VAGS: VELOCITY ADAPTIVE GUIDANCE SCALE FOR IMAGE EDITING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifier-free guidance (CFG) is the primary control over how strongly text semantics move a flow-based sampler, yet standard practice holds its scale fixed across the entire ODE trajectory. This is a fundamental mismatch: early steps are noise-dominated and carry weak semantic signal, while late steps commit image structure and demand stronger directional commitment; more critically, the value of any guidance strength depends on whether the guided velocity is consistent with the model's current dynamics or working against them. We propose *Velocity-Adaptive Guidance Scale* (VAGS), a training-free replacement that multiplies the nominal scale by a bounded factor combining a temporal signal-level term with the cosine similarity between task-relevant velocity fields. For inversion-free editing, VAGS measures the alignment between source- and target-guided velocities, so edit strength at each step reflects local compatibility between preservation and transformation. For generation, VAGS-Gen uses the alignment between unconditional and conditional velocities as the analogous signal. Neither variant requires fine-tuning, auxiliary networks, or extra forward passes, and fixed CFG is recovered as a special case. On PIE-Bench and DIV2K for editing, and COCO17, CUB-200, and Flickr30K for generation, VAGS consistently improves structural fidelity and generation quality over fixed CFG and recent training-free guidance variants. The codebase is available in the supplementary materials.

1 INTRODUCTION

Diffusion and flow-based generative models now define the dominant paradigm for text-conditioned image synthesis, from iterative denoising (Ho et al., 2020; Song et al., 2021b) and latent diffusion (Rombach et al., 2022) to DDIM-style sampling (Song et al., 2021a), flow matching (Lipman et al., 2023), rectified flow (Liu et al., 2023), and recent large-scale text-to-image systems (Esser et al., 2024; Black Forest Labs, 2024). Across these models, classifier-free guidance (CFG) (Ho & Salimans, 2022) is the standard mechanism for steering generation. It extrapolates from unconditional to conditional predictions, compressing the entire trade-off between prompt adherence and sample fidelity into a single scalar. In flow-matching and rectified-flow models, that scalar directly modulates the velocity field driving the sampling ODE, and therefore governs the full trajectory from noise to image. The same control has become central to image editing, including inversion-based editors (Hertz et al., 2023; Tumanyan et al., 2023; Mokady et al., 2023; Wallace et al., 2023; Brooks et al., 2023), rectified-flow inversion methods (Rout et al., 2024; Wang et al., 2024a; Deng et al., 2024; Xu et al., 2025), and recent inversion-free editors (Xu et al., 2023; Kulikov et al., 2024; Yang et al., 2025; Yoon et al., 2025), where edits are produced by integrating differences between source- and target-conditioned velocities. CFG is not a peripheral sampling detail: it is the primary control over how strongly text semantics move a flow-based sampler. Fig. 1 illustrates this fixed-scale failure mode: a constant guidance scale can accumulate semantic drift along the sampling trajectory, whereas VAGS adapts the scale using velocity alignment to reinforce reliable directions and suppress harmful ones.

Holding this scalar constant across the entire ODE trajectory is the standard practice, and it is the wrong one. Flow-based sampling does not evolve uniformly. In the noisy regime, velocity differences carry weak or unstable semantic signal, and strong guidance amplifies uninformative discrepancies that pull the trajectory off the data manifold. In the clean regime, structure has accumulated, semantic directions are sharp, and strong guidance is needed to commit the visual details a prompt demands.

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.

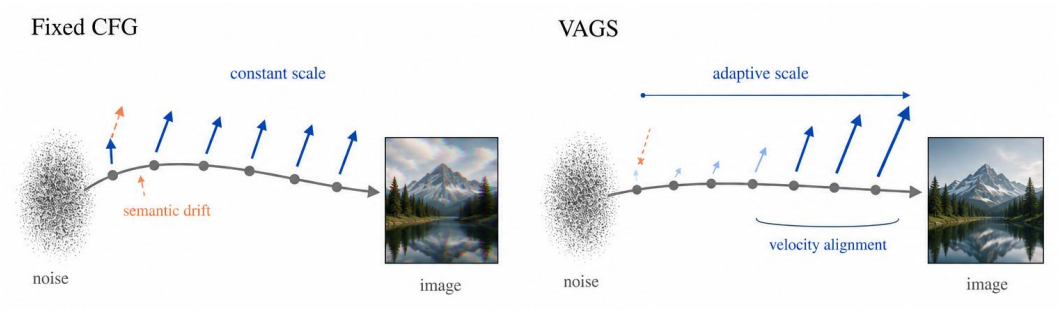


Figure 1: Overview of Velocity-Adaptive Guidance Scale (VAGS). Fixed CFG uses a constant scale along the sampling trajectory, which can cause semantic drift and structural distortion. VAGS instead adaptively adjusts the guidance scale according to denoising progress and velocity alignment, improving semantic control while preserving image structure without extra training.

Dynamic guidance schedules partially recognize this nonuniformity (Chang et al., 2023; Wang et al., 2024b; Kynkäänniemi et al., 2024a; Papalampidi et al., 2025; Galashov et al., 2025), and recent rectified-flow analyses connect CFG to off-manifold drift (Fan et al., 2025; Saini et al., 2025). However, monotone or searched schedules still bypass the question that actually matters at each step: whether the guided direction is consistent with the model's current dynamics, or whether it is working against structure that has already formed.

The consequences are concrete. In text-driven editing, methods must transform the requested content while preserving unrelated source structure. Existing approaches expose this tension through attention control, inversion refinement, or source-target velocity transport (Hertz et al., 2023; Tumanyan et al., 2023; Mokady et al., 2023; Kulikov et al., 2024; Yoon et al., 2025). A large fixed target scale achieves the requested object change but also displaces background texture, identity, pose, and layout from the source image; reducing it preserves the source but weakens or loses the edit. In generation, a single scalar must serve both early global composition and late local refinement, so the value that improves prompt adherence simultaneously introduces saturation, duplicate objects, or brittle fine structure. The problem is not finding the right average guidance strength; it is recognizing that no fixed scalar can be right at every step.

Flow-based samplers already expose the signal needed to resolve this. Each model call returns a velocity: the direction in latent space the sampler is about to follow. Once image structure has emerged, alignment between relevant velocities supports stronger guidance, whereas conflict warns that guidance is extrapolating against the reference dynamics, the condition under which unrelated structure drifts and artifacts emerge. Cosine similarity directly quantifies this angular relationship. It is insensitive to the timestep-dependent scale of the ODE and responds instead to the compatibility that varies with prompt difficulty, image content, and sampling stage. This is a clean, principled signal, and it has gone unused.

We introduce *Velocity-Adaptive Guidance Scale* (VAGS), a training-free, plug-and-play replacement for fixed CFG. Instead of applying one constant scale throughout sampling, VAGS modulates the nominal scale at each step with a bounded factor determined by denoising progress and velocity alignment. The signed temporal term changes how alignment is interpreted across the trajectory: in the noisy regime, agreement attenuates guidance and disagreement amplifies it; in the clean regime, agreement amplifies guidance and disagreement attenuates it. Fixed CFG is recovered as a special case, and the added cost per step is a single inner product.

The formulation applies directly to both tasks. For inversion-free editing, VAGS adapts the target scale from the alignment between the source-guided velocity and a pilot target-guided velocity, so edit strength at each step reflects the local compatibility between preservation and transformation. For text-to-image generation, VAGS-Gen uses the alignment between unconditional and conditional velocities as the analogous measure of compatibility with the data prior. Neither variant requires extra model evaluations, external evaluators, fine-tuning, or auxiliary networks.

Experiments confirm that velocity-adaptive guidance improves both tasks. On PIE-Bench (Ju et al., 2023) and DIV2K (Agustsson & Timofte, 2017), VAGS improves source preservation while retaining competitive edit alignment when plugged into FlowEdit (Kulikov et al., 2024) and Split-

Flow (Yoon et al., 2025). On COCO17 (Lin et al., 2014), CUB-200 (Welinder et al., 2010), and Flickr30K (Plummer et al., 2015), VAGS-Gen improves FID and Inception Score over fixed CFG and recent training-free guidance variants (Jin et al., 2025; Li et al., 2025b). Ablations against interval (Kynkäänniemi et al., 2024b), monotone (Wang et al., 2024b), zero-initialized (Fan et al., 2025), and matched-mean CFG schedules isolate step-level velocity regulation as the source of the gains, ruling out any benefit from a shifted average scale.

Our contributions are:

- We identify fixed CFG as a shared bottleneck in flow-based image editing and generation, where a uniform scale ignores both the temporal structure of the ODE and the step-wise compatibility of the guided velocity.
- We propose VAGS, a unified adaptive guidance rule that combines signal level with velocity alignment at negligible computational overhead.
- We instantiate VAGS for editing and generation without additional training, networks, or model forward passes.
- We demonstrate consistent gains across five benchmarks and confirm through scheduler and matched-mean ablations that step-level adaptivity, not average scale, drives the improvement.

2 Related Work

Diffusion and Flow-based Generative Models. Diffusion models (Ho et al., 2020; Song et al., 2021b) generate images by iterative denoising, with later improvements including latent diffusion (Rombach et al., 2022), classifier-free guidance (Ho & Salimans, 2022), DDIM sampling (Song et al., 2021a), and consistency distillation (Song et al., 2023). Flow Matching (Lipman et al., 2023) and Rectified Flow (Liu et al., 2023) instead learn an ODE velocity field that transports noise to data, yielding a direct geometric object for sampling control. This formulation underlies recent high-quality text-to-image systems such as Stable Diffusion 3 (Esser et al., 2024) and FLUX (Black Forest Labs, 2024).

Text-driven Image Editing. Text-driven image editing modifies a real image according to a target prompt while preserving unrelated content. Inversion-based diffusion editors first map the image to a noisy latent and then regenerate it, often with attention or inversion refinements (Hertz et al., 2023; Tumanyan et al., 2023; Mokady et al., 2023; Wallace et al., 2023; Brooks et al., 2023). Similar inversion-based strategies have been adapted to rectified flows (Rout et al., 2024; Wang et al., 2024a; Deng et al., 2024; Xu et al., 2025), but inversion error continues to trade editability against reconstruction. Inversion-free editors avoid this by integrating source/target velocity differences directly, as in InfEdit (Xu et al., 2023), FlowEdit (Kulikov et al., 2024), iRFDS (Yang et al., 2025), and SplitFlow (Yoon et al., 2025). VAGS targets this latter family by adapting the target guidance scale rather than changing the editing ODE.

Guidance Scheduling. CFG is commonly applied with a fixed scale, although dynamic schedules often improve the fidelity/adherence trade-off. Muse (Chang et al., 2023) and subsequent analyses (Wang et al., 2024b; Kynkäänniemi et al., 2024a) show that the best guidance strength varies across noise levels, while other methods learn or search for dynamic scales (Papalampidi et al., 2025; Galashov et al., 2025). For rectified flows, recent work studies CFG-induced off-manifold drift and proposes corrected guidance updates (Fan et al., 2025; Saini et al., 2025); for editing, FGS (Cho & Kim, 2025) uses scheduled faithfulness guidance. VAGS is complementary: it is training-free, uses no external controller, and adapts guidance through the velocity alignment already available inside the sampler.

Adaptive Classifier-Free Guidance. Recent methods adapt CFG without an external controller, but use different signals. Temporal methods such as C^2FG (Gao et al., 2026) and β-CFG (Malarz et al., 2025) prescribe timestep-only schedules, without conditioning the scale on the current sample or velocity geometry. S-CFG (Shen et al., 2024) instead assigns spatially varying scales using cross-attention, while APG (Sadat et al., 2025) modifies the guidance update by down-weighting its component parallel to the conditional score. Related approaches reduce CFG evaluations after convergence (Castillo et al., 2025), alter the unconditional input in masked-diffusion language models (Li et al., 2025a), or modulate negative prompting (Koulischer et al., 2025). VAGS instead

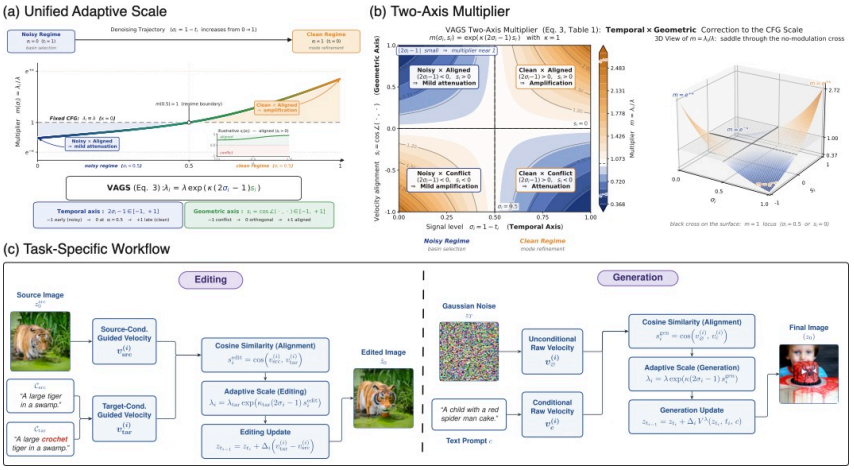


Figure 2: Workflow of Velocity-Adaptive Guidance Scale (VAGS). VAGS replaces fixed CFG with $\lambda_i = \lambda \exp(\kappa(2\sigma_i - 1)s_i)$, where σ_i tracks denoising progress and s_i measures velocity alignment. The same rule applies to editing with source/target alignment and to generation with unconditional/conditional alignment.

combines denoising progress with step-specific velocity alignment already available from the sampler, requiring no auxiliary network, attention map, or additional model evaluation.

3 Velocity-Adaptive Guidance Scale

We propose *Velocity-Adaptive Guidance Scale* (VAGS) to replace the constant scale with a per-step value controlled by two signals already exposed by the ODE: the current signal level and the angular alignment of relevant velocity fields. The same rule applies to inversion-free image editing and text-to-image generation. Full pseudocodes are provided in Algorithms 1 and 2 in the appendix. Fig. 2 summarises the adaptive scale by a two-axis multiplier and the task-specific workflows.

Flow matching and CFG. Flow-matching models (Lipman et al., 2023; Liu et al., 2023) learn a velocity field v_θ that transports noise to data along the interpolant $z_t = (1-t)x + t\epsilon$, with $z_0 = x$ and $z_1 = \epsilon$. Sampling integrates $\mathrm{d}z_t/\mathrm{d}t = v_\theta(z_t, t, c)$ from $t = 1$ to $t = 0$ over $\{t_i\}_{i=1}^N$. At step i, vanilla CFG combines the unconditional and conditional velocities as

$$V^\lambda\big(z_{t_i}, t_i, c\big) = v_\varnothing^{(i)} + \lambda\left(v_c^{(i)} - v_\varnothing^{(i)}\right), \tag{1}$$

where λ is held fixed in standard rectified-flow samplers such as SD3 (Esser et al., 2024) and FLUX (Black Forest Labs, 2024).

