by the pre-trained diffusion model. On the other hand, MUG remains relatively robust to η on the MMDiT-based Flux-schnell model, maintaining competitive performance even at much larger guidance scales.

Sensitivity to merge ratio. We also explore the effect of merge ratio r on the quality of images generated using the proposed guidance. Results in Table 14 clearly show that increasing the merge ratio brings about consistent performance gains. The reason is that the latent shall be sufficiently "perturbed" in order to serve as an effective negative signal, as discussed in Section 1. We therefore set $r = 0.75$ as the default configuration. Note that $r = 0.75$ corresponds to merging all *src* tokens into their respective *dst* tokens, where 1 *dst* token is sampled from each 2×2 token grid and 75% of all tokens are chosen as *src* tokens. Increasing merge ratio beyond 0.75 means that we must start to also discard *dst* tokens.

Effect of window-based *dst* sampling. Empirically, we find that using a window-based *dst* token selection policy is key to MUG's performance. In diffusion models, latent tokens correspond to spatial

Table 13: **Performance of MUG as a dual-branch guidance method on SDXL.**

Method	#Steps	NFE	CLIP↑	ImageReward↑	PickScore↑	AES↑
Vanilla	50	50	0.281	-0.834	20.22	5.676
SAG (Hong et al., 2023)	50	100	0.295	-0.480	20.66	5.757
SEG (Hong, 2024)	50	100	0.303	-0.0226	21.38	5.888
PAG (Ahn et al., 2024)	50	100	0.306	0.00260	21.57	5.828
TPG (Rajabi et al., 2025)	50	100	**0.313**	**0.225**	21.84	5.817
MUG (ours)	50	100	0.310	0.221	**22.01**	**5.946**

Table 14: **Effect of varying merge ratio.**

Merge ratio	CLIP↑	ImageReward↑	PickScore↑
0.25	0.329	0.711	22.23
0.50	0.331	0.797	22.28
0.75	**0.333**	**0.890**	**22.47**

Table 15: **Effect of different *dst* sampling policies.**

dst Sampling Policy	CLIP↑	ImageReward↑	PickScore↑
Random	0.327	0.633	22.29
Window	**0.333**	**0.890**	**22.47**

positions on a feature grid rather than arbitrary sequence elements (*e.g.*, the 4096-token sequence of SDXL maps to a 64×64 latent map). Selecting *dst* tokens completely at random can lead to clustering in certain regions and leave other areas underrepresented, resulting in uneven spatial coverage and unstable aggregation during merging. In contrast, ensuring that each $S_x \times S_y$ window contains one *dst* token guarantees a more uniform spatial distribution. This uniform coverage preserves representative anchors across the entire latent grid, so that even when each *dst* token is randomly merged with *src* tokens globally, the merged and subsequently unmerged representation retains structured spatial information while compressing fine details. This spatially coherent perturbation produces a stable and meaningful difference from the original attention, providing a more reliable guidance signal.

Effect of renormalization. Although both our method and PLADIS degrade at very large guidance scales on SDXL, we verify that this issue can be substantially mitigated by our proposed renormalization post-processing. The improvement is illustrated by the red curves in Figure 8. Notably, applying renormalization not only alleviates degradation at high guidance scales but can also further improve the performance of our method, as measured by the metrics. For Flux, however, as shown in Figure 9, applying renormalization does not consistently provide extra benefits, and our guidance remains sufficiently stable on its own. Therefore, we do not apply renormalization for Flux.

MUG on self-attention. We followed the practice in PLADIS (Kim & Sim, 2025) to integrate guidance with cross-attention modules. We consider the alternative option of applying it to self-attention, and find that it renders the generated images significantly degraded or even collapsed, which aligns with the observation made in PLADIS. Some examples are visualized in Figure 10.

Self-attention operates purely on latent image tokens and is responsible for maintaining global spatial coherence—propagating structure, enforcing long-range consistency, and stabilizing the iterative denoising dynamics. It critically determines "how" the image forms. Cross-attention, by contrast, connects latent tokens to text embeddings and primarily injects semantic conditioning, controlling "what" content appears and how strongly regions align with prompt concepts without defining the internal spatial topology of the latent itself.

Because of their different roles, we argue that perturbing cross-attention mainly reshapes semantic strength or alignment and can enhance image fidelity or prompt adherence, while perturbing self-attention directly distorts the spatial propagation mechanism that the diffusion process critically depends on; even small structural changes there accumulate across timesteps and destabilize denoising, an effect exacerbated the single-branch nature of our method, readily leading to noisy or collapsed images.

