Images Talk Back: Guiding MMDiTs by Suppressing Image-to-Text Feedback

Anonymous authors
Paper under double-blind review

Abstract

Multimodal diffusion transformers (MMDiTs) allow the evolving image to update internal text representations through joint attention. These representations then guide subsequent image updates, raising a question: does this feedback help correct deviations from the prompt, or preserve them? We find that image-to-text (I2T) feedback can preserve the objects and attributes already represented in the evolving image, including errors. Transplanting I2T messages can rescue failures or disrupt successes. At matched intervention strengths, suppressing the influence of I2T feedback improves generation, whereas opposite controls degrade it. These findings motivate Semantic Feedback Guidance (SFG), which suppresses the influence of I2T feedback by guiding away from an auxiliary branch that strengthens this feedback and weakens text-to-image conditioning. The in-place variant, iSFG, directly suppresses I2T feedback and strengthens text-to-image conditioning, requiring no extra model evaluations. Temporal interventions motivate running the SFG branch during the first 20% of image sampling steps, requiring about $1.2\times$ the base NFE without classifier-free guidance (CFG). Experiments across three image-generation backbones show gains in alignment and preference-model scores. On FLUX, adding SFG to CFG raises GenEval from 0.6661 to 0.7479; on SD3.5M, iSFG adds $+1.33$ HPSv2.1 over CFG at zero extra NFE. The control also transfers to text-to-video and image-to-video generation, improving VBench scores.

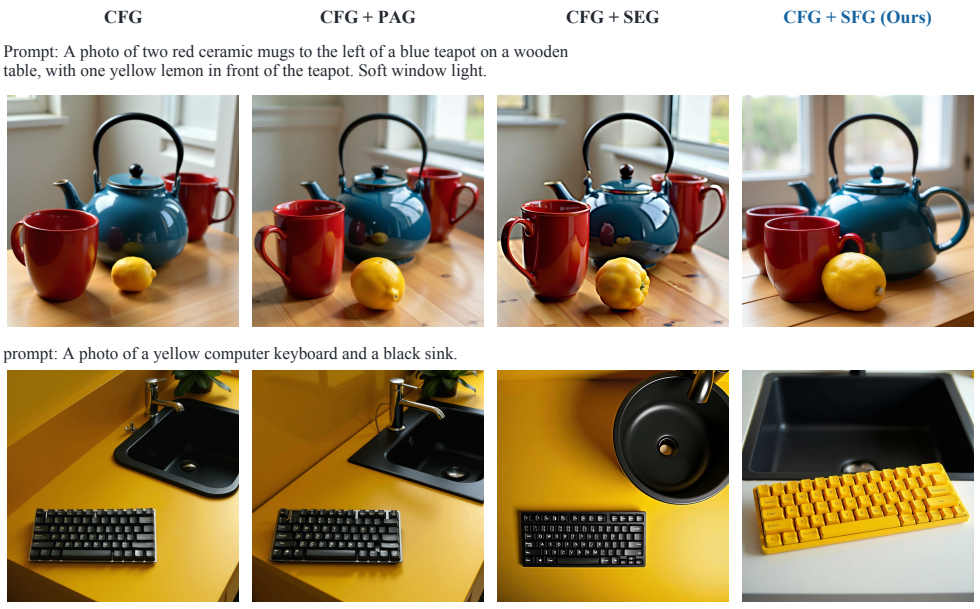


Figure 1: **Matched qualitative comparisons on FLUX.** When dealing with more complex prompts, our method aligns better with the objects and attributes specified in the complex prompts.