Shared sampler form. Both generation and inversion-free editing can be written as an Euler update using a reference and target guided velocity:

$$z_{t_{i-1}} = z_{t_i} + \Delta_i\, \mathcal{F}\big(V_{t_i}^{\mathrm{ref}}, V_{t_i}^{\mathrm{tgt}}\big), \qquad \Delta_i = t_{i-1} - t_i < 0. \tag{2}$$

For generation, $V^{\mathrm{ref}} = 0$ and $\mathcal{F}(0, V^{\mathrm{tgt}}) = V^{\mathrm{tgt}}$. For inversion-free editing (Kulikov et al., 2024; Yoon et al., 2025), V^{ref} and V^{tgt} are source- and target-prompt guided velocities, and $\mathcal{F}(V^{\mathrm{ref}}, V^{\mathrm{tgt}}) = V^{\mathrm{tgt}} - V^{\mathrm{ref}}$. Thus CFG controls the strength of the same target field in both tasks.

Fixed-scale bottleneck. A constant λ treats all steps as equally trustworthy. This conflicts with the trajectory geometry: early latents are noisy and weakly semantic, whereas late latents contain committed structure and sharper prompt directions. A constant scale is also content-blind. It cannot distinguish a step where the target direction agrees with the model's current dynamics from one where guidance fights the trajectory, the latter being the setting where high CFG produces off-manifold artefacts in rectified flows (Fan et al., 2025; Saini et al., 2025).

Per-step adaptive scale. The adaptive rule should satisfy three constraints. First, it should preserve the familiar CFG scale as the nominal strength, so existing sampler settings remain meaningful.

Second, it should vary smoothly across the trajectory, because abrupt scale changes can introduce artifacts of their own. Third, it should modulate symmetrically around fixed CFG, with a regime-dependent reading of agreement and conflict and a range controlled by one parameter.

Let $\sigma_i = 1 - t_i$ denote signal level and let $s_i \in [-1, 1]$ be the cosine similarity between two task-specific velocities at step i. VAGS sets

$$\lambda_i = \lambda \exp\Big(\kappa(2\sigma_i - 1)s_i\Big), \tag{3}$$

where κ controls modulation strength (Fig. 2a). The cosine measures whether the local velocity geometry supports stronger guidance, and the temporal factor $(2\sigma_i - 1)$ changes sign at the trajectory midpoint ($\sigma_i = 0.5$), which reverses the role of that signal. In the clean regime ($\sigma_i > 0.5$), aligned directions are amplified and conflicting directions are damped, as intended. In the noisy regime ($\sigma_i < 0.5$), this polarity is reversed: aligned directions are instead damped and conflicting directions are amplified. The multiplier stays bounded by $[e^{-\kappa}, e^{\kappa}]$ at every step and equals 1 exactly at $\sigma_i = 0.5$, so this reversal cannot produce unbounded guidance, but it is a genuine change in polarity rather than a negligible effect; Fig. 2b visualises the resulting four quadrants of the temporal–geometric multiplier. Setting $\kappa = 0$ recovers fixed CFG exactly. The exponential form makes the adaptation multiplicative rather than an absolute schedule, keeping the base scale interpretable while letting local velocity geometry reshape the trajectory.

VAGS for image editing. Given a source image x_{src}, source prompt c_{src}, and target prompt c_{tar}, inversion-free editing couples noisy source and target latents as

$$\hat{z}^{\mathrm{src}}_{t_i} = (1 - t_i)x_{\mathrm{src}} + t_i\epsilon_i, \qquad \hat{z}^{\mathrm{tar}}_{t_i} = z^{\mathrm{edit}}_{t_i} + \hat{z}^{\mathrm{src}}_{t_i} - x_{\mathrm{src}}, \tag{4}$$

with $\epsilon_i \sim \mathcal{N}(0, I)$, then advances the edit by

$$z^{\mathrm{edit}}_{t_{i-1}} = z^{\mathrm{edit}}_{t_i} + (t_{i-1} - t_i)\big(V^{\mathrm{tar}}_{t_i} - V^{\mathrm{src}}_{t_i}\big). \tag{5}$$

VAGS keeps the source scale fixed and adapts only the target scale, since the target field controls the requested semantic change.

This choice separates the two roles in the editing update. The source-guided velocity acts as the preservation anchor: it describes how the original image would move under its own prompt. The target-guided velocity supplies the requested semantic change. Adapting the source scale would move the anchor itself, making preservation changes harder to attribute. We therefore keep the source branch fixed and ask whether the target branch is locally compatible with that anchor.

The editing alignment signal is the cosine between the source guided velocity and a one-shot pilot target velocity. Let $u^{\mathrm{src}}_i, p^{\mathrm{src}}_i$ and $u^{\mathrm{tar}}_i, p^{\mathrm{tar}}_i$ be unconditional and conditional raw predictions for the source and target latents. We form

$$\lambda^{\mathrm{tar,base}}_i = \lambda_{\mathrm{tar}} \exp\big(\kappa_{\mathrm{tar}}(2\sigma_i - 1)\big), \qquad V^{\mathrm{tar,pilot}}_{t_i} = u^{\mathrm{tar}}_i + \lambda^{\mathrm{tar,base}}_i(p^{\mathrm{tar}}_i - u^{\mathrm{tar}}_i), \tag{6}$$

and then set

$$s^{\mathrm{edit}}_i = \frac{V^{\mathrm{src}}_{t_i} \cdot V^{\mathrm{tar,pilot}}_{t_i}}{\|V^{\mathrm{src}}_{t_i}\| \, \|V^{\mathrm{tar,pilot}}_{t_i}\|}, \qquad \lambda^{\mathrm{tar}}_i = \lambda_{\mathrm{tar}} \exp\Big(\kappa_{\mathrm{tar}}(2\sigma_i - 1)s^{\mathrm{edit}}_i\Big). \tag{7}$$

The pilot avoids iterative recomputation, and all quantities come from the same four-way batch already used by the editing baseline.

Why cosine similarity. Cosine similarity depends only on the angle between $V^{\mathrm{src}}_{t_i}$ and $V^{\mathrm{tar,pilot}}_{t_i}$, not their magnitude, which matters because a norm-based alternative such as $\|V^{\mathrm{tar,pilot}}_{t_i} - V^{\mathrm{src}}_{t_i}\|$ conflates how much the trajectory changes with which direction it changes in, while only the latter is informative about local stability. $V^{\mathrm{src}}_{t_i}$ is the direction the sample would follow under pure preservation, and $V^{\mathrm{tar,pilot}}_{t_i}$ is the direction the edit requests, so s^{edit}_i measures whether the requested displacement is geometrically compatible with staying close to the source or opposes it outright. Crucially, the same low s^{edit}_i admits two readings that depend on spatial location: outside the edited region it predicts corruption, while inside the edited region it is the expected signature of the edit succeeding, since a real edit necessarily disagrees with pure preservation somewhere. This is why Eq. 7 uses a multiplicative exponential rather than a hard gate on s^{edit}_i: it tolerates moderate in-mask disagreement

while still damping the extreme late-stage cases in which disagreement manifests as a visible artifact rather than a successful edit. Sec. D.4 decomposes s_i^{edit} spatially inside and outside the edit mask and correlates each component with the corresponding preservation and edit success metrics.

VAGS for Text-to-Image Generation. For generation, the latent starts from $z_{t_N} \sim \mathcal{N}(0, I)$ and follows

$$z_{t_{i-1}} = z_{t_i} + (t_{i-1} - t_i) V^{\lambda}\big(z_{t_i}, t_i, c\big). \tag{8}$$

There is no source field, so VAGS-Gen uses the unconditional and conditional raw velocities already required by CFG:

$$s_i^{\text{gen}} = \frac{v_{\varnothing}^{(i)} \cdot v_c^{(i)}}{\|v_{\varnothing}^{(i)}\| \, \|v_c^{(i)}\|}, \qquad \lambda_i = \lambda \exp\Big(\kappa(2\sigma_i - 1) s_i^{\text{gen}}\Big). \tag{9}$$

Each step therefore performs the same two-way unconditional/conditional batch as vanilla CFG, followed by one cosine similarity and scalar re-weighting. VAGS-Gen is a drop-in replacement for fixed CFG with no auxiliary network, fine-tuning, or extra forward pass.

4 Experiments

Evaluation scope. The evaluation tests VAGS on two uses of classifier-free guidance: inversion-free editing and text-to-image generation. The main editing criterion is the preservation/editability trade-off, while the generation study tests whether the same velocity-alignment signal improves sample quality without weakening prompt alignment. Scheduler and matched-mean ablations isolate adaptive guidance from simple changes to the average CFG scale, and κ sweeps measure modulation sensitivity.

Datasets and metrics. For editing, we evaluate on PIE-Bench (Ju et al., 2023) and DIV2K (Agustsson & Timofte, 2017). PIE-Bench provides 700 image-prompt pairs across ten edit categories with foreground masks, and DIV2K stresses preservation on high-resolution natural images with fine texture and detail. Editing metrics cover structure distance (Tumanyan et al., 2023), background PSNR/LPIPS (Zhang et al., 2018)/MSE/SSIM, and whole-image and edited-region CLIP similarity (Wu et al., 2021). For generation, we use COCO17 (Lin et al., 2014), CUB-200 (Welinder et al., 2010), and Flickr30K (Plummer et al., 2015), spanning broad captions, fine-grained bird species, and everyday scenes. We report FID (Heusel et al., 2017), IS (Salimans et al., 2016), CLIPScore (Hessel et al., 2021), and caption metrics BLEU-4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), and CLAIR.

Implementation. All experiments use Stable Diffusion 3.5 Large. Images are resized and center-cropped to 512×512 before encoding. Runs use PyTorch (Paszke et al., 2019) and `diffusers` (von Platen et al., 2022) on one NVIDIA H100 80GB GPU. For editing, VAGS is plugged into FlowEdit (Kulikov et al., 2024) and SplitFlow (Yoon et al., 2025) with $N = 50$, $N_{\max} = 33$, $\lambda_{\text{src}} = 3.5$, $\lambda_{\text{tar}} = 13.5$, and $\kappa_{\text{tar}} = 0.9$. For generation, VAGS-Gen uses $N = 25$, base CFG $\lambda = 7.0$, $\kappa = 1.0$ on COCO17 and CUB-200, and $\kappa = 0.9$ on Flickr30K.

Baselines. Editing baselines cover diffusion inversion and attention editors, including DDIM+P2P (Song et al., 2021a; Hertz et al., 2023), DDIM+PnP (Song et al., 2021a; Tumanyan et al., 2023), and Null-text+P2P (Mokady et al., 2023; Hertz et al., 2023); rectified-flow inversion methods (Rout et al., 2024; Wang et al., 2024a; Deng et al., 2024; Xu et al., 2025); and inversion-free editors (Kulikov et al., 2024; Yoon et al., 2025; Yang et al., 2025). Scheduler ablations compare Interval (Kynkäänniemi et al., 2024b), Monotone (Wang et al., 2024b), and Zero-Init (Fan et al., 2025) across both editors, plus C^{2}FG (Gao et al., 2026) and β-CFG (Malarz et al., 2025) for FlowEdit. For generation, we compare fixed CFG SDv3.5, A-Euler (Jin et al., 2025), and Self-Guidance (Li et al., 2025b).

Quantitative Evaluation for Editing. Tab. 1 shows that VAGS substantially improves preservation with competitive edit alignment. On PIE-Bench, FlowEdit+VAGS reduces Dist from 30.31 to 13.84 and MSE by 62%; CLIP-I rises from 83.69 to 87.44, while Whole/Edited CLIP decrease modestly from 27.62/23.80 to 26.92/23.08. SplitFlow+VAGS follows the same preservation trend (Dist $28.03 \rightarrow 14.67$, Tab. 3). On DIV2K, VAGS ranks first on all five preservation metrics. Thus, velocity alignment shifts the preservation and edit-alignment trade-off toward better structural fidelity.

Dataset	Method	Model	Structure	Background Preservation					CLIP Similarity	
			Dist. $\times 10^3 \downarrow$	PSNR $\uparrow$	LPIPS $\times 10^3 \downarrow$	MSE $\times 10^4 \downarrow$	SSIM $\times 10^2 \uparrow$	CLIP-I $\uparrow$	Whole $\uparrow$	Edited $\uparrow$
PIE-Bench	*Diffusion-based*									
	DDIM Song et al. (2021a) + P2P Hertz et al. (2023)	SD 1.4	71.12	17.84	207.62	220.79	71.62	82.44	25.91	22.75
	DDIM Song et al. (2021a) + PnP Tumanyan et al. (2023)	SD 1.4	27.27	22.36	110.22	81.94	79.65	86.93	26.13	22.79
	Null-text Mokady et al. (2023) + P2P Hertz et al. (2023)	SD 1.4	15.69	**26.92**	**66.42**	35.09	82.67	85.42	25.11	22.01
	Flow-based / Rectified Flow									
	RF-Inversion Rout et al. (2024)	FLUX.1	47.19	20.64	193.68	123.33	76.34	80.23	26.16	22.38
	RF-Solver Wang et al. (2024a)	FLUX.1	48.04	20.52	181.15	128.54	77.81	82.20	26.63	23.46
	FireFlow Deng et al. (2024) + RF-Solver Wang et al. (2024a)	FLUX.1	25.88	22.90	129.27	74.73	82.03	84.76	25.96	22.02
	FlowEdit Kulikov et al. (2024)	FLUX.1	27.93	21.74	113.83	98.75	83.46	86.54	26.19	22.19
	FTEdit Xu et al. (2025) + AdaLN	SD 3.5	**13.81**	23.69	77.69	36.10	**92.62**	**92.40**	25.01	21.81
	iRFDS Yang et al. (2025)	SD 3.5	67.64	19.42	150.68	147.86	79.47	77.69	25.45	21.73
	FlowEdit Kulikov et al. (2024)	SD 3.5	30.31	21.87	117.06	92.70	82.54	83.69	**27.62**	**23.80**
	SplitFlow Yoon et al. (2025)	SD 3.5	28.03	22.32	111.49	85.31	83.17	84.10	27.52	23.77
	Ours									
	FlowEdit + VAGS	SD 3.5	13.84	26.38	70.38	**34.86**	87.68	87.44	26.92	23.08
DIV2K	*Diffusion-based*									
	DDIM Song et al. (2021a) + P2P Hertz et al. (2023)	SD 1.4	64.46	15.00	323.57	359.10	60.93	81.89	26.45	-
	DDIM Song et al. (2021a) + PnP Tumanyan et al. (2023)	SD 1.4	30.66	18.75	205.72	154.73	68.61	83.83	28.37	-
	Null-text Mokady et al. (2023) + P2P Hertz et al. (2023)	SD 1.4	15.82	21.35	137.28	93.17	74.37	88.32	27.61	-
	Flow-based / Rectified Flow									
	RF-Inversion Rout et al. (2024)	FLUX.1	31.87	20.05	252.75	115.45	73.37	79.57	29.60	-
	RF-Solver Wang et al. (2024a)	FLUX.1	34.63	18.56	265.56	161.60	72.98	79.54	29.90	-
	FireFlow Deng et al. (2024) + RF-Solver Wang et al. (2024a)	FLUX.1	20.52	21.02	197.26	92.75	77.88	82.86	28.77	-
	FlowEdit Kulikov et al. (2024)	FLUX.1	24.84	18.73	173.21	152.33	78.81	83.93	29.07	-
	FTEdit Xu et al. (2025) + AdaLN	SD 3.5	14.43	23.02	125.97	56.65	83.73	**88.97**	27.03	-
	iRFDS Yang et al. (2025)	SD 3.5	61.14	15.86	317.41	320.07	63.02	76.79	27.47	-
	FlowEdit Kulikov et al. (2024)	SD 3.5	18.27	20.27	145.55	103.37	82.15	82.31	**30.06**	-
	SplitFlow Yoon et al. (2025)	SD 3.5	17.61	20.53	141.99	97.19	82.63	82.44	29.95	-
	Ours									
	FlowEdit + VAGS	SD 3.5	**8.62**	**23.61**	**98.43**	**50.22**	**84.41**	84.07	29.61	-

Table 1: Quantitative comparison on PIE-Bench and DIV2K. FlowEdit + VAGS ranks first on all background preservation metrics while maintaining competitive edit strength.