Effect of guided layers. By default, we apply MUG to the cross-attention blocks of all layers in the U-Net backbone of SDXL, or all joint attention blocks in Flux. This saves us from the laborious heuristic layer selection done in existing works (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025). In this section, we explore the effect of applying MUG to specific layers of the diffusion model, taking SDXL as an example. We split SDXL's U-Net layers into 3 groups, namely downsampling, middle, and upsampling layers. They consist of 24, 10, and 36 layers, respectively. In

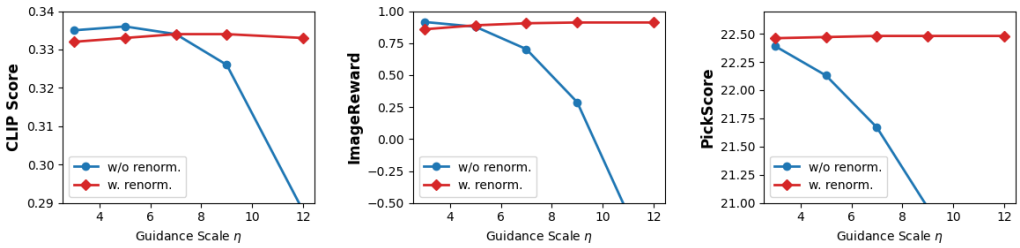


Figure 8: **Qualitative effect of varying guidance scale and the use of the renormalization technique on SDXL.**

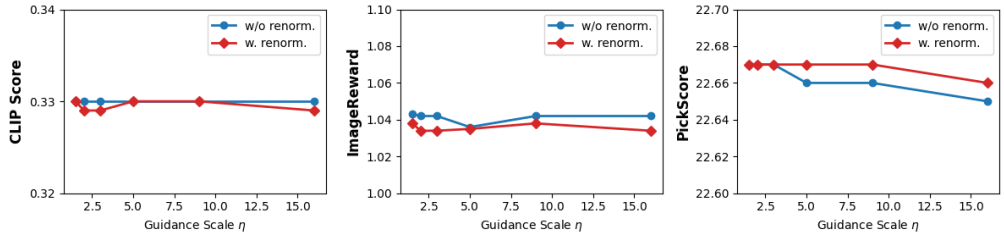


Figure 9: **Qualitative effect of varying guidance scale and the use of the renormalization technique on Flux-schnell.**

Table 19, we report the quantitative results of applying MUG to different groups or combinations of groups. We find that in all cases, MUG successfully guides the diffusion sampling to produce better images. When all layers are guided, the highest performance is achieved across all three metrics. Hence, we simply integrate MUG into all layers of the diffusion model as our default design option, which not only avoids heuristic layer selection but also yields optimal performance. By removing MUG from specific layers (*e.g.*, Down or Mid), we can still achieve competitive results while incurring a lower computational cost.

Ablations on merging policy. In its original context of model acceleration, token merging is designed to merge the most similar tokens in order to eliminate redundancy (Bolya et al., 2023; Bolya & Hoffman, 2023). However, for our sampling guidance problem, we reframe token merging as a means of introducing token-level perturbations. Therefore, we do not necessarily merge redundant tokens. We conduct experiments to ablate this choice of merge policy, where we additionally consider merging the most similar tokens, as usually done for model acceleration (Bolya et al., 2023; Bolya & Hoffman, 2023), as well as merging the most dissimilar tokens. The comparison in Table 16 and Figure 11 reveals that random merging yields the best performance. Inferior results are obtained when we merge most similar tokens like in model acceleration tasks, which is likely due to insufficient perturbation strength and weak guidance direction.

Statistical significance of performance gains. To ensure that the performance gains achieved by MUG are statistically significant, we conduct three independent runs for all methods, each using a different seed, and report the standard deviation of the quantitative measurements. As shown in Tables 17 and 18, all methods exhibit relatively small standard deviations, indicating stable behavior across different random initializations. More importantly, the gains obtained by our method are consistently larger than, or at least comparable to, the corresponding run-to-run fluctuations. For instance, on SDXL, our method improves ImageReward from 0.635 to 0.890 while maintaining a standard deviation of only 0.0398, demonstrating that the observed gain is substantially beyond random variance. Similar trends can also be observed on Flux-schnell, where our method consistently attains the best or tied-best mean performance with tightly bounded deviations. These results verify that the superiority of the proposed method is reproducible and statistically reliable, rather than being caused by favorable random seeds.

Further efficiency analysis. To identify the source of the additional latency, we further profile the major components of the Flux-schnell denoising backbone. As shown in Table 20, the runtime increase is localized almost entirely to the double-stream joint-attention modules, where MUG constructs the degraded attention representation. In contrast, the runtimes of the single-stream joint



Figure 10: **Effect of applying guidance over self-attention.**

Table 16: **Quantitative comparison of different merge policies.**

Policy	CLIP↑	ImageReward↑	PickScore↑
Vanilla	0.327	0.635	22.29
Similar	0.328	0.679	22.28
Dissimilar	0.332	0.859	22.44
Random	**0.333**	**0.890**	**22.47**

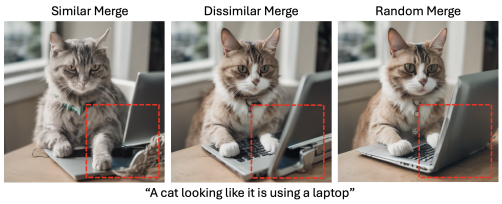


Figure 11: **Visualized effect of different merge policies.**

self-attention blocks, FFNs, and the remaining normalization, gating, and projection operations are essentially unchanged. Overall, MUG increases the end-to-end pipeline latency by only 10.5%.