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

Text prompts specify what an image generator should produce, but in MMDiTs, the internal text stream also depends on what the model is drawing. Text queries read image keys and values before later blocks use the updated text stream to condition the image. This image-to-text-to-image loop underlies Stable Diffusion 3/3.5, FLUX, Qwen-Image, and HunyuanVideo (Esser et al., 2024; Black Forest Labs, 2024; Wu et al., 2025; Kong et al., 2024). When the image departs from the prompt, does this feedback correct the departure or preserve the existing image content? For the prompt "an apple and an orange," an image containing only an apple may write that incomplete scene back into the text stream. Figure 2 illustrates the two possible effects: recovering the orange or reinforcing its omission. The answer determines whether early errors are corrected or persist through sampling.

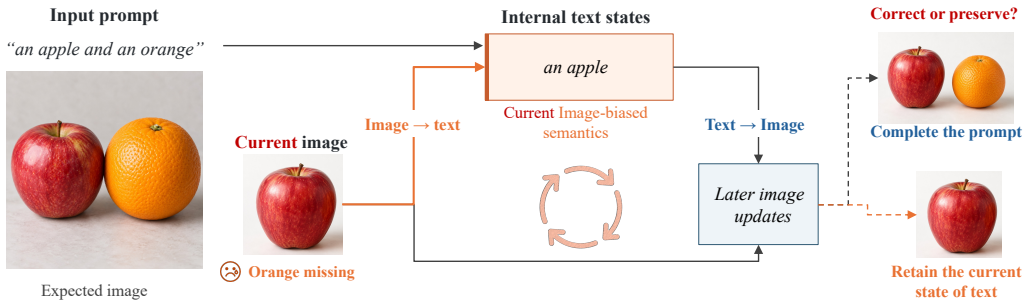


Figure 2: **Image-to-text feedback.** The current image writes into internal text states through $B_{i\to t}$, which later blocks read to update the image. If the image misses an object, does this feedback help recover it or preserve the incomplete scene? Text labels and AI-generated fruit images illustrate schematic states and possible outcomes; the input prompt stays fixed.

Existing analyses describe the image-to-text path as a register for image content and object identity, or as the weaker cross-modal direction (Toker et al., 2025; Li et al., 2026; Shin et al., 2025). They leave open how this content affects generation. We use *image content* to mean the objects and attributes represented in the evolving image. Such content can conflict with the prompt; we test the effect of preserving it by intervening on this path.

To address these questions, we intervene on the image-to-text path and measure its effect on generation. We find that this feedback can preserve image content that conflicts with the prompt, and identify suppression as a useful direction. Specifically, we design two experiments to examine feedback content and local response direction. Message transplantation shows that feedback can rescue failures or disrupt successes. At the same failed state, feedback suppression is more aligned with the response to successful-message transplantation than with a norm-matched random control. Our first ablation tests finite interventions and shows that suppression improves generation, whereas the opposite control degrades it.

We construct a *negative branch*, the reference prediction that guidance extrapolates away from, by increasing I2T feedback. Unlike SAG and PAG, which blur image regions or disrupt attention (Hong et al., 2023; Ahn et al., 2024), this branch preserves the attention computation and gives the current image more influence. Guiding away from it suppresses that influence. We call this control SFG-X. A complementary control, SFG-S, weakens text-to-image conditioning in the same negative branch, so guiding away strengthens conditioning. Together they form Semantic Feedback Guidance (SFG). The in-place variant, iSFG, applies feedback suppression and stronger conditioning directly in the conditional forward, with no extra model evaluations.

Early message transplants have the greatest effect on the final outcome, consistent with object layout being largely fixed early in sampling, motivating a SFG guidance window covering the first 20% of image sampling steps. This costs about $1.2\times$ the base NFE without CFG, compared with $2\times$ for an auxiliary branch at every step. SFG corrections remain nearly orthogonal to CFG through sampling, and the two methods combine effectively. Experiments on SD3M, SD3.5M, and FLUX show gains in alignment, preference scores, and distribution quality; transfer to HunyuanVideo improves text-to-video and image-to-video VBench scores over the base model and CFG.