Dataset	Method	FID $\downarrow$	IS $\uparrow$	CLIPScore $\uparrow$	Caption-based Metrics			
					BLEU-4 $\uparrow$	METEOR $\uparrow$	ROUGE-L $\uparrow$	CLAIR $\uparrow$
COCO17 Lin et al. (2014)	SDv3.5 (CFG=7)	28.46	33.64	**32.84**	7.99	29.17	35.11	71.45
	SDv3.5 w/ A-Euler Jin et al. (2025)	27.64	34.63	**32.84**	7.96	29.09	35.17	71.23
	SDv3.5 w/ Self-Guidance Li et al. (2025b)	40.98	28.54	31.21	6.52	26.18	31.59	62.26
	SDv3.5 w/ VAGS (CFG=7, κ=1.0)	**26.07**	**35.15**	**32.84**	**8.28**	**29.53**	**35.21**	**71.62**
CUB-200 Welinder et al. (2010)	SDv3.5 (CFG=7)	24.92	4.81	32.40	0.13	16.57	17.64	**66.69**
	SDv3.5 w/ A-Euler Jin et al. (2025)	22.98	5.05	**32.50**	0.12	16.39	17.51	66.22
	SDv3.5 w/ Self-Guidance Li et al. (2025b)	61.16	4.98	31.25	**0.18**	**17.85**	**17.78**	62.28
	SDv3.5 w/ VAGS (CFG=7, κ=1.0)	**20.60**	**5.17**	32.45	0.15	17.00	17.52	66.52
Flickr30K Plummer et al. (2015)	SDv3.5 (CFG=7)	79.58	17.95	**33.79**	4.12	**23.88**	**29.30**	**68.03**
	SDv3.5 w/ A-Euler Jin et al. (2025)	78.57	17.83	33.71	3.91	22.99	29.00	67.56
	SDv3.5 w/ Self-Guidance Li et al. (2025b)	93.07	14.44	31.60	3.35	20.80	26.27	57.35
	SDv3.5 w/ VAGS (CFG=7, κ=0.9)	**75.65**	**18.21**	33.77	**4.17**	23.43	29.25	67.65

Table 2: Text-to-image generation on COCO17, CUB-200, and Flickr30K. BLIP-2 Li et al. (2023) is used to generate captions from the generated images, enabling evaluation with the caption-based metrics. VAGS-Gen yields consistent FID and IS gains.

Because a trivial CFG reduction usually improves preservation at the cost of edit strength, these large preservation gains alongside modest semantic changes support step-level regulation rather than global weakening.

Quantitative Evaluation Generation. Tab. 2 evaluates VAGS-Gen on three datasets. On COCO17, VAGS-Gen improves FID from 28.46 to 26.07 and IS from 33.64 to 35.15, outperforming fixed CFG, A-Euler, and Self-Guidance. The gain is larger on CUB-200, where FID drops from 24.92 to 20.60, and remains positive on Flickr30K (79.58 to 75.65). Caption metrics stay competitive, indicating that the FID/IS gains do not come from ignoring the prompt. Appendix trajectory diagnostics show that VAGS-Gen reshapes the CFG path using unconditional/conditional velocity alignment, while prompt-wise mean scales remain concentrated.

Scheduler ablation. Tab. 3 isolates the role of velocity-aware adaptivity. Interval scheduling gives small gains, Monotone improves preservation but underperforms VAGS, and Zero-Init collapses structure. VAGS provides the strongest overall trade-off among schedules evaluated for both editors. Matched-mean controls in Secs. D.4 and D.6 further show that VAGS is not merely a lower or higher average CFG scale: editing with fixed CFG 14.05 remains near the FlowEdit baseline (Dist 30.55 vs. 30.31), while VAGS reaches 13.84; generation with fixed CFG 5.955 only improves FID from 28.46 to 28.28, far short of VAGS-Gen at 26.07.

Dataset	Editor	Scheduler	Structure	Background Preservation					CLIP Similarity	
			Dist. $\times 10^3 \downarrow$	PSNR $\uparrow$	LPIPS $\times 10^3 \downarrow$	MSE $\times 10^4 \downarrow$	SSIM $\times 10^2 \uparrow$	CLIP-I $\uparrow$	Whole $\uparrow$	Edited $\uparrow$
PIE-Bench	FlowEdit	No Scheduler	30.31	21.87	117.06	92.70	82.54	83.69	27.62	23.80
		Interval Kynkäänniemi et al. (2024b)	29.82	21.93	114.14	91.72	82.88	84.03	27.58	23.75
		Monotone Wang et al. (2024b)	15.91	25.43	**69.87**	43.47	87.67	84.73	27.05	23.26
		Zero-Init Fan et al. (2025)	194.67	11.69	354.57	867.59	57.64	75.28	**28.44**	**24.24**
		C^2FG Gao et al. (2026)	76.98	13.84	500.49	456.78	48.61	78.71	27.42	27.37
		β-CFG Malarz et al. (2025)	14.07	21.82	185.25	79.51	80.11	88.76	26.73	26.53
		VAGS (Ours)	**13.84**	**26.38**	70.38	**34.86**	**87.68**	**87.44**	26.92	23.08
	SplitFlow Yoon et al. (2025)	No Scheduler	28.03	22.32	111.49	85.31	83.17	84.10	**27.52**	**23.77**
		Interval Kynkäänniemi et al. (2024b)	27.28	23.05	106.26	77.64	84.17	83.25	27.25	23.22
		Monotone Wang et al. (2024b)	16.06	**26.25**	**66.73**	38.77	**88.12**	85.94	26.83	22.76
		Zero-Init Fan et al. (2025)	21.65	23.57	94.01	65.07	84.90	85.36	27.30	23.57
		VAGS (Ours)	**14.67**	25.99	73.43	**38.41**	87.31	**87.22**	27.06	23.28
DIV2K	FlowEdit	No Scheduler	18.27	20.27	145.55	103.37	82.15	82.31	**30.06**	-
		Interval Kynkäänniemi et al. (2024b)	18.11	19.94	144.88	111.55	79.51	82.32	30.03	-
		Monotone Wang et al. (2024b)	10.89	22.47	102.91	63.91	83.64	**84.10**	29.60	-
		Zero-Init Fan et al. (2025)	194.67	11.69	354.57	867.59	57.64	75.28	28.44	-
		C^2FG Gao et al. (2026)	38.14	16.50	355.19	242.93	65.94	79.37	30.34	-
		β-CFG Malarz et al. (2025)	10.15	23.33	148.65	53.93	86.55	85.32	29.16	-
		VAGS (Ours)	**8.62**	**23.61**	**98.43**	**50.22**	**84.41**	84.07	29.61	-
	SplitFlow Yoon et al. (2025)	No Scheduler	17.61	20.53	141.99	97.19	82.63	82.44	29.95	-
		Interval Kynkäänniemi et al. (2024b)	16.01	20.53	136.04	97.62	80.55	82.65	29.91	-
		Monotone Wang et al. (2024b)	9.63	23.01	**97.16**	56.81	**84.23**	**84.43**	29.55	-
		Zero-Init Fan et al. (2025)	46.72	14.83	378.94	345.27	48.95	75.93	**31.58**	-
		VAGS (Ours)	**9.32**	**23.28**	102.58	**53.57**	84.01	84.08	29.61	-

Table 3: Ablation of guidance scheduling strategies on FlowEdit and SplitFlow. VAGS gives the strongest overall preservation and edit-quality trade-off, improving substantially over constant CFG and prior adaptive schedules.

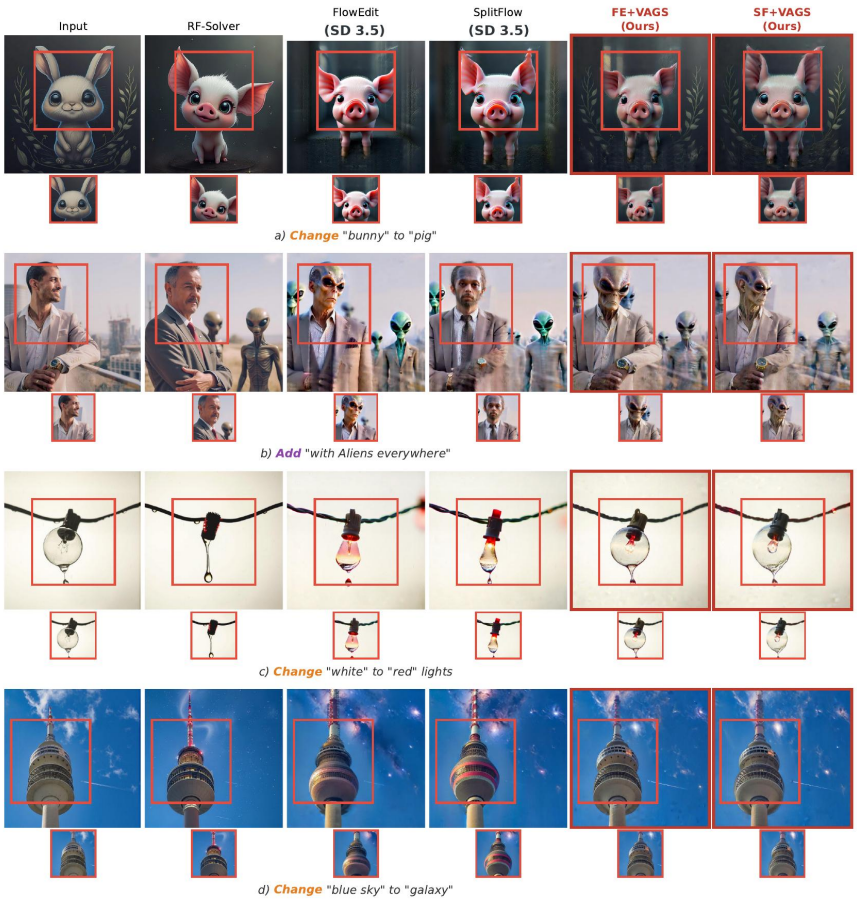


Figure 3: Qualitative editing comparison on PIE-Bench. Columns show, from left to right: Input, RF-Solver, FlowEdit (SD 3.5), SplitFlow (SD 3.5), FE+VAGS (Ours), and SF+VAGS (Ours). Fixed-guidance editors often corrupt backgrounds or under-edit the target region, while VAGS produces sharper semantic changes with stronger preservation of unedited content.

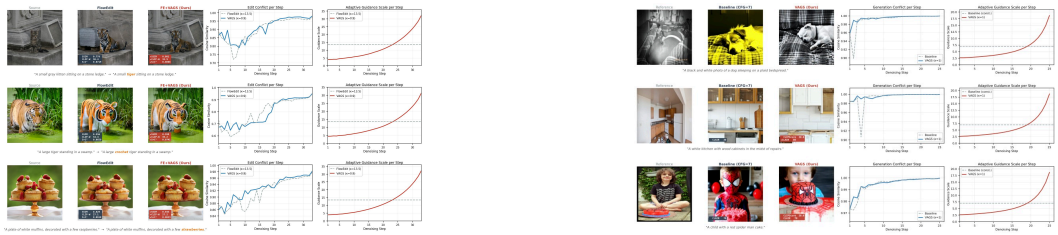


Figure 4: Editing (left) and generation (right) trajectories with per-step cosine similarity and adaptive guidance scale.

CLIP-scaled tables in Sec. D.8 preserve the same conclusions after penalizing methods whose preservation or fidelity gains come with weaker text-image alignment.

Qualitative results on editing. Fig. 3 shows representative PIE-Bench and DIV2K edits. Fixed guidance often under-edits or corrupts backgrounds, while time-only schedules still drift on fine texture. VAGS produces sharper target changes with stronger source preservation, matching the quantitative trends. The editing panel of Fig. 4 links these gains to step-wise source/target alignment; aggregate diagnostics are provided in Sec. D.4.

Qualitative results on generation. Qualitative generation examples are provided in Fig. 6 and Sec. D.6. VAGS-Gen better preserves object counts, fine-grained species cues, and compositional layout than fixed CFG and the training-free guidance baselines. The generation panel of Fig. 4 shows the analogous unconditional/conditional alignment, while appendix progressive comparisons distinguish VAGS-Gen from static CFG even at a matched mean scale.

Analysis of modulation strength. Generation and editing κ ablations are reported in Tabs. 7 and 8. Moderate modulation improves FID and editing preservation, while overly large values suppress or amplify guidance too aggressively. We therefore use $\kappa = 1.0$ for COCO17/CUB-200, $\kappa = 0.9$ for Flickr30K, and $\kappa_{\text{tar}} = 0.9$ for editing. Full plots are in Sec. D.7. The chosen values prioritize balanced quality: larger values can improve isolated fidelity metrics, but caption or preservation metrics degrade under stronger modulation.

Time complexity & computational cost. The dominant cost of every flow-based sampler considered here is the per-step forward pass through the velocity network v_θ, whose cost we denote T_{net}; all other operations are $\mathcal{O}(d)$ in the latent dimension d and negligible by comparison. For editing, the inversion-free baseline performs a single batched forward pass of size four per step, covering unconditional and conditional predictions for both source and target prompts. VAGS reuses these raw predictions to assemble the pilot target velocity, compute one cosine similarity, and re-weight the target velocity in closed form. The total complexity therefore remains $\mathcal{O}(N\,T_{\text{net}})$ with no extra network evaluations.