This localized overhead follows directly from the design of MUG. MUG introduces an additional cross-attention computation at each modified layer but reuses the remaining backbone computation. The additional operations are lightweight compared to the expensive sparse attention used by PLADIS. In contrast, conventional CFG requires separate conditional and unconditional evaluations, duplicating all backbone components, including attention layers, FFNs, normalization layers, and projections. Therefore, although MUG introduces additional attention computation, it avoids a second full-backbone evaluation.

For completeness, in Table 21 we also report the end-to-end latency T_{E2E} and incremental GPU memory cost Δ_{Mem} for the SDXL model. As can be seen, standalone MUG remains close to the vanilla model in both latency and incremental memory cost, while being 18% faster and more memory-efficient than PLADIS. MUG also incurs negligible additional memory overhead when combined with CFG and remains substantially more efficient than CFG+PLADIS. The performance-efficiency comparisons on both SDXL and Flux models show that MUG may work as a more effective yet more efficient alternative to PLADIS.

Effect of alternative token perturbation operations. We have presented token merging and unmerging as a novel way of constructing degraded latents. But these operations still require dedicated calculations. This raises the question: are the benefits worth the complexity? We are interested in exploring several other mechanisms to perturb the latents, including:

- Token shuffle, which randomly rearranges the spatial location of each token without changing its values;
- Token downsampling & upsampling, which performs 2×2 spatial average pooling to downsample the token features, followed by 2×2 spatial interpolation to restore the original spatial resolution;
- Token blurring, which applies a Gaussian kernel of size 3×3 and $\sigma = 1.5$ to blur the latent.

Note that for the sake of fairness, we enable latent renormalization for all policies. The comparisons are presented in Table 22.

We first observe that token downsampling & upsampling and token blurring both underperform, even though they involve mixing token features like MUG. The reason is that these operations act as data-agnostic low-pass filters: by averaging neighboring feature vectors, they reduce the variance of the latent manifold and effectively erase the high-frequency structural information. Hence, the guidance direction $\Delta\mathbf{Z} = \mathbf{Z} - \mathbf{Z}'$ acts merely as a localized sharpening filter. Because these methods

Table 17: **Main results on SDXL with standard deviations.**

Method	#Steps	NFE	CLIP↑	ImageReward↑	PickScore↑
CFG	50	100	0.327±0.0015	0.635±0.0561	22.29±0.0611
CFG+PLADIS	50	100	0.332±0.0010	0.796±0.0341	22.24±0.0100
CFG+MUG (ours)	50	100	**0.333**±0.0015	**0.890**±0.0398	**22.47**±0.0436

Table 18: **Main results on Flux-schnell with standard deviations.**

Method	#Steps	NFE	CLIP↑	ImageReward↑	PickScore↑
Vanilla	4	4	0.329±0.0015	1.036±0.0130	22.66±0.0608
Vanilla+PLADIS	4	4	0.331 ±0.0015	1.040±0.0188	22.65±0.0693
Vanilla+MUG (ours)	4	4	**0.333**±0.0021	**1.108**±0.0177	**22.85**±0.0635

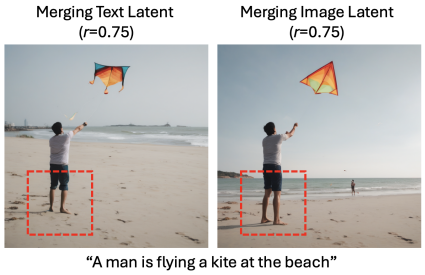


Figure 12: **Visualized effect of merging text embeddings.**

introduce interpolation artifacts and spatial phase shifts, the resulting direction often lacks semantic coherence and is less structured.

Token shuffling performs better because it is a distribution-preserving transformation. While blurring destroys the statistical profile of the features, shuffling maintains the exact histogram of the feature tokens while only permuting their spatial assignments. This ensures that the guidance remains stable and grounded in the model's learned feature manifold. However, shuffling remains suboptimal because it is globally stochastic. By fully decoupling tokens from their spatial positions, the model is forced to handle a highly unstable gradient with no spatial reference, effectively requiring the denoising process to recover the global spatial layout from a completely randomized token arrangement.

In contrast, MUG performs best by combining fixed spatial anchors with controlled semantic perturbation. Window-based *dst* sampling ensures that, after unmerging, features can be projected back onto spatially consistent locations, preserving global coverage. Meanwhile, mixing these anchors with random global *src* tokens creates a bottleneck where local context is mixed with global variation. The resulting guidance signal specifically isolates the difference between original local structure and injected global noise, producing a stable and more structured guidance signal, allowing the sampling to prioritize the recovery of local semantic details rather than merely correcting blurred or randomly scrambled features.

Theoretical grounding. Recent inference-time guidance methods (Hong, 2024; Kim & Sim, 2025) can be interpreted through the lens of associative memory and energy-based attention dynamics (Park et al., 2023; Ramsauer et al., 2020). Smoothed Energy Guidance (SEG) (Hong, 2024) reduces the *curvature* of an attention-induced energy landscape by smoothing the corresponding energy, thereby stabilizing the update. PLADIS (Kim & Sim, 2025) instead modifies attention *connectivity* at inference time by sparsifying attention weights, effectively changing the associative retrieval rule. In contrast, MUG constructs a *worse* retrieval target by *coarsening* image tokens through merging and broadcasting the coarsened representation back through unmerging, and then performs a CFG-style *repulsion* away from this degraded attractor, all within a single model forward pass.

Inspired by SEG's energy-based interpretation of self-attention blurring guidance, we likewise formulate cross-attention as modern Hopfield retrieval. Consider a single cross-attention layer with image queries $\mathbf{Q} \in \mathbb{R}^{N\times d}$ and text keys and values $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{M\times d}$. With scaling $\beta = 1/\sqrt{d}$, the

Table 19: **Effect of applying MUG to different layers of SDXL.**

Down	Mid	Up	CLIP↑	ImageReward↑	PickScore↑
w/o MUG			0.325	0.565	22.11
✓			0.328	0.665	22.14
	✓		0.326	0.574	22.13
		✓	0.330	0.778	22.29
✓	✓		0.328	0.679	22.16
✓		✓	**0.332**	0.838	22.29
	✓	✓	0.331	0.794	**22.32**
✓	✓	✓	**0.333**	**0.890**	**22.47**

Table 20: **Component-level runtime breakdown on Flux-schnell.**

Component	Vanilla (ms)	MUG (ms)
Double-stream joint attention	391	637
Single-stream joint self-attention	655	654
FFN / MLP	496	492
Normalization, gating, projections	470	469
Total pipeline	2,254	2,490 ↑ 10.5%

attention output is

$$\mathbf{Z} = \mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\big(\beta \mathbf{Q}\mathbf{K}^\top\big)\, \mathbf{V} \in \mathbb{R}^{N\times d}. \tag{7}$$

Following the modern Hopfield-network view (Park et al., 2023), $\mathrm{Softmax}(\beta\mathbf{Q}\mathbf{K}^\top)$ implements a content-addressable memory lookup: each query $\mathbf{q}_i$ retrieves a convex combination of the value patterns $\{\mathbf{v}_j\}$, with retrieval weights exponentially tilted by query–key similarity.

A convenient energy associated with a single query $\mathbf{q}$ (omitting index i) and fixed memory $\{(\mathbf{k}_j, \mathbf{v}_j)\}_{j=1}^M$ is

$$\mathcal{E}(\mathbf{q}) = -\frac{1}{\beta} \log \sum_{j=1}^{M} \exp\big(\beta \mathbf{q}^\top \mathbf{k}_j\big)\,, \tag{8}$$

whose gradient is

$$\nabla_{\mathbf{q}} \mathcal{E}(\mathbf{q}) = -\sum_{j=1}^{M} a_j(\mathbf{q})\, \mathbf{k}_j = -\mathbf{K}^\top \mathbf{a}(\mathbf{q}), \tag{9}$$

where $\mathbf{a}(\mathbf{q}) \in \mathbb{R}^M$ denotes the attention-weight vector for query $\mathbf{q}$, with

$$a_j(\mathbf{q}) = \frac{\exp\big(\beta \mathbf{q}^\top \mathbf{k}_j\big)}{\sum_{m=1}^{M} \exp(\beta \mathbf{q}^\top \mathbf{k}_m)}, \qquad \mathbf{a}(\mathbf{q}) = \mathrm{Softmax}(\beta \mathbf{K}\mathbf{q})\,. \tag{10}$$

Thus, attention can be viewed as an energy-driven associative alignment of queries with keys, while its output is the corresponding value readout. Under this interpretation, SEG reduces the effective Hessian, *i.e.*, the curvature, of the attention energy to avoid overly sharp minima during sampling, whereas PLADIS changes the effective memory interactions by enforcing sparsity in $\mathbf{a}(\mathbf{q})$.

We next show that MUG's token merging and unmerging can be interpreted as a low-rank coarsening operation. Let $\mathbf{H} \in \mathbb{R}^{N\times d}$ denote the latent tokens before cross-attention, and let $\mathbf{Q} = \mathbf{H}\mathbf{W}_Q$ be the corresponding query projection. MUG's token-merging operator reduces the token count to $N' = (1-r)N$:

$$\mathbf{H}^m = \mathrm{Merge}(\mathbf{H}, r) \in \mathbb{R}^{N'\times d}, \qquad \mathbf{Q}^m = \mathbf{H}^m \mathbf{W}_Q \in \mathbb{R}^{N'\times d}. \tag{11}$$

Cross-attention is then applied to the merged queries, followed by an unmerging, or broadcasting, operation:

$$\mathbf{Z}^m = \mathrm{Attn}(\mathbf{Q}^m, \mathbf{K}, \mathbf{V}) \in \mathbb{R}^{N'\times d}, \qquad \mathbf{Z}' = \mathrm{Unmerge}(\mathbf{Z}^m) \in \mathbb{R}^{N\times d}. \tag{12}$$