For generation, constant CFG performs one batched forward pass of size two per step. VAGS-Gen computes s_i^{gen} directly from the same two raw predictions, again adding only one inner product per step. In measured runtime on a single NVIDIA H100, editing takes $7.5\,\text{s}$ per image for the baseline and $7.6\,\text{s}$ for VAGS, a 1.3% overhead. Generation takes $11.3\,\text{s}$ for the baseline and $11.5\,\text{s}$ for VAGS-Gen, a 1.8% overhead.

5 Conclusion

We presented VAGS, a training-free replacement for fixed CFG in flow-based editing and generation. Rather than holding guidance constant, VAGS combines a temporal signal-level term with local velocity alignment, amplifying guidance where the sampler's dynamics support it and restraining it where stronger extrapolation would work against committed structure. The method is plug-and-play, adds no forward passes, and delivers consistent gains in editing fidelity and generation quality across five benchmarks. The broader implication is clear: guidance in flow models is not a global scalar to be chosen before sampling begins. It is a local control problem over velocities, one that can and should be solved at every step.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR*, 2017.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*, 2005.

Black Forest Labs. FLUX.1. `https://blackforestlabs.ai`, 2024. Open-weight rectified-flow text-to-image model.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Angela Castillo, Jonas Kohler, Juan C Pérez, Juan Pablo Pérez, Albert Pumarola, Bernard Ghanem, Pablo Arbeláez, and Ali Thabet. Adaptive guidance: Training-free acceleration of conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1962–1970, 2025.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Hansam Cho and Seoung Bum Kim. Improving diffusion-based image editing faithfulness via guidance and scheduling. *arXiv preprint arXiv:2506.21045*, 2025.

Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. FireFlow: Fast inversion of rectified flow for image semantic editing. *arXiv preprint arXiv:2412.07517*, 2024.

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. *arXiv preprint arXiv:2403.03206*, 2024.

Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

Alexandre Galashov, Ashwini Pokle, Arnaud Doucet, Arthur Gretton, Mauricio Delbracio, and Valentin De Bortoli. Learn to guide your diffusion model. *arXiv preprint arXiv:2510.00815*, 2025.

Jiayang Gao, Tianyi Zheng, Jiayang Zou, Fengxiang Yang, Shice Liu, Luyao Fan, Zheyu Zhang, Hao Zhang, Jinwei Chen, Peng-Tao Jiang, Bo Li, and Jia Wang. C^2FG: Control classifier-free guidance via score discrepancy analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34398–34407, 2026.

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 *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 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 *Adv. Neural Inform. Process. Syst.*, 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. *NeurIPS*, 2020.

Cheng Jin, Zhenyu Xiao, and Yuantao Gu. A-FloPS: Accelerating diffusion models via adaptive flow path sampler. *arXiv preprint arXiv:2509.00036*, 2025. Also appeared in AAAI 2026.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023.

Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic negative guidance of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. FlowEdit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arXiv preprint arXiv:2404.07724*, 2024a.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *NeurIPS*, 2024b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Pengxiang Li, Shilin Yan, Joey Tsai, Renrui Zhang, Ruichuan An, Ziyu Guo, and Xiaowei Gao. Adaptive classifier-free guidance via dynamic low-confidence masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Tiancheng Li, Weijian Luo, Zhiyang Chen, Liyuan Ma, and Guo-Jun Qi. Self-guidance: Boosting flow and diffusion generation on their own. *IEEE TPAMI*, 2025b.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL Workshop*, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

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.

Dawid Malarz, Artur Kasymov, Maciej Zięba, Jacek Tabor, and Przemysław Spurek. Classifier-free guidance with adaptive scaling. *arXiv preprint arXiv:2502.10574*, 2025.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023.

Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajić, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. *arXiv preprint arXiv:2509.16131*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

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

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Shreshth Saini, Shashank Gupta, and Alan C Bovik. Rectified-cfg++ for flow based models. *arXiv preprint arXiv:2510.07631*, 2025.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Adv. Neural Inform. Process. Syst.*, 2016.

Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9370–9379, 2024.

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

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*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Bram Wallace, Akash Gokul, and Nikhil Naik. EDICT: Exact diffusion inversion via coupled transformations. In *CVPR*, 2023.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024a.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *arXiv preprint arXiv:2404.13040*, 2024b.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

Pengcheng Xu, Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, Charles Ling, and Boyu Wang. Unveil inversion and invariance in flow transformer for versatile image editing. *arXiv preprint arXiv:2411.15843*, 2025.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. *arXiv preprint arXiv:2312.04965*, 2023.

Xiaofeng Yang, Cheng Chen, Xulei Yang, Fayao Liu, and Guosheng Lin. Text-to-image rectified flow as plug-and-play priors. *arXiv preprint arXiv:2406.03293*, 2025.

Sung-Hoon Yoon, Minghan Li, Gaspard Beaudouin, Congcong Wen, Muhammad Rafay Azhar, and Mengyu Wang. Splitflow: Flow decomposition for inversion-free text-to-image editing. *arXiv preprint arXiv:2510.25970*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

A USE OF LARGE LANGUAGE MODELS

We used large language models for grammar checking and refining the text of this paper. The authors take full responsibility for the paper's contents.

B REPRODUCIBILITY STATEMENT

Sec. 4 (Implementation) lists the base model, datasets, GPU, and all VAGS hyperparameters (N, $N_{\max}$, λ_{src}, λ_{tar}, κ_{tar}, and κ) used for every reported result. Algs. 1 and 2 in Sec. D.2 give complete step-by-step pseudocode for the editing and generation variants, and Sec. D.3 records the formulas and verified FlowEdit hyperparameters for the Interval, Monotone, and Zero-Init schedulers, together with the DIV2K evaluation protocol. Matched-mean controls (Sec. D.4, Sec. D.6) and the κ sensitivity sweeps (Sec. D.7) isolate step-level adaptivity as the source of the reported gains. Code is withheld during review to preserve anonymity, as stated in the abstract, and will be made publicly available upon publication.

C ETHICS STATEMENT

No personal or human-subject data is used in this work.

D TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

This appendix collects implementation pseudocode, diagnostic evidence, and secondary quantitative tables. The goal is to support the main claim without interrupting the paper flow: VAGS works because it changes the shape of the guidance trajectory in response to velocity alignment, not because it simply changes the average CFG value.

Appendix contents.

- Sec. D.1: adaptive scale with two axes.
- Sec. D.2: editing and generation pseudocode.
- Sec. D.4: editing traces, alignment diagnostics, and matched-mean controls.
- Sec. D.5: operational definition and measurement of semantic drift.
- Sec. D.6: generation examples, trajectories, and matched-mean controls.
- Sec. D.7: sensitivity to modulation strength.
- Sec. D.8: CLIP-scaled metric tables.

D.1 ADAPTIVE SCALE WITH TWO AXES.

VAGS resolves both issues by replacing the fixed target scale λ with a step-dependent scale λ_i that is jointly modulated by a *temporal* factor and a *geometric* factor, both computed from quantities the sampler already produces. Let $\sigma_i = 1 - t_i \in [0, 1]$ be the *signal level* at step i, which grows monotonically from 0 (pure noise) to 1 (clean image), and let $s_i \in [-1, 1]$ be a *velocity-alignment* signal, defined as the cosine similarity between two task-specific velocity fields available at step i (made concrete in Sec. 3). VAGS sets

$$\lambda_i \;=\; \lambda \cdot \exp\Big(\kappa \, (2\sigma_i - 1) \, s_i \Big), \tag{10}$$

where $\kappa \geq 0$ controls the modulation strength. The two factors play complementary roles: the temporal factor encodes *which signal to trust* at the current stage of sampling, and the geometric factor encodes *whether the local velocity geometry corroborates that trust enough to warrant a correction* (Fig. 2).

- **Temporal axis: $(2\sigma_i - 1)$.** Negative in the noisy regime ($\sigma_i < \frac{1}{2}$) and positive in the clean regime ($\sigma_i > \frac{1}{2}$), this signed factor passes through zero at $\sigma_i = \frac{1}{2}$, where λ_i recovers the nominal scale exactly. Its sign reflects a qualitative shift in the sampler's task across the trajectory. Early steps perform *basin selection*: from a near-isotropic latent the sampler must commit to a region of the data manifold, and the unconditional prior is the more reliable signal because raw conditional velocities are dominated by noise. Late steps perform *mode refinement*: coarse content is fixed, and the prompt-conditioned signal must sharpen semantic detail without disturbing committed structure. The monotone ramp is consistent with prior empirical findings that increasing CFG schedules outperform constant guidance (Chang et al., 2023; Wang et al., 2024b), but VAGS treats it not as a magnitude schedule alone, rather, as a *regime indicator* that flips how the geometric axis should be read.
- **Geometric axis: s_i.** The cosine similarity between the two velocity fields is invariant to the magnitude fluctuations of the raw predictions across timesteps and prompts, isolating their angular relationship. Critically, the *informative event* differs between regimes. In the clean regime, where raw velocities are high-fidelity, agreement ($s_i > 0$) means the prompt-conditioned direction is consistent with established structure: amplifying guidance sharpens the desired direction without disturbing unrelated content. Disagreement ($s_i < 0$) means guidance is fighting committed structure and should be damped to avoid overrides or off-manifold drift. The polarity inverts in the noisy regime, where s_i is itself a low-fidelity quantity computed from two noisy vectors. The modulation vanishes at the trajectory midpoint and grows with distance from it, reaching the same bounded range near the noisy and clean endpoints. Within this bounded envelope, agreement indicates the prompt is broadly redundant with the unconditional prior (high λ buys little adherence at the cost of amplifying noise, motivating slight attenuation), while systematic disagreement is the most plausible early evidence that the prompt targets a region of the manifold the unconditional default would not naturally select, warranting a small early commitment toward that region.

The product $(2\sigma_i - 1)\, s_i$ is the indicator that implements this regime-dependent reading of s_i. Its sign across the four (σ_i, s_i) quadrants is summarised in Table 4, together with the sampler's local objective.

Table 4: Behaviour of the multiplicative correction across regimes. Disagreement carries different information depending on which signal the sampler should trust at the current stage; the temporal factor flips the polarity, and the exponential bound controls the correction in both regimes.

Regime	**Role of disagreement** $(s_i < 0)$	**Effect on** λ_i
Noisy ($\sigma_i < \frac{1}{2}$): basin selection	Plausible signal of required basin shift	Amplification; agreement attenuates
Clean ($\sigma_i > \frac{1}{2}$): mode refinement	Conflict with committed structure	Attenuation; agreement amplifies

This design is well suited to the training-free, plug-and-play setting in which VAGS operates. The properties it relies on—that the unconditional prior dominates trajectory shape under high noise, that prompt-conditioned signals dominate under low noise, and that high CFG late in sampling produces

off-manifold artefacts when guidance fights structure—are documented behaviours of pretrained flow models (Chang et al., 2023; Wang et al., 2024b; Fan et al., 2025; Saini et al., 2025), not behaviours VAGS induces. The temporal axis selects which of these documented regimes the sampler is in; the geometric axis reads the local velocity geometry to gate the correction. Crucially, in the noisy regime where s_i is least reliable, $(2\sigma_i - 1)$ is small in magnitude, so the multiplier stays close to one and the cost of misreading s_i is bounded by construction. In the clean regime where the geometric reading is most trustworthy, $(2\sigma_i - 1)$ approaches its maximum, allowing the correction to take full effect exactly where it is most reliable. The training-free nature of VAGS is therefore not a limitation that the design tolerates; it is what the design is calibrated for.

Two structural properties of Eq. (3) are worth noting. First, because s_i is bounded in $[-1, 1]$, the multiplier is bounded in $[\, e^{-\kappa}, e^{\kappa}\,]$, so a single κ controls the entire range of admissible scales and the final λ_i is not guaranteed to be monotone in i—adaptivity is traded for strict monotonicity by design. Second, setting $\kappa = 0$ yields $\lambda_i \equiv \lambda$, recovering the fixed-CFG baseline exactly; VAGS is therefore a strict generalisation. The remaining task is to specify the velocity pair from which s_i is computed in each setting.

D.2 Pseudocode

Algorithms 1 and 2 expand the compact description in Sec. 3. Both procedures reuse the raw velocity predictions required by the underlying CFG sampler; VAGS only adds cosine similarity and scalar re-weighting.

Algorithm 1 Velocity-Adaptive Guidance Scale for Editing (VAGS)

Input: source image x_{src}; prompts $c_{\mathrm{src}}, c_{\mathrm{tar}}$; base scales $\lambda_{\mathrm{src}}, \lambda_{\mathrm{tar}}$; modulation strength κ_{tar}; schedule $\{t_i\}_{i=1}^{N}$ with $0 \approx t_1 < \cdots < t_N \approx 1$.

Output: edited image $z_{t_1}^{\mathrm{edit}}$.

Notation: $\hat{z}_i^{\mathrm{s}} \equiv \hat{z}_{t_i}^{\mathrm{src}}$, $\hat{z}_i^{\mathrm{t}} \equiv \hat{z}_{t_i}^{\mathrm{tar}}$; $V_i^{\mathrm{s}},\ \widetilde{V}_i^{\mathrm{t}},\ V_i^{\mathrm{t}}$ denote source, pilot-target, and adaptive-target guided velocities.