Table 21: **End-to-end latency and incremental memory cost on SDXL.**

Metric	Vanilla	CFG	PLADIS/+CFG	MUG/+CFG
T_{E2E} (s/img) ↓	4.216	6.814	5.895 / 9.276	**4.826 / 7.943**
Δ_{Mem} (MB) ↓	166.96	327.67	198.12 / 392.01	**167.68 / 328.95**

Table 22: **Effect of different token perturbation strategies.**

Perturbation Policy	**CLIP↑**	**ImageReward↑**	**PickScore↑**
Token Shuffling	0.331	0.822	22.25
Token Downsampling & Upsampling	0.328	0.638	22.04
Token Blurring	0.327	0.624	22.04
Token Merging & Unmerging (ours)	**0.333**	**0.890**	**22.47**

Abstractly, merging and unmerging can be modeled as linear maps

$$\mathbf{P} \in \mathbb{R}^{N' \times N} \quad \text{and} \quad \mathbf{U} \in \mathbb{R}^{N \times N'}, \tag{13}$$

where $\mathbf{P}$ performs token aggregation and $\mathbf{U}$ broadcasts the aggregated representations back to their original positions. Ignoring minor implementation details, we have

$$\mathbf{H}^m \approx \mathbf{PH}, \qquad \mathbf{Z}' \approx \mathbf{U}\,\mathrm{Attn}(\mathbf{PHW}_Q, \mathbf{K}, \mathbf{V})\,. \tag{14}$$

For average merging and unmerging, $\mathbf{P}$ locally averages tokens and $\mathbf{U}$ replicates each averaged representation across the corresponding merged members. Consequently, $\mathbf{UP}$ behaves as a token-coarsening projector that suppresses fine-grained token variations:

$$\mathbf{UP} \approx \mathbf{\Pi}_{\text{coarse}}, \qquad \mathbf{I} - \mathbf{\Pi}_{\text{coarse}} \approx \mathbf{\Pi}_{\text{detail}}. \tag{15}$$

The resulting representation therefore serves as a controlled "bad" latent: it loses discriminative spatial details while largely retaining the global semantic structure and the original text condition.

From the Hopfield perspective, MUG performs repulsion from this degraded associative retrieval. Specifically, it modifies the cross-attention output by extrapolating away from the coarsened and unmerged retrieval:

$$\mathbf{Z}_{\text{MUG}} = \mathbf{Z} + \eta(\mathbf{Z} - \mathbf{Z}') = (1+\eta)\mathbf{Z} - \eta\mathbf{Z}'. \tag{16}$$

Here, $\mathbf{Z}$ is the layer's original retrieval, whereas $\mathbf{Z}'$ is produced by coarsened queries, or equivalently by queries constrained to a coarse token subspace. Equation 16 can also be viewed as a negative-gradient step on the local repulsive energy

$$\mathcal{L}_{\text{rep}}(\mathbf{Z}; \mathbf{Z}') = -\frac{1}{2}\,\|\mathbf{Z} - \mathbf{Z}'\|_F^2\,, \tag{17}$$

where $\mathbf{Z}'$ is treated as the fixed degraded reference for the current layer. Indeed,

$$\mathbf{Z} - \eta\nabla_{\mathbf{Z}}\mathcal{L}_{\text{rep}}(\mathbf{Z}; \mathbf{Z}') = \mathbf{Z} + \eta(\mathbf{Z} - \mathbf{Z}') = \mathbf{Z}_{\text{MUG}}. \tag{18}$$

When the degraded retrieval approximately corresponds to a projection of the original retrieval, *i.e.*, $\mathbf{Z}' \approx \mathbf{\Pi}_{\text{coarse}}(\mathbf{Z})$, this update becomes

$$\mathbf{Z}_{\text{MUG}} \approx \mathbf{Z} + \eta\mathbf{\Pi}_{\text{detail}}(\mathbf{Z}), \tag{19}$$

showing that MUG amplifies the residual components not preserved by the coarse projector. Through the Hopfield lens, $\mathbf{Z}'$ behaves as an easy but overly smooth associative retrieval—a broad, low-resolution attractor—and MUG pushes the representation away from this attractor toward a sharper, more discriminative retrieval that better preserves fine details.

The analysis above identifies a meaningful *layer-local* guidance direction, but it does not by itself imply that extrapolation in an intermediate feature space is equivalent to extrapolation in the final score space. We therefore characterize how these layer-local perturbations propagate to the model prediction. Let

$$\widehat{\mathbf{y}}_t^0 = F_{\theta,t}(\mathbf{x}_t, \mathbf{c}) \tag{20}$$

denote the original model prediction at time t, where $\widehat{\mathbf{y}}_t$ represents the predicted noise for an ϵ-prediction model or the predicted velocity for a flow-based model. At a modified cross-attention layer l, define the MUG perturbation as

$$\boldsymbol{\delta}_l = \eta_l \left(\mathbf{Z}_l - \mathbf{Z}'_l \right). \tag{21}$$

Let $G_{l,t}$ denote the downstream mapping from the output of layer l to the final model prediction, with the remaining inputs and network state held implicit. For a single modified layer, the corresponding output change admits the exact integral form

$$G_{l,t}(\mathbf{Z}_l + \boldsymbol{\delta}_l) - G_{l,t}(\mathbf{Z}_l) = \int_0^1 \mathbf{J}_{G_{l,t}}(\mathbf{Z}_l + \tau \boldsymbol{\delta}_l)\, \boldsymbol{\delta}_l \, d\tau, \tag{22}$$

where $\mathbf{J}_{G_{l,t}}$ is the Jacobian of the downstream network mapping.