$z_{t_N}^{\mathrm{edit}} \leftarrow x_{\mathrm{src}}$ // initialise editing trajectory

for $i = N, N-1, \ldots, 2$ **do**

/* 1. Couple noisy source/target latents (4) */

$\epsilon_i \sim \mathcal{N}(0, I)$ $\hat{z}_i^{\mathrm{s}} \leftarrow (1 - t_i)\, x_{\mathrm{src}} + t_i\, \epsilon_i$ $\hat{z}_i^{\mathrm{t}} \leftarrow z_{t_i}^{\mathrm{edit}} + \hat{z}_i^{\mathrm{s}} - x_{\mathrm{src}}$

/* 2. One forward pass on the four-way batch */

$\big(u_i^{\mathrm{src}},\, p_i^{\mathrm{src}},\, u_i^{\mathrm{tar}},\, p_i^{\mathrm{tar}}\big) \leftarrow v_\theta\big([\hat{z}_i^{\mathrm{s}}, \hat{z}_i^{\mathrm{s}}, \hat{z}_i^{\mathrm{t}}, \hat{z}_i^{\mathrm{t}}],\, t_i,\, [\varnothing, c_{\mathrm{src}}, \varnothing, c_{\mathrm{tar}}]\big)$

/* 3. Source velocity and pilot target velocity (6) */

$\sigma_i \leftarrow 1 - t_i$ $\lambda_i^{\mathrm{tar,base}} \leftarrow \lambda_{\mathrm{tar}} \exp\big(\kappa_{\mathrm{tar}}(2\sigma_i - 1)\big)$ $V_i^{\mathrm{s}} \leftarrow u_i^{\mathrm{src}} + \lambda_{\mathrm{src}}\,(p_i^{\mathrm{src}} - u_i^{\mathrm{src}})$ $\widetilde{V}_i^{\mathrm{t}} \leftarrow u_i^{\mathrm{tar}} + \lambda_i^{\mathrm{tar,base}}\,(p_i^{\mathrm{tar}} - u_i^{\mathrm{tar}})$

/* 4. Alignment signal and adaptive target scale (7) */

$s_i^{\mathrm{edit}} \leftarrow \langle V_i^{\mathrm{s}},\, \widetilde{V}_i^{\mathrm{t}} \rangle \,/\, \big(\|V_i^{\mathrm{s}}\|\, \|\widetilde{V}_i^{\mathrm{t}}\|\big)$ $\lambda_i^{\mathrm{tar}} \leftarrow \lambda_{\mathrm{tar}} \exp\big(\kappa_{\mathrm{tar}}(2\sigma_i - 1)\, s_i^{\mathrm{edit}}\big)$

/* 5. Re-weight target velocity and advance the ODE (5) */

$V_i^{\mathrm{t}} \leftarrow u_i^{\mathrm{tar}} + \lambda_i^{\mathrm{tar}}\,(p_i^{\mathrm{tar}} - u_i^{\mathrm{tar}})$ $z_{t_{i-1}}^{\mathrm{edit}} \leftarrow z_{t_i}^{\mathrm{edit}} + (t_{i-1} - t_i)\big(V_i^{\mathrm{t}} - V_i^{\mathrm{s}}\big)$

end

return $z_{t_1}^{\mathrm{edit}}$

Algorithm 2 Velocity-Adaptive Guidance Scale for Generation (VAGS-Gen)

Input: prompt c; base guidance scale λ; modulation strength κ; schedule $\{t_i\}_{i=1}^{N}$ with $0 \approx t_1 < \cdots < t_N \approx 1$.

Output: generated image z_{t_1}.

Notation: $v_{\varnothing}^{(i)}$ and $v_c^{(i)}$ denote the unconditional and conditional raw velocity predictions at z_{t_i}; V_i is the CFG-guided velocity used to advance the ODE.

$z_{t_N} \sim \mathcal{N}(0, I)$ // initialise generation trajectory

for $i = N, N-1, \ldots, 2$ **do**

/* 1. One forward pass on the two-way batch */

$\big(v_{\varnothing}^{(i)},\, v_c^{(i)}\big) \leftarrow v_\theta\big([z_{t_i}, z_{t_i}],\, t_i,\, [\varnothing,\, c]\big)$

/* 2. Alignment signal (9) */

$\sigma_i \leftarrow 1 - t_i \;\; s_i^{\text{gen}} \leftarrow \langle v_{\varnothing}^{(i)},\, v_c^{(i)} \rangle \,/\, \big(\|v_{\varnothing}^{(i)}\|\, \|v_c^{(i)}\|\big)$

/* 3. Adaptive guidance scale (9) */

$\lambda_i \leftarrow \lambda \, \exp\big(\kappa(2\sigma_i - 1)\, s_i^{\text{gen}}\big)$

/* 4. Assemble guided velocity and advance the ODE (1), (8) */

$V_i \leftarrow v_{\varnothing}^{(i)} + \lambda_i\big(v_c^{(i)} - v_{\varnothing}^{(i)}\big) \;\; z_{t_{i-1}} \leftarrow z_{t_i} + (t_{i-1} - t_i)\, V_i$

end

return z_{t_1}

D.3 Dataset and baseline scheduler details

DIV2K construction. Unlike PIE-Bench, DIV2K (Agustsson & Timofte, 2017) does not provide edit annotations or foreground masks. We therefore evaluate DIV2K without a mask: background and preservation metrics are computed over the full image rather than a masked region, and edit prompts are constructed directly for each source image rather than drawn from a pre-existing annotation set.

Baseline scheduler formulas. Tab. 3 compares VAGS against three schedule-shape baselines evaluated across FlowEdit and SplitFlow. It additionally reports C^{2}FG (Gao et al., 2026) and β-CFG (Malarz et al., 2025) for FlowEdit, for which results are available. Writing $\tau_i = (N - i)/N$ for normalized progress through the active sampling window ($\tau = 0$ at the noisiest active step, $\tau = 1$ at the cleanest):

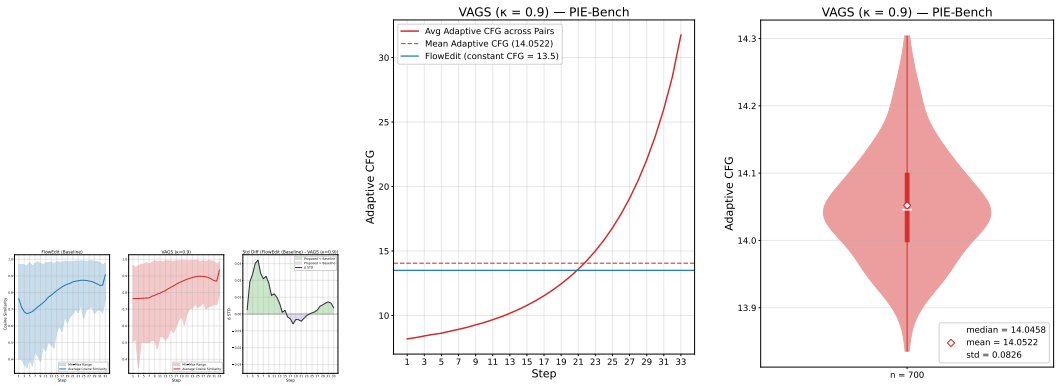
- **Interval** (Kynkäänniemi et al., 2024b) applies the target scale only inside a fixed window and otherwise disables guidance: $\lambda_i = \lambda_{\text{tar}}$ for $0.5 \leq \tau_i \leq 0.7$, and $\lambda_i = 1$ outside it.
- **Monotone** (Wang et al., 2024b) ramps the scale with a cosine curve from the noisiest to the cleanest active step: $\lambda_i = \max\big(1,\ \lambda_{\text{tar}} \cdot \frac{1}{2}(1 - \cos(\pi\tau_i))\big)$.
- **Zero-Init** (Fan et al., 2025) disables guidance only at the first active step ($\lambda_i = 0$) and applies the target scale with star correction at every subsequent step.

For FlowEdit, all three baselines use the same base target scale as VAGS ($\lambda_{\text{tar}} = 13.5$) rather than an independent grid search per scheduler. Corresponding C^{2}FG and β-CFG results are not reported for SplitFlow, so they are not included among the three cross-editor schedule-shape comparisons.

D.4 EDITING DIAGNOSTICS

This section checks whether the editing gains are tied to the intended source/target velocity geometry. Fig. 5 summarises the alignment signal and the resulting adaptive target scale across PIE-Bench. Within this grouped diagnostic, Fig. 5a reports the step-wise cosine signal, Fig. 5b reports the step-wise adaptive target scale, and Fig. 5c reports the pair-wise mean scale.

Matched-mean editing control. Tab. 5 compares VAGS with a constant target scale set to the mean adaptive value ($\bar{\lambda} = 14.05$). The matched constant scale stays close to the baseline, while VAGS substantially improves preservation, confirming that the gain comes from step-wise modulation.



(a) Step-wise alignment (b) Step-wise target CFG (c) Pair-wise mean CFG

Figure 5: Editing diagnostics on PIE-Bench. The cosine signal tracks the source/target velocity relation, while the resulting VAGS target scale is stable across image-prompt pairs and rises toward clean-image steps.

Dataset	Method	Dist. $\times 10^3 \downarrow$	PSNR ↑	LPIPS $\times 10^3 \downarrow$	MSE $\times 10^4 \downarrow$	SSIM $\times 10^2 \uparrow$	CLIP-I ↑	CLIP Similarity	
								Whole ↑	Edited ↑
PIE-Bench	FlowEdit (CFG=13.5)	30.31	21.87	117.06	92.70	82.54	83.69	**27.62**	23.80
	FlowEdit (CFG=14.05, κ=0)	30.55	21.83	117.40	93.10	82.50	83.65	27.60	**23.82**
	FlowEdit + VAGS (CFG=13.5, κ=0.9)	**13.84**	**26.38**	**70.38**	**34.86**	**87.68**	**87.44**	26.92	23.08

Table 5: Constant-CFG vs. VAGS editing on PIE-Bench. A fixed scale at the mean adaptive value ($\bar{\lambda} = 14.05$) performs on par with the baseline, confirming that the gains stem from dynamic modulation rather than a shifted average guidance level.

D.5 OPERATIONALIZING SEMANTIC DRIFT

The main text refers to constant guidance scales accumulating *semantic drift* along the sampling trajectory. We operationalize this term as three measurable modes rather than a single umbrella quantity. *Spatial/background drift* covers pixel or feature changes outside the edit mask; it is already captured by the background LPIPS, SSIM, PSNR, and MSE metrics reported in Tab. 1, here explicitly reframed as drift indicators, together with a background feature distance $\|\phi_{\mathrm{DINOv2}}(x_{\mathrm{src}} \odot (1 - M)) - \phi_{\mathrm{DINOv2}}(x_{\mathrm{edit}} \odot (1 - M))\|_2$. *Count/compositional drift* covers unintended changes in the number of detected objects in the preserved region, measured via an open-vocabulary detector applied before and after editing. *Identity/structural drift* covers the edited subject losing its identity or the preserved subject shifting in layout or pose; we measure identity drift with the same DINOv2 feature distance restricted to the edit mask rather than its complement, and note that a full accounting of layout and pose displacement is left to future work.

D.6 GENERATION DIAGNOSTICS

This section provides the generation-side analogue of the editing diagnostics. The evidence is organised into qualitative comparisons, trajectory views, and aggregate alignment statistics so that each figure group has nearby context.

Qualitative comparisons. Figs. 6 and 7 show where VAGS-Gen helps with counts, fine-grained attributes, and composition while preserving sample diversity.

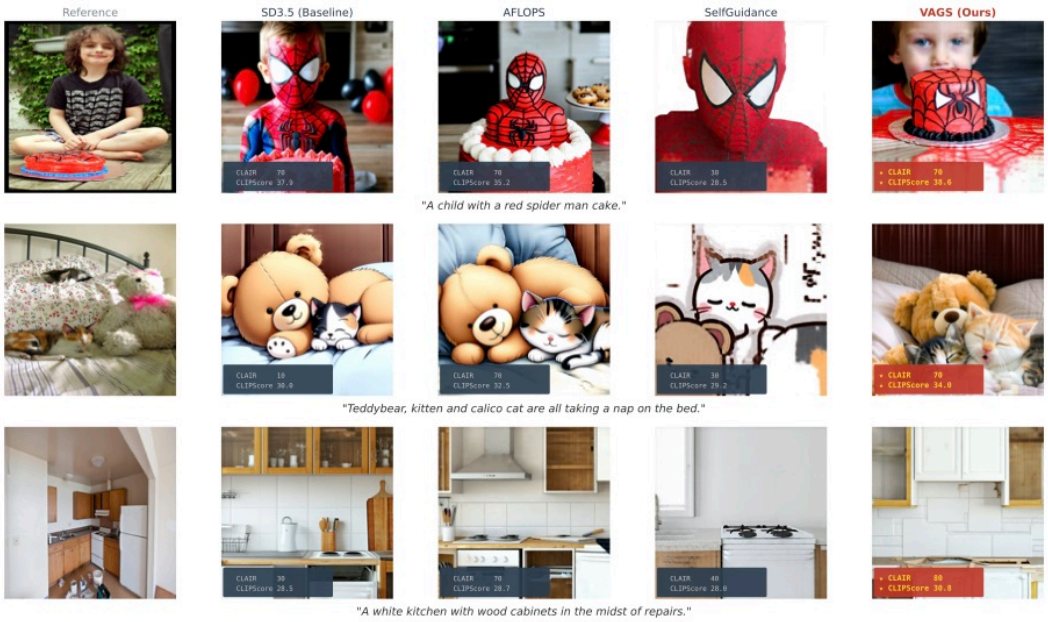


Figure 6: Qualitative generation comparison on COCO17 and CUB-200. VAGS-Gen improves prompt alignment and visual fidelity over fixed-CFG generation while preserving sample diversity.

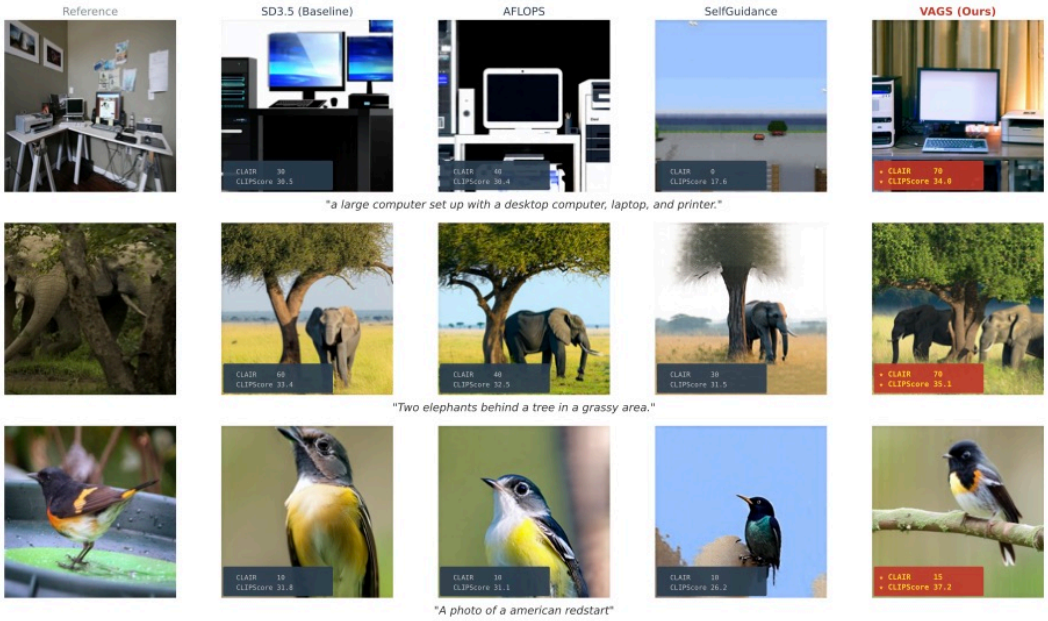


Figure 7: Generation qualitative comparison: additional examples on COCO17 and CUB-200.

Trajectory diagnostics. The generation panel of Fig. 4, together with Figs. 8 and 9, verifies that the qualitative gains correspond to a reshaped CFG schedule driven by unconditional/conditional velocity alignment.

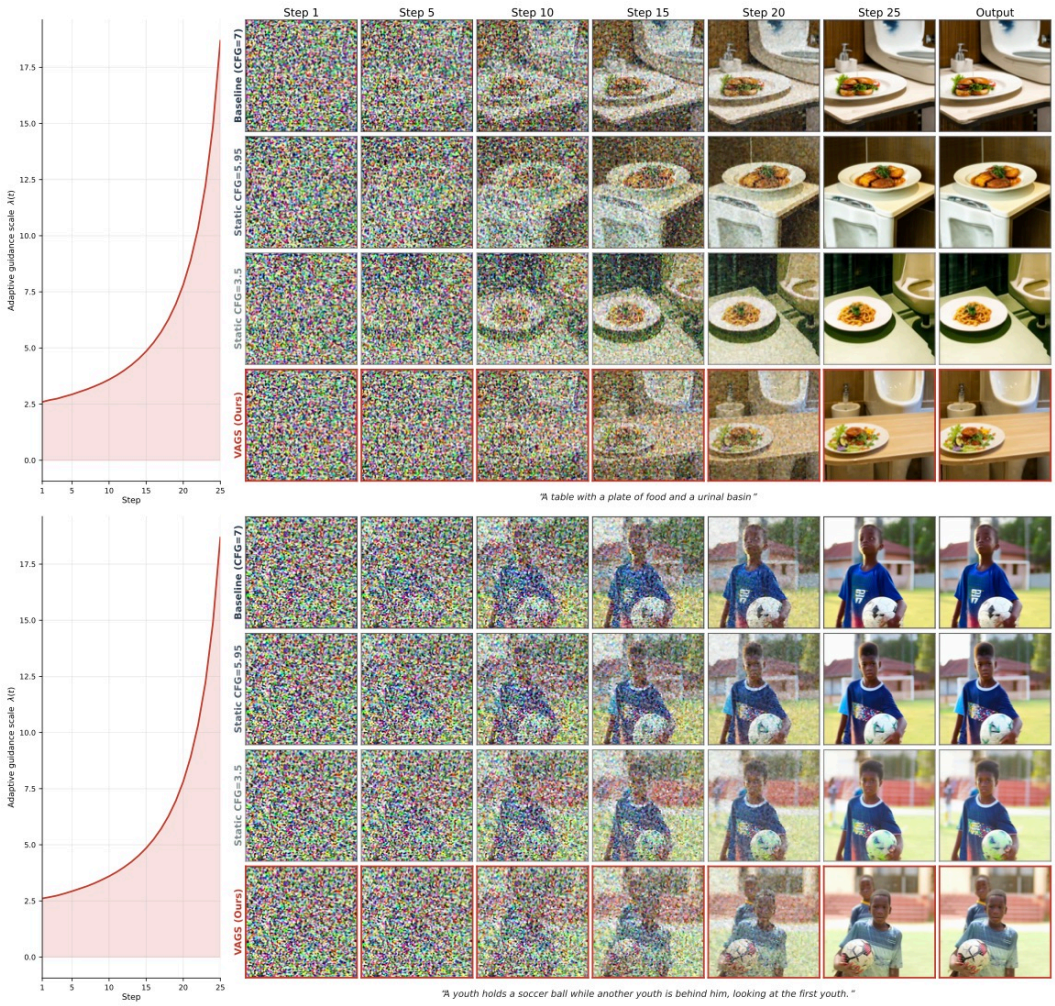


Figure 8: Progressive denoising on COCO17: per-step intermediate images for VAGS (κ=1) versus three static-CFG baselines (CFG=7 default, CFG=3.5 low, CFG=5.95 matched-mean of VAGS adaptive λ). Leftmost column shows the VAGS adaptive $\lambda(t)$ curve. Even when matched on mean guidance, static CFG denoises along a distinguishable trajectory from VAGS.

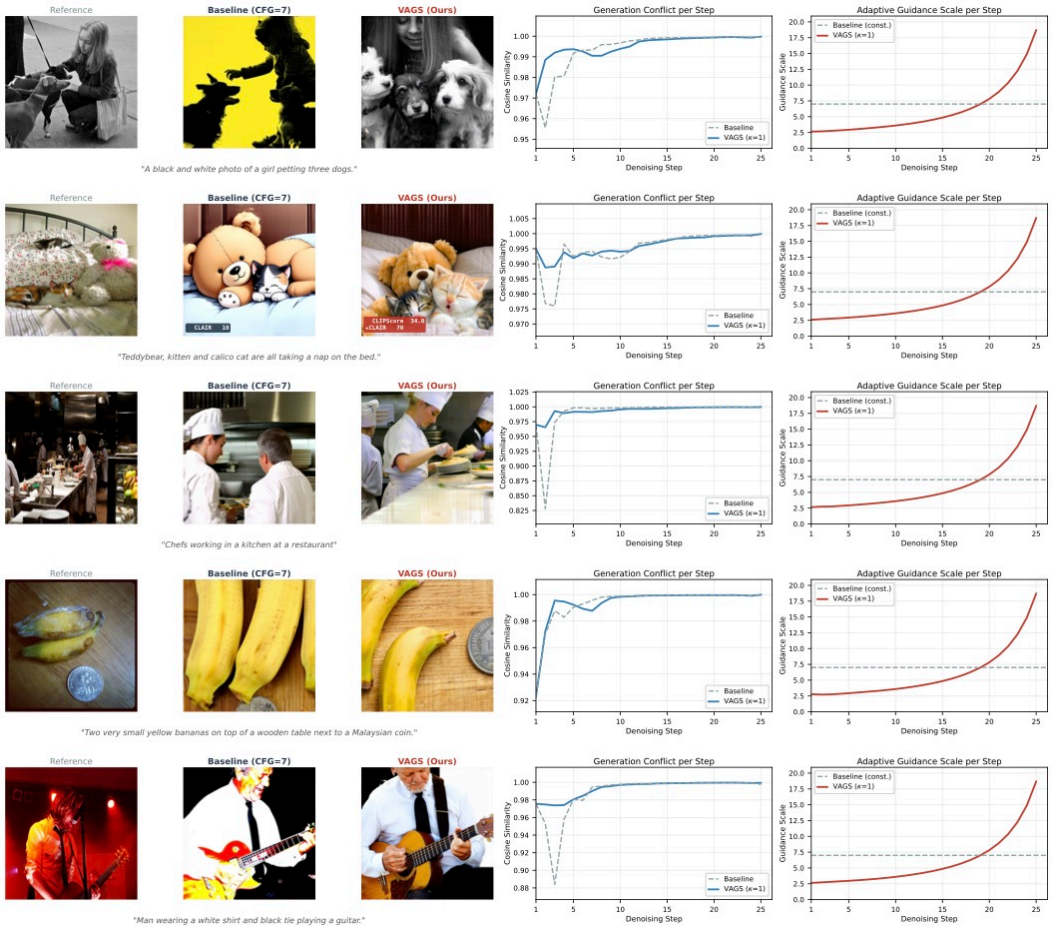
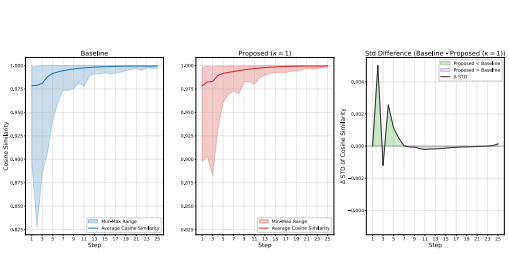


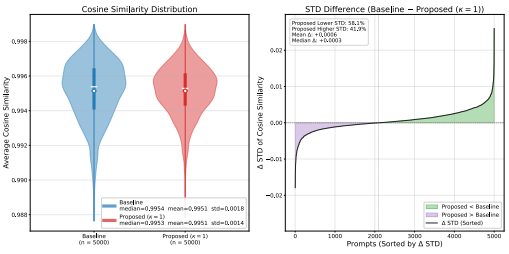
Figure 9: Generation trajectory analysis: additional per-prompt traces.

Aggregate alignment statistics. Fig. 10 summarises the alignment and adaptive-scale statistics across COCO17. Within this grouped diagnostic, Figs. 10a and 10b report step-wise and prompt-wise alignment, while Figs. 10c and 10d report the corresponding adaptive scales.

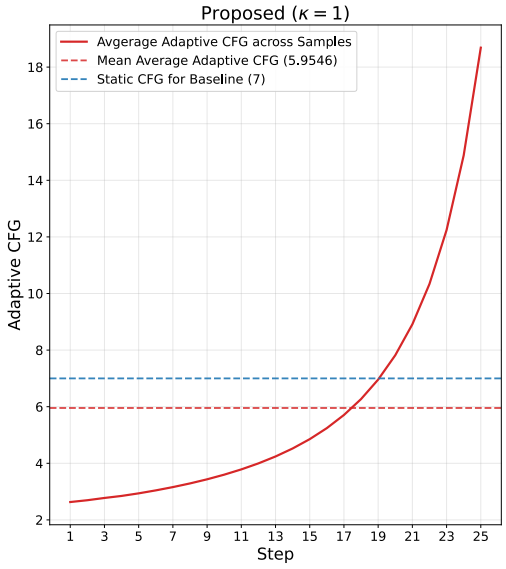
Matched-mean generation control. Tab. 6 compares VAGS-Gen with a constant-CFG baseline set to the mean adaptive value ($\bar{\lambda} = 5.955$). The matched-mean constant scale yields only a small FID improvement over fixed CFG, whereas VAGS-Gen obtains a much larger gain.



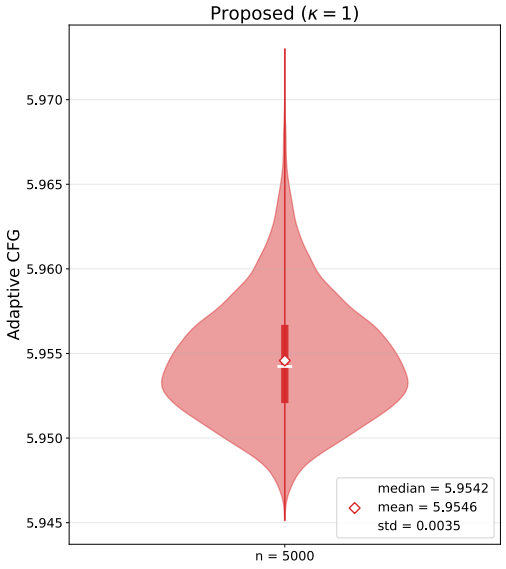
(a) Step-wise velocity alignment



(b) Prompt-wise alignment statistics



(c) Step-wise adaptive CFG



(d) Prompt-wise mean CFG

Figure 10: Generation diagnostics on COCO17. VAGS-Gen uses the unconditional/conditional velocity cosine to reshape the CFG trajectory while keeping the prompt-wise mean scale concentrated.

Dataset	Method	FID ↓	IS ↑	CLIPScore ↑	Caption-based Metrics			
					BLEU-4 ↑	METEOR ↑	ROUGE-L ↑	CLAIR ↑
COCO17 Lin et al. (2014)	SDv3.5 (CFG=7)	28.46	33.64	**32.84**	7.99	29.17	35.11	71.45
	SDv3.5 w/ VAGS (CFG=5.955, κ=0)	28.28	34.63	**32.84**	8.20	**29.69**	**35.37**	**72.27**
	SDv3.5 w/ VAGS (CFG=7, κ=1.0)	**26.07**	**35.15**	**32.84**	**8.28**	29.53	35.21	71.62

Table 6: Constant-CFG vs. VAGS-Gen on COCO17. A fixed scale at the mean adaptive value ($\bar{\lambda} = 5.955$) barely outperforms the baseline, confirming that the gains stem from dynamic modulation rather than a shifted average guidance level.

D.7 KAPPA SENSITIVITY

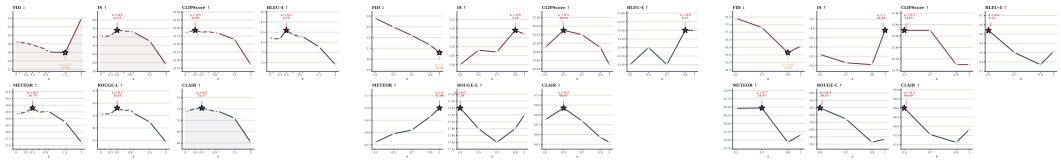
The modulation strength κ controls the range of adaptive scales. Tabs. 7 and 8 report the raw metrics, while Figs. 11, 11a, 11b, 11c, and 12 visualise the trends. Moderate modulation improves preservation or FID; overly large values begin to suppress or amplify guidance too aggressively.

Generation trends. On COCO17, FID decreases monotonically from 28.58 at κ=0 (constant CFG) to 26.07 at κ=1.0, reaching its absolute minimum of 25.93 at κ=1.5. IS, BLEU-4, METEOR, ROUGE-L, and CLAIR all peak earlier at κ=0.5 (IS 35.55, BLEU-4 8.59, CLAIR 72.17) and degrade past κ=1.0; κ=2.0 is a clear collapse on every measure (FID 33.99, IS 27.54, CLAIR 66.13). We adopt κ=1.0 rather than the slightly lower κ=1.5 (FID 25.93) because moving from κ=1.0 to κ=1.5 saves only 0.14 FID while losing 2.12 on IS, 0.50 on BLEU-4, 0.78 on METEOR, 0.74 on ROUGE-L, and 1.24 on CLAIR. On CUB-200, FID decreases monotonically from 23.78 at κ=0.3 to 20.60 at κ=1.0 over the tested range, with no observed collapse, suggesting that fine-grained generation benefits from the full bound allowed by κ=1.0. On Flickr30K the optimum sits slightly lower at κ=0.9 (FID 75.65 vs. 76.08 at κ=1.0); the more diverse Flickr30K scenes appear to favour a marginally milder correction, hence the per-dataset choice.

Editing trends. On PIE-Bench, structure-oriented metrics reach their optimum at κ_{tar}=0.9 (Dist 13.84, PSNR 26.38, MSE 34.86), whereas perceptual metrics peak earlier at κ_{tar}=0.5 (LPIPS 67.74, SSIM 87.86). We adopt κ_{tar}=0.9 since the preservation/editability trade-off central to inversion-free editing is governed primarily by structural fidelity, and the perceptual loss between κ_{tar}=0.5 and κ_{tar}=0.9 is small: LPIPS rises by 2.64 and SSIM drops by 0.18 across that range. Beyond κ_{tar}=0.9 every metric degrades: at κ_{tar}=2.0, Dist rises from 13.84 to 20.30, LPIPS rises to 109.25, and SSIM drops to 84.07, the same clean-regime over-amplification observed on the generation side.

Table 7: Ablation of κ on COCO17, CUB-200, and Flickr30K. Top results are in bold and second-best results are underlined. We adopt κ=1.0 as the default.