Applying a telescoping decomposition over all modified layers and taking a first-order expansion gives

$$\widehat{\mathbf{y}}_t^{\mathrm{MUG}} - \widehat{\mathbf{y}}_t^{\,0} = \sum_l \mathbf{J}_{l\to\mathrm{out},t} \boldsymbol{\delta}_l + \mathcal{O}\left(\left(\sum_l \|\boldsymbol{\delta}_l\|_F \right)^2 \right), \tag{23}$$

where $\mathbf{J}_{l\to\mathrm{out},t}$ denotes the Jacobian mapping a perturbation at layer l to the final output, evaluated along the unperturbed forward trajectory. Thus, the pretrained network's downstream Jacobians transform and aggregate MUG's layer-local associative-memory directions into a change in the final prediction field. This analysis does not require the intermediate extrapolation itself to be identical to an extrapolation in score space.

For an ϵ-prediction diffusion model, the predicted score satisfies

$$\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{c}) = -\frac{\widehat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, \mathbf{c})}{\sigma_t}, \tag{24}$$

and therefore MUG induces the score change

$$\Delta \mathbf{s}_t = -\frac{\Delta \widehat{\boldsymbol{\epsilon}}_t}{\sigma_t}, \qquad \Delta \widehat{\boldsymbol{\epsilon}}_t = \widehat{\boldsymbol{\epsilon}}_t^{\mathrm{MUG}} - \widehat{\boldsymbol{\epsilon}}_t^{\,0}. \tag{25}$$

For a sampler whose update can be expressed as

$$\mathbf{x}_{t-1} = A_t \mathbf{x}_t + B_t \widehat{\boldsymbol{\epsilon}}_t, \tag{26}$$

the corresponding change in the next latent is

$$\Delta \mathbf{x}_{t-1} = B_t \Delta \widehat{\boldsymbol{\epsilon}}_t. \tag{27}$$

For a velocity-prediction or flow-based model, $\widehat{\mathbf{y}}_t$ directly parameterizes the sampling vector field, so Equation 23 directly characterizes the MUG-induced change in that field; a discretized flow update inherits the corresponding step-size-scaled perturbation. Equations 21–27 therefore establish the connection from MUG's intermediate cross-attention perturbations to the final model prediction and sampling trajectory.

This theoretical view also clarifies MUG's relationship to recent guidance methods. SEG modifies the energy landscape by smoothing the attention energy and reducing curvature effects, thereby stabilizing retrieval updates. PLADIS modifies the retrieval rule by imposing sparsity on attention weights, changing which memories participate in the Hopfield update. MUG instead preserves the original retrieval rule and text condition, constructs an explicit degraded attractor $\mathbf{Z}'$ by coarsening the image queries, and performs an anti-attractor step that amplifies components discarded by the coarse representation. The associative-memory analysis explains why $\mathbf{Z} - \mathbf{Z}'$ provides a meaningful layer-local direction, while the Jacobian analysis explains how this direction propagates through the downstream network to alter the final score or velocity field.

The "repel from coarse memory" interpretation also helps explain why random merging can outperform similarity-based merging in our ablations. Similarity-based merging primarily groups already similar tokens, making the resulting perturbation relatively weak. Random merging introduces a larger mismatch among fine-scale token assignments and thus produces a stronger and more informative repulsion direction.

In summary, under the associative-memory interpretation of attention, MUG performs a single-pass, layer-local guidance step that increases the distance from a low-resolution Hopfield attractor induced by token coarsening. The resulting perturbations are propagated by the pretrained network's downstream Jacobians to modify the final prediction and sampling vector field. MUG therefore provides an efficient guidance signal compatible with native single-branch settings, including distilled or inherently single-pass backbones, while remaining conceptually complementary to curvature smoothing (*i.e.*, SEG) and sparsity-based retrieval (*i.e.*, PLADIS).

D More Visualized Samples

In the remainder of the appendix, Figures 13 to 16 provide more visualized examples of images generated by MUG and other methods using different configurations.