Dataset	Method	FID ↓	IS ↑	CLIPScore ↑	Caption-based Metrics			
					BLEU-4 ↑	METEOR ↑	ROUGE-L ↑	CLAIR ↑
COCO17 Lin et al. (2014)	SDv3.5	28.46	33.64	32.84	7.99	29.17	35.11	71.45
	VAGS, κ=0	28.58	34.35	32.85	8.24	29.35	35.15	71.63
	VAGS, κ=0.1	28.36	34.03	32.88	8.17	29.37	35.13	71.65
	VAGS, κ=0.3	28.17	34.34	**32.93**	8.18	29.49	35.14	72.09
	VAGS, κ=0.5	27.80	**35.55**	32.87	**8.59**	**29.79**	**35.61**	**72.17**
	VAGS, κ=0.7	27.23	35.43	32.89	8.38	29.47	35.40	71.87
	VAGS, κ=0.9	26.77	35.27	32.85	8.28	29.53	35.46	71.55
	VAGS, κ=1.0	26.07	35.15	32.84	8.28	29.53	35.21	71.62
	VAGS, κ=1.5	**25.93**	33.03	32.65	7.78	28.75	34.47	70.38
	VAGS, κ=2.0	33.99	27.54	31.86	6.90	27.23	32.84	66.13
CUB-200 Welinder et al. (2010)	SDv3.5	24.92	4.81	32.40	0.13	16.57	**17.64**	66.69
	VAGS, κ=0.3	23.78	5.00	32.49	0.13	16.72	17.54	66.75
	VAGS, κ=0.5	22.97	5.08	**32.53**	0.14	16.79	17.48	**66.87**
	VAGS, κ=0.7	22.19	5.07	32.52	0.13	16.82	17.44	66.74
	VAGS, κ=0.9	21.30	**5.19**	32.49	**0.15**	16.93	17.48	66.57
	VAGS, κ=1.0	**20.60**	5.17	32.45	**0.15**	**17.00**	17.52	66.52
Flickr30K Plummer et al. (2015)	SDv3.5	79.58	17.95	33.79	4.12	23.88	29.30	68.03
	VAGS, κ=0.5	77.92	18.32	**33.83**	**4.54**	24.46	**30.19**	**68.40**
	VAGS, κ=0.7	77.32	18.23	**33.83**	4.30	**24.47**	29.89	67.82
	VAGS, κ=0.9	**75.65**	18.21	33.77	4.17	23.43	29.25	67.65
	VAGS, κ=1.0	76.08	**18.58**	33.77	4.31	23.66	29.34	67.93



(a) COCO17 (b) CUB-200 (c) Flickr30K

Figure 11: Effect of modulation strength κ on generation metrics. Moderate modulation improves fidelity, while overly large values begin to suppress or amplify guidance too aggressively.

Table 8: Ablation of κ for image editing on PIE-Bench (700 pairs). Top results are in bold and second-best results are underlined. We adopt κ=0.9 as the default.

Dataset	Method	Dist ↓	PSNR ↑	LPIPS ↓	MSE ↓	SSIM ↑	CLIP Similarity		
							Whole ↑	Edited ↑	CLIP-I ↑
PIE-Bench	FlowEdit	30.31	21.87	117.06	92.70	82.54	**27.62**	**23.80**	83.69
	FlowEdit + Monotone (κ=0)	15.91	25.43	69.87	43.47	87.67	27.05	23.26	84.73
	VAGS, κ=0.5	14.32	26.07	**67.74**	37.70	**87.86**	27.00	23.25	**87.55**
	VAGS, κ=0.7	13.97	26.25	68.58	36.16	87.82	26.99	23.14	87.52
	VAGS, κ=0.9	**13.84**	**26.38**	70.38	**34.86**	87.68	26.92	23.08	87.44
	VAGS, κ=1.5	15.62	26.15	85.04	35.90	86.32	27.01	23.00	86.41
	VAGS, κ=2.0	20.30	25.30	109.25	42.06	84.07	26.97	23.03	84.81

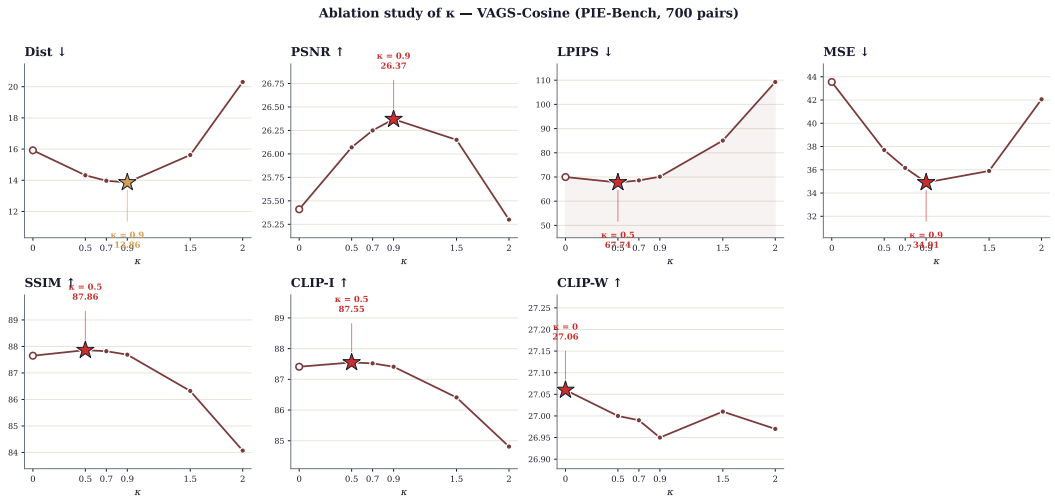


Figure 12: Effect of κ on PIE-Bench editing metrics.

Dataset	Method	Model	Structure	Background Preservation					CLIP Similarity	
			Dist. $\times 10^3 \downarrow$	PSNR ↑	LPIPS $\times 10^3 \downarrow$	MSE $\times 10^4 \downarrow$	SSIM $\times 10^2 \uparrow$	CLIP-I ↑	Whole ↑	Edited ↑
PIE-Bench	*Diffusion-based*									
	DDIM Song et al. (2021a) + P2P Hertz et al. (2023)	SD 1.4	71.12 (1059.39)	17.84 (1.20)	207.62 (3092.68)	220.79 (3288.85)	71.62 (4.81)	82.44 (5.53)	25.91	22.75
	DDIM Song et al. (2021a) + PnP Tumanyan et al. (2023)	SD 1.4	27.27 (399.40)	22.36 (1.53)	110.22 (1614.29)	81.94 (1200.10)	79.65 (5.44)	86.93 (5.94)	26.13	22.79
	Null-text Mokady et al. (2023) + P2P Hertz et al. (2023)	SD 1.4	15.69 (248.85)	**26.92 (1.70)**	**66.42 (1053.43)**	35.09 (556.53)	82.67 (5.21)	85.42 (5.39)	25.11	22.01
	Flow-based / Rectified Flow									
	RF-Inversion Rout et al. (2024)	FLUX.1	47.19 (689.56)	20.64 (1.41)	193.68 (2830.15)	123.33 (1802.16)	76.34 (5.22)	80.23 (5.49)	26.16	22.38
	RF-Solver Wang et al. (2024a)	FLUX.1	48.04 (677.42)	20.52 (1.46)	181.15 (2554.44)	128.54 (1812.57)	77.81 (5.52)	82.20 (5.83)	26.63	23.46
	FireFlow Deng et al. (2024) + RF-Solver Wang et al. (2024a)	FLUX.1	25.88 (384.02)	22.90 (1.54)	129.27 (1918.18)	74.73 (1108.88)	82.03 (5.53)	84.76 (5.71)	25.96	22.02
	FlowEdit Kulikov et al. (2024)	FLUX.1	27.93 (407.19)	21.74 (1.49)	113.83 (1659.53)	98.75 (1439.68)	83.46 (5.72)	86.54 (5.94)	26.19	22.19
	FTEdit Xu et al. (2025) + AdaLN	SD 3.5	**13.81 (220.78)**	23.69 (1.48)	77.69 (1242.05)	36.10 (577.14)	**92.62 (5.79)**	**92.40 (5.78)**	25.01	21.81
	iRFDS Yang et al. (2025)	SD 3.5	67.64 (1044.31)	19.42 (1.26)	150.68 (2326.38)	147.86 (2282.84)	79.47 (5.15)	77.69 (5.03)	25.45	21.73
	FlowEdit Kulikov et al. (2024)	SD 3.5	30.31 (397.32)	21.87 (1.67)	117.06 (1534.48)	92.70 (1215.16)	82.54 (6.30)	83.69 (6.38)	**27.62**	**23.80**
	SplitFlow Yoon et al. (2025)	SD 3.5	28.03 (370.11)	22.32 (1.69)	111.49 (1472.11)	85.31 (1126.43)	83.17 (6.30)	84.10 (6.37)	27.52	23.77
	Ours									
	FlowEdit + VAGS	SD 3.5	13.84 (190.98)	26.38 (1.91)	70.38 (971.18)	**34.86 (481.04)**	87.68 (6.35)	87.44 (6.34)	26.92	23.08
DIV2K	*Diffusion-based*									
	DDIM Song et al. (2021a) + P2P Hertz et al. (2023)	SD 1.4	64.46 (921.38)	15.00 (1.05)	323.57 (4625.05)	359.10 (5132.91)	60.93 (4.26)	81.89 (5.73)	26.45	-
	DDIM Song et al. (2021a) + PnP Tumanyan et al. (2023)	SD 1.4	30.66 (380.94)	18.75 (1.51)	205.72 (2555.98)	154.73 (1922.45)	68.61 (5.52)	83.83 (6.75)	28.37	-
	Null-text Mokady et al. (2023) + P2P Hertz et al. (2023)	SD 1.4	15.82 (207.53)	21.35 (1.63)	137.28 (1800.84)	93.17 (1222.20)	74.37 (5.67)	88.32 (6.73)	27.61	-
	Flow-based / Rectified Flow									
	RF-Inversion Rout et al. (2024)	FLUX.1	31.87 (363.75)	20.05 (1.76)	252.75 (2884.75)	115.45 (1317.68)	73.37 (6.43)	79.57 (6.97)	29.60	-
	RF-Solver Wang et al. (2024a)	FLUX.1	34.63 (387.36)	18.56 (1.66)	265.56 (2970.44)	161.60 (1807.59)	72.98 (6.52)	79.54 (7.11)	29.90	-
	FireFlow Deng et al. (2024) + RF-Solver Wang et al. (2024a)	FLUX.1	20.52 (247.91)	21.02 (1.74)	197.26 (2383.19)	92.75 (1120.56)	77.88 (6.45)	82.86 (6.86)	28.77	-
	FlowEdit Kulikov et al. (2024)	FLUX.1	24.84 (293.94)	18.73 (1.58)	173.21 (2049.67)	152.33 (1802.58)	78.81 (6.66)	83.93 (7.09)	29.07	-
	FTEdit Xu et al. (2025) + AdaLN	SD 3.5	14.43 (197.50)	23.02 (1.68)	125.97 (1724.15)	56.65 (775.37)	83.73 (6.12)	**88.97 (6.50)**	27.03	-
	iRFDS Yang et al. (2025)	SD 3.5	61.14 (810.23)	15.86 (1.20)	317.41 (4206.33)	320.07 (4241.58)	63.02 (4.76)	76.79 (5.79)	27.47	-
	FlowEdit Kulikov et al. (2024)	SD 3.5	18.27 (202.19)	20.27 (1.83)	145.55 (1610.77)	103.37 (1143.98)	82.15 (7.42)	82.31 (7.44)	**30.06**	-
	SplitFlow Yoon et al. (2025)	SD 3.5	17.61 (196.32)	20.53 (1.84)	141.99 (1582.94)	97.19 (1083.50)	82.63 (7.41)	82.44 (7.39)	29.95	-
	Ours									
	FlowEdit + VAGS	SD 3.5	**8.62 (98.32)**	**23.61 (2.07)**	**98.43 (1122.67)**	**50.22 (572.80)**	**84.41 (7.40)**	84.07 (7.37)	29.61	-

Table 9: [CLIP-scaled] Quantitative comparison on PIE-Bench and DIV2K. FlowEdit + VAGS ranks first on all background preservation metrics while maintaining competitive edit strength.

D.8 CLIP-SCALED METRIC TABLES

Tables 9, 10, 11, 12, and 13 report the same comparisons as the main paper and sensitivity studies after normalising non-CLIP metrics by whole-image CLIP similarity. The purpose is to penalise methods whose low distortion or high fidelity scores are partly explained by weak text-image alignment. For every cell except the CLIP scores themselves, the CLIP-scaled value in parentheses is computed as $\text{scaled} = \text{raw}/(\text{CLIP}_\text{W}/100)^2$ for lower-is-better metrics and $\text{scaled} = \text{raw} \cdot (\text{CLIP}_\text{W}/100)^2$ for higher-is-better metrics.

Editing SOTA. On PIE-Bench, FTEdit's raw Structure Distance 13.81 sits within 0.03 of FlowEdit+VAGS at 13.84, but its lower CLIP-W (25.01 vs. 26.92) yields a substantially larger scaled value (220.78 vs. 190.98); VAGS therefore leads by roughly 13% on scaled Dist where the raw gap is essentially zero. FTEdit's bold raw SSIM (92.62) and CLIP-I (92.40) likewise fall below the VAGS row when scaled (5.79 and 5.78 vs. 6.35 and 6.34). On DIV2K, VAGS is bold on every non-CLIP metric in raw form and retains the lead under scaling.

Editing scheduler ablation. The scaled column reorders the failure modes more sharply than the raw column. Zero-Init's PIE-Bench Structure Distance of 194.67 becomes 2406.80 under scaling on FlowEdit, roughly a 12× penalty over VAGS' 190.98. The penalty comes overwhelmingly from the catastrophic raw Distance (194.67 vs. 13.84), since Zero-Init's CLIP-W is comparable to the other rows. The bold ranking on Dist, PSNR, MSE, SSIM, and CLIP-I is otherwise unchanged: VAGS leads on both editors and on both datasets in raw and scaled form.

Generation SOTA. Scaling preserves the raw conclusion that VAGS-Gen leads on COCO17, CUB-200, and Flickr30K. CLIPScore is the divisor and is therefore reported unchanged; the scaled FID, IS, BLEU-4, METEOR, ROUGE-L, and CLAIR rows keep VAGS-Gen at the top or tied position on all three datasets. Self-Guidance's degradation, already visible in the raw rows (FID 40.98 on COCO17), is amplified under scaling (420.71) since its CLIPScore is the lowest in its block (31.21 vs. 32.84 for the other COCO17 rows).

Dataset	Editor	Scheduler	Structure	Background Preservation					CLIP Similarity	
			Dist. $\times 10^3 \downarrow$	PSNR $\uparrow$	LPIPS $\times 10^3 \downarrow$	MSE $\times 10^4 \downarrow$	SSIM $\times 10^2 \uparrow$	CLIP-I $\uparrow$	Whole $\uparrow$	Edited $\uparrow$
PIE-Bench	FlowEdit	No Scheduler	30.31 (397.32)	21.87 (1.67)	117.06 (1534.48)	92.70 (1215.16)	82.54 (6.30)	83.69 (6.38)	27.62	23.80
		Interval Kynkäänniemi et al. (2024b)	29.82 (392.03)	21.93 (1.67)	114.14 (1500.55)	91.72 (1205.80)	82.88 (6.30)	84.03 (6.39)	27.58	23.75
		Monotone Wang et al. (2024b)	15.91 (217.44)	25.43 (1.86)	**69.87 (954.90)**	43.47 (594.09)	87.67 (6.41)	84.73 (6.20)	27.05	23.26
		Zero-Init Fan et al. (2025)	194.67 (2406.80)	11.69 (0.946)	354.57 (4383.72)	867.59 (10726.43)	57.64 (4.66)	75.28 (6.09)	**28.44**	**24.24**
		VAGS (Ours)	**13.84 (190.98)**	**26.38 (1.91)**	70.38 (971.18)	**34.86 (481.04)**	**87.68 (6.35)**	**87.44 (6.34)**	26.92	23.08
	SplitFlow Yoon et al. (2025)	No Scheduler	28.03 (370.11)	22.32 (1.69)	111.49 (1472.11)	85.31 (1126.43)	83.17 (6.30)	84.10 (6.37)	**27.52**	**23.77**
		Interval Kynkäänniemi et al. (2024b)	27.28 (367.38)	23.05 (1.71)	106.26 (1430.99)	77.64 (1045.57)	84.17 (6.25)	83.25 (6.18)	27.25	23.22
		Monotone Wang et al. (2024b)	16.06 (223.10)	**26.25 (1.89)**	**66.73 (927.00)**	38.77 (538.59)	**88.12 (6.34)**	85.94 (6.19)	26.83	22.76
		Zero-Init Fan et al. (2025)	21.65 (290.49)	23.57 (1.76)	94.01 (1261.39)	65.07 (873.08)	84.90 (6.33)	85.36 (6.36)	27.30	23.57
		VAGS (Ours)	**14.67 (200.34)**	25.99 (1.90)	73.43 (1002.81)	**38.41 (524.55)**	87.31 (6.39)	**87.22 (6.39)**	27.06	23.28
DIV2K	FlowEdit	No Scheduler	18.27 (202.19)	20.27 (1.83)	145.55 (1610.77)	103.37 (1143.98)	82.15 (7.42)	82.31 (7.44)	**30.06**	-
		Interval Kynkäänniemi et al. (2024b)	18.11 (200.82)	19.94 (1.80)	144.88 (1606.56)	111.55 (1236.97)	79.51 (7.17)	82.32 (7.42)	30.03	-
		Monotone Wang et al. (2024b)	10.89 (124.29)	22.47 (1.97)	102.91 (1174.56)	63.91 (729.43)	83.64 (7.33)	**84.10 (7.37)**	29.60	-
		Zero-Init Fan et al. (2025)	194.67 (2406.80)	11.69 (0.946)	354.57 (4383.72)	867.59 (10726.43)	57.64 (4.66)	75.28 (6.09)	28.44	-
		VAGS (Ours)	**8.62 (98.32)**	**23.61 (2.07)**	**98.43 (1122.67)**	**50.22 (572.80)**	**84.41 (7.40)**	84.07 (7.37)	29.61	-
	SplitFlow Yoon et al. (2025)	No Scheduler	17.61 (196.32)	20.53 (1.84)	141.99 (1582.94)	97.19 (1083.50)	82.63 (7.41)	82.44 (7.39)	29.95	-
		Interval Kynkäänniemi et al. (2024b)	16.01 (178.96)	20.53 (1.84)	136.04 (1520.67)	97.62 (1091.20)	80.55 (7.21)	82.65 (7.39)	29.91	-
		Monotone Wang et al. (2024b)	9.63 (110.28)	23.01 (2.01)	**97.16 (1112.69)**	56.81 (650.59)	**84.23 (7.35)**	**84.43 (7.37)**	29.55	-
		Zero-Init Fan et al. (2025)	46.72 (468.47)	14.83 (1.48)	378.94 (3799.67)	345.27 (3462.06)	48.95 (4.88)	75.93 (7.57)	**31.58**	-
		VAGS (Ours)	**9.32 (106.30)**	**23.28 (2.04)**	102.58 (1170.00)	**53.57 (611.01)**	84.01 (7.37)	84.08 (7.37)	29.61	-

Table 10: [CLIP-scaled] Ablation of guidance scheduling strategies on FlowEdit and SplitFlow. VAGS gives the strongest overall preservation and edit-quality trade-off under the CLIP-scaled normalization.

Dataset	Method	FID $\downarrow$	IS $\uparrow$	CLIPScore $\uparrow$	Caption-based Metrics			
					BLEU-4 $\uparrow$	METEOR $\uparrow$	ROUGE-L $\uparrow$	CLAIR $\uparrow$
COCO17 Lin et al. (2014)	SDv3.5 (CFG=7)	28.46 (263.89)	33.64 (3.63)	**32.84**	7.99 (0.862)	29.17 (3.15)	35.11 (3.79)	71.45 (7.71)
	SDv3.5 w/ A-Euler Jin et al. (2025)	27.64 (256.29)	34.63 (3.73)	**32.84**	7.96 (0.858)	29.09 (3.14)	35.17 (3.79)	71.23 (7.68)
	SDv3.5 w/ Self-Guidance Li et al. (2025b)	40.98 (420.71)	28.54 (2.78)	31.21	6.52 (0.635)	26.18 (2.55)	31.59 (3.08)	62.26 (6.06)
	SDv3.5 w/ VAGS (CFG=7, κ=1.0)	**26.07 (241.73)**	**35.15 (3.79)**	**32.84**	**8.28 (0.893)**	**29.53 (3.18)**	**35.21 (3.80)**	**71.62 (7.72)**
CUB-200 Welinder et al. (2010)	SDv3.5 (CFG=7)	24.92 (237.39)	4.81 (0.505)	32.40	0.13 (0.014)	16.57 (1.74)	17.64 (1.85)	**66.69 (7.00)**
	SDv3.5 w/ A-Euler Jin et al. (2025)	22.98 (217.56)	5.05 (0.533)	**32.50**	0.12 (0.013)	16.39 (1.73)	17.51 (1.85)	66.22 (6.99)
	SDv3.5 w/ Self-Guidance Li et al. (2025b)	61.16 (626.28)	4.98 (0.486)	31.25	**0.18 (0.018)**	**17.85 (1.74)**	**17.78 (1.74)**	62.28 (6.08)
	SDv3.5 w/ VAGS (CFG=7, κ=1.0)	**20.60 (195.63)**	**5.17 (0.544)**	32.45	0.15 (0.016)	17.00 (1.79)	17.52 (1.84)	66.52 (7.00)
Flickr30K Plummer et al. (2015)	SDv3.5 (CFG=7)	79.58 (696.99)	17.95 (2.05)	**33.79**	4.12 (0.470)	**23.88 (2.73)**	**29.30 (3.35)**	**68.03 (7.77)**
	SDv3.5 w/ A-Euler Jin et al. (2025)	78.57 (691.42)	17.83 (2.03)	33.71	3.91 (0.444)	22.99 (2.61)	29.00 (3.30)	67.56 (7.68)
	SDv3.5 w/ Self-Guidance Li et al. (2025b)	93.07 (932.04)	14.44 (1.44)	31.60	3.35 (0.335)	20.80 (2.08)	26.27 (2.62)	57.35 (5.73)
	SDv3.5 w/ VAGS (CFG=7, κ=0.9)	**75.65 (663.36)**	**18.21 (2.08)**	33.77	**4.17 (0.476)**	23.43 (2.67)	29.25 (3.34)	67.65 (7.71)

Table 11: [CLIP-scaled] Text-to-image generation on COCO17, CUB-200, and Flickr30K. BLIP-2 Li et al. (2023) is used to generate captions from the generated images, enabling evaluation with the caption-based metrics. VAGS-Gen yields consistent FID and IS gains, with the largest improvement on fine-grained generation.

Generation κ ablation. The optimal κ on each dataset is unchanged under scaling (κ=1.0 on COCO17 and CUB-200, κ=0.9 on Flickr30K), and the same collapse at κ=2.0 on COCO17 is visible in the scaled column (FID 334.86 vs. VAGS at κ=1.0 at 241.73). The chosen modulation strength therefore does not depend on whether non-CLIP metrics are reported raw or scaled by alignment.

Matched-mean control. Setting CFG to the per-trajectory mean adaptive value ($\bar{\lambda}$=5.955, κ=0) improves COCO17 scaled FID marginally over fixed CFG (262.22 vs. 263.89), while VAGS-Gen at κ=1.0 reaches 241.73, a further 7.8% reduction. Scaling agrees with the raw conclusion: the gain comes from step-level modulation, not from a shifted average scale.

Table 12: [CLIP-scaled] Ablation of κ on COCO17, CUB-200, and Flickr30K. Top results are in bold and second-best results are underlined. We adopt κ=1.0 as the default.

Dataset	Method	FID ↓	IS ↑	CLIPScore ↑	Caption-based Metrics			
					BLEU-4 ↑	METEOR ↑	ROUGE-L ↑	CLAIR ↑
COCO17 Lin et al. (2014)	SDv3.5	28.46 (263.89)	33.64 (3.63)	32.84	7.99 (0.862)	29.17 (3.15)	35.11 (3.79)	71.45 (7.71)
	VAGS, κ=0	28.58 (264.84)	34.35 (3.71)	32.85	8.24 (0.889)	29.35 (3.17)	35.15 (3.79)	71.63 (7.73)
	VAGS, κ=0.1	28.36 (262.33)	34.03 (3.68)	32.88	8.17 (0.883)	29.37 (3.18)	35.13 (3.80)	71.65 (7.75)
	VAGS, κ=0.3	28.17 (259.78)	34.34 (3.72)	**32.93**	8.18 (0.887)	29.49 (3.20)	35.14 (3.81)	72.09 (7.82)
	VAGS, κ=0.5	27.80 (257.30)	**35.55 (3.84)**	32.87	**8.59 (0.928)**	**29.79 (3.22)**	**35.61 (3.85)**	**72.17 (7.80)**
	VAGS, κ=0.7	27.23 (251.72)	35.43 (3.83)	32.89	8.38 (0.907)	29.47 (3.19)	35.40 (3.83)	71.87 (7.77)
	VAGS, κ=0.9	26.77 (248.07)	35.27 (3.81)	32.85	8.28 (0.894)	29.53 (3.19)	35.46 (3.83)	71.55 (7.72)
	VAGS, κ=1.0	26.07 (241.73)	35.15 (3.79)	32.84	8.28 (0.893)	29.53 (3.18)	35.21 (3.80)	71.62 (7.72)
	VAGS, κ=1.5	**25.93 (243.24)**	33.03 (3.52)	32.65	7.78 (0.829)	28.75 (3.06)	34.47 (3.67)	70.38 (7.50)
	VAGS, κ=2.0	33.99 (334.86)	27.54 (2.80)	31.86	6.90 (0.700)	27.23 (2.76)	32.84 (3.33)	66.13 (6.71)
CUB-200 Welinder et al. (2010)	SDv3.5	24.92 (237.39)	4.81 (0.505)	32.40	0.13 (0.014)	16.57 (1.74)	**17.64 (1.85)**	66.69 (7.00)
	VAGS, κ=0.3	23.78 (225.27)	5.00 (0.528)	32.49	0.13 (0.014)	16.72 (1.76)	17.54 (1.85)	66.75 (7.05)
	VAGS, κ=0.5	22.97 (217.07)	5.08 (0.538)	**32.53**	0.14 (0.015)	16.79 (1.78)	17.48 (1.85)	**66.87 (7.08)**
	VAGS, κ=0.7	22.19 (209.82)	5.07 (0.536)	32.52	0.13 (0.014)	16.82 (1.78)	17.44 (1.84)	66.74 (7.06)
	VAGS, κ=0.9	21.30 (201.78)	**5.19 (0.548)**	32.49	**0.15 (0.016)**	16.93 (1.79)	17.48 (1.85)	66.57 (7.03)
	VAGS, κ=1.0	**20.60 (195.63)**	5.17 (0.544)	32.45	**0.15 (0.016)**	**17.00 (1.79)**	17.52 (1.84)	66.52 (7.00)
Flickr30K Plummer et al. (2015)	SDv3.5	79.58 (696.99)	17.95 (2.05)	33.79	4.12 (0.470)	23.88 (2.73)	29.30 (3.35)	68.03 (7.77)
	VAGS, κ=0.5	77.92 (680.84)	18.32 (2.10)	**33.83**	**4.54 (0.520)**	24.46 (2.80)	**30.19 (3.46)**	**68.40 (7.83)**
	VAGS, κ=0.7	77.32 (675.60)	18.23 (2.09)	**33.83**	4.30 (0.492)	**24.47 (2.80)**	29.89 (3.42)	67.82 (7.76)
	VAGS, κ=0.9	**75.65 (663.36)**	18.21 (2.08)	33.77	4.17 (0.476)	23.43 (2.67)	29.25 (3.34)	67.65 (7.71)
	VAGS, κ=1.0	76.08 (667.13)	**18.58 (2.12)**	33.77	4.31 (0.492)	23.66 (2.70)	29.34 (3.35)	67.93 (7.75)

Dataset	Method	FID ↓	IS ↑	CLIPScore ↑	Caption-based Metrics			
					BLEU-4 ↑	METEOR ↑	ROUGE-L ↑	CLAIR ↑
COCO17 Lin et al. (2014)	SDv3.5 (CFG=7)	28.46 (263.89)	33.64 (3.63)	**32.84**	7.99 (0.862)	29.17 (3.15)	35.11 (3.79)	71.45 (7.71)
	SDv3.5 w/ VAGS (CFG=5.955, κ=0)	28.28 (262.22)	34.63 (3.73)	**32.84**	8.20 (0.884)	**29.69 (3.20)**	**35.37 (3.81)**	**72.27 (7.79)**
	SDv3.5 w/ VAGS (CFG=7, κ=1.0)	**26.07 (241.73)**	**35.15 (3.79)**	**32.84**	**8.28 (0.893)**	29.53 (3.18)	35.21 (3.80)	71.62 (7.72)

Table 13: [CLIP-scaled] Constant-CFG vs. VAGS-Gen on COCO17. A fixed scale at the mean adaptive value ($\bar{\lambda} = 5.955$) barely outperforms the baseline, confirming that the gains stem from dynamic modulation rather than a shifted average guidance level.