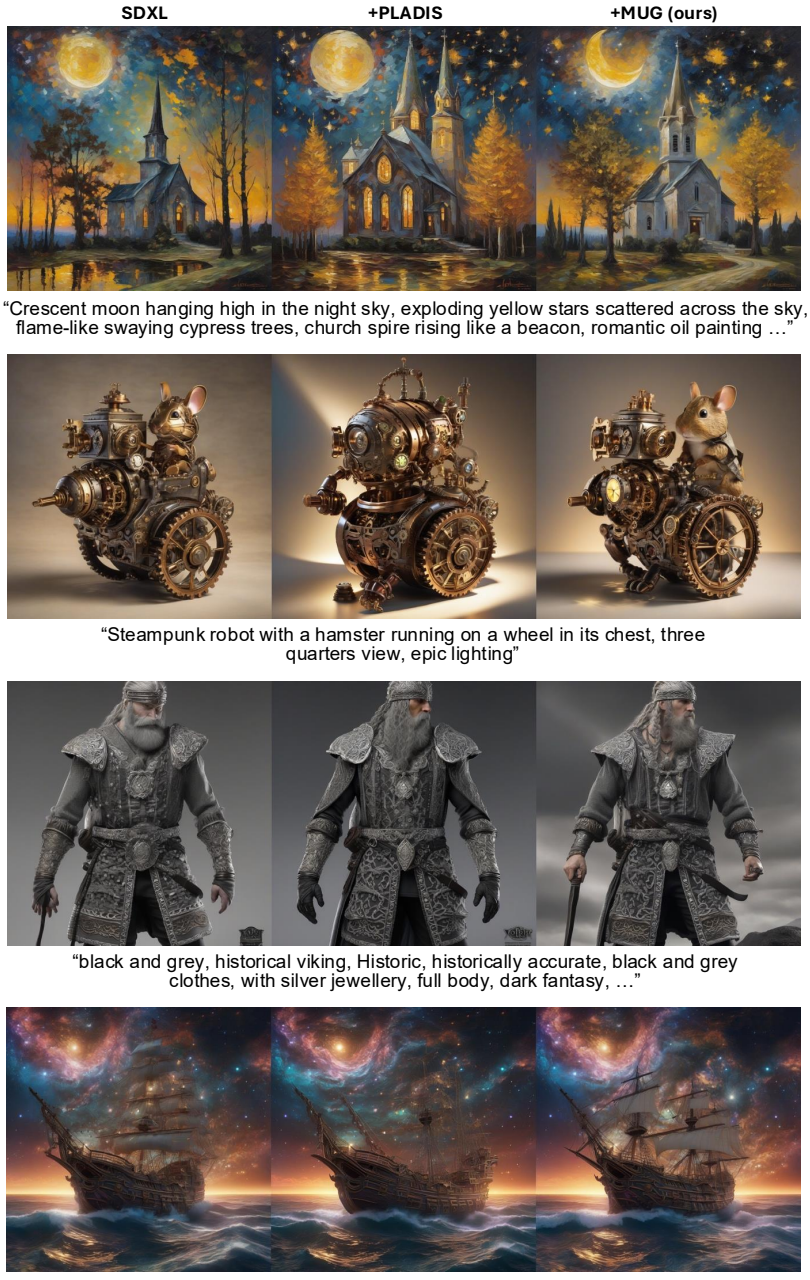
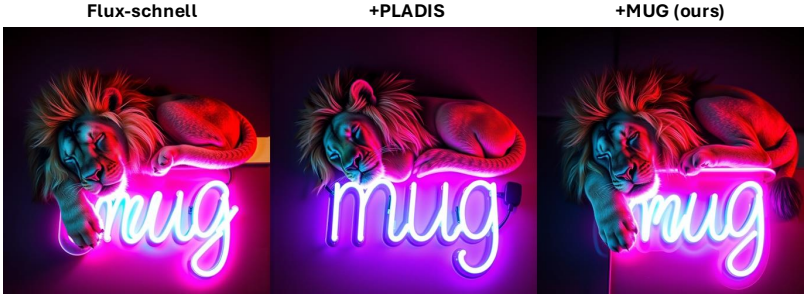


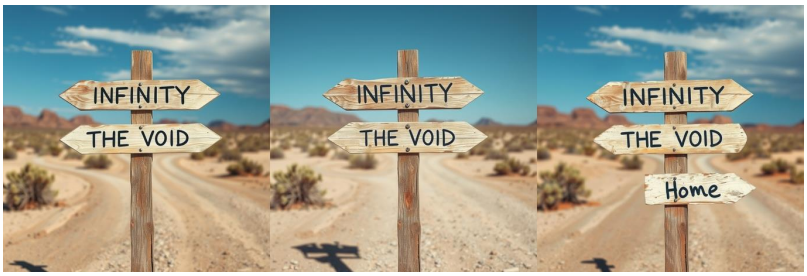
Figure 13: **More qualitative comparisons on SDXL generation.**



“A lion sleeping on a neon sign that writes ‘mug’”



“A mysterious underwater city with bioluminescent corals and towering glass domes”



“A weathered wooden signpost at a desert crossroads, with three arrows labeled: 'INFINITY', 'THE VOID', and 'Home'.”



“A dragon attacking burning medieval hobbit homes in a picturesque landscape with a waterfall and a bridge”



Figure 14: **More qualitative comparisons on Flux-schnell generation.**

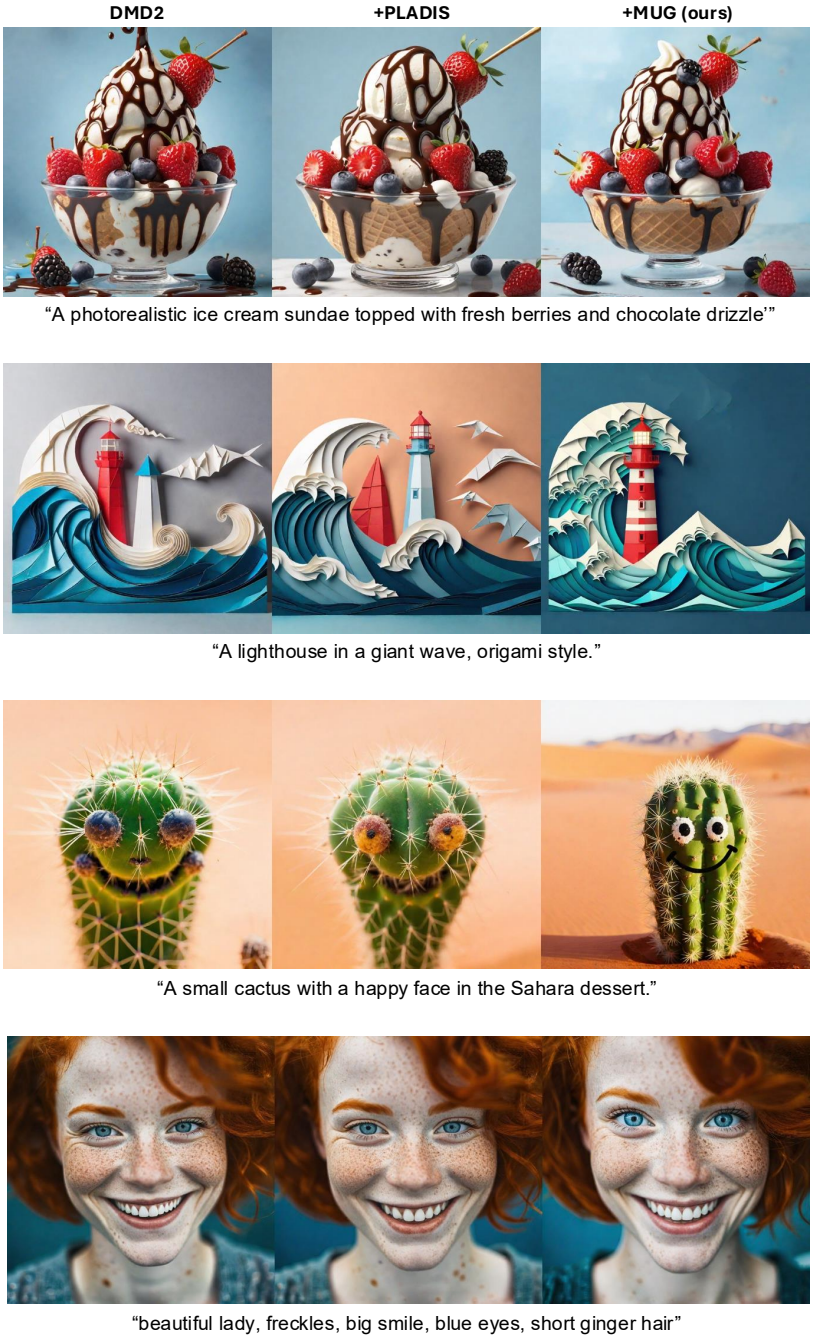
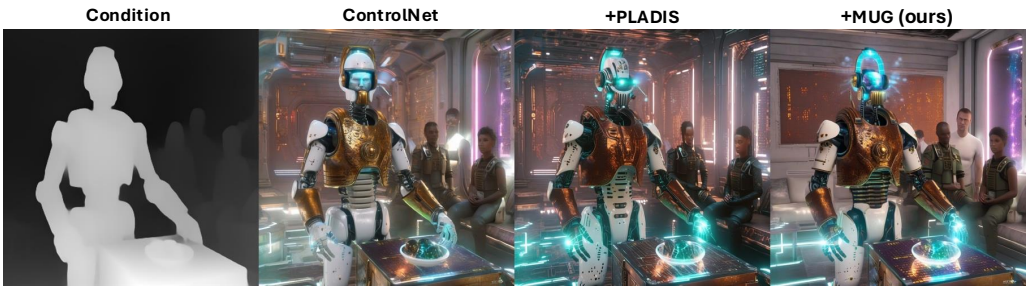


Figure 15: **More qualitative comparisons on DMD2 (4-step) generation.**



"A slightly rusted cyberspace marine warrior is manipulating a floating matrix hologram"



"A realistic owl standing on the grass, wearing a pair of black sunglasses over its eyes."



"An orange cat is lying on the train engine."



"A cube carved from white porcelain, along with a bouquet of flowers in a gold-rimmed vase, is placed in a dreamy castle."

Figure 16: **More qualitative comparisons on depth-guided ControlNet generation.**