Our contributions are:

- **How I2T feedback preserves image content.** Bidirectional transplantation tests the content carried by I2T feedback. Local response analysis and matched two-sided interventions identify suppression as a useful direction on failed trajectories.
- **A negative branch derived from feedback measurements.** SFG-X strengthens I2T feedback in its negative branch and guides against it. SFG combines this control with stronger T2I conditioning; temporal interventions motivate its guidance window.
- **Branch and in-place controls.** iSFG applies the controls directly with zero extra NFE. Experiments evaluate alignment and preference gains, combination with CFG, and transfer to both video tasks.

2 From mechanism analysis to signed guidance

2.1 Background: guidance and writable MMDiT states

CFG extrapolates between conditional and unconditional predictions; training-free self-guidance extrapolates away from a perturbed prediction of the same model. The formulas and perturbation choices are summarized in App. A. We choose the perturbation by measuring a native feedback path.

Writable MMDiT states. Generation integrates a learned velocity field backward from noise along $x_t = (1-t)x_0 + t\epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$ (Esser et al., 2024). The encoded prompt $c = E(p)$ stays fixed. Conventional cross-attention and deep fusion only let image queries read prompt-derived features (Liu et al., 2024; Tang et al., 2025). MMDiT instead maintains image (i) and text (t) residual streams, both querying concatenated image–text keys through a joint softmax:

$$O_i = A_{i\to i}V_i + \underbrace{A_{t\to i}V_t}_{B_{t\to i}}, \qquad O_t = A_{t\to t}V_t + \underbrace{A_{i\to t}V_i}_{B_{i\to t}}. \tag{1}$$

We call $B_{i\to t}$ the *I2T message* and $B_{t\to i}$ the *text-to-image (T2I) message*. The I2T message writes image values into the text stream; the T2I message conditions the image on text. Later blocks can read this updated text back into the image, closing a feedback loop. Each denoiser call resets text from $E(p)$, so the stream changes with depth and the current x_t without persisting across timesteps. We scale these messages after softmax, retaining joint normalization.

2.2 Analysis: feedback content, direction, and timing

We test what the I2T message carries, which intervention sign helps, and when it affects the final result. Figure 3 illustrates message transplantation and the local response directions. We then use those measurements to construct SFG and iSFG.

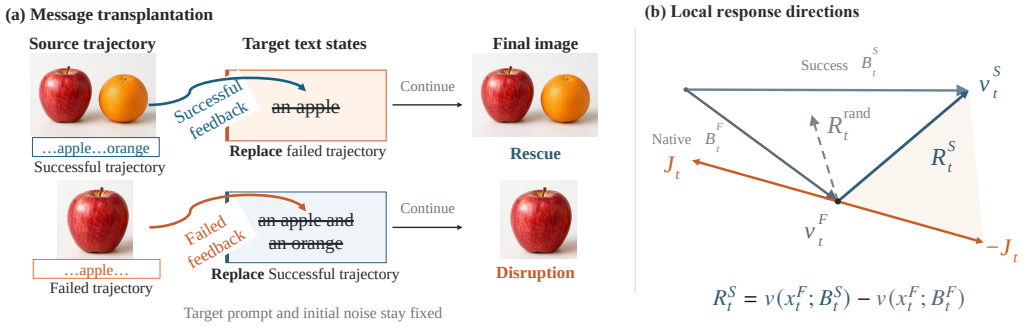


Figure 3: **Message transplantation and local response directions.** (a) Transplant donor image-to-text messages into recipient text states, keeping the recipient prompt and initial noise fixed. Success-to-failure and failure-to-success transplants test rescue and disruption. (b) J_t and $-J_t$ are opposite local responses; R_t^S and R_t^{rand} are velocity changes from successful-donor and random-message interventions. Text labels, AI-generated fruit images, and vector angles are schematic.

Paired trajectories. We use 400 same-prompt pairs, each containing one successful and one failed trajectory from unmodified SD3.5M or FLUX runs, differing only in generation seed. Labels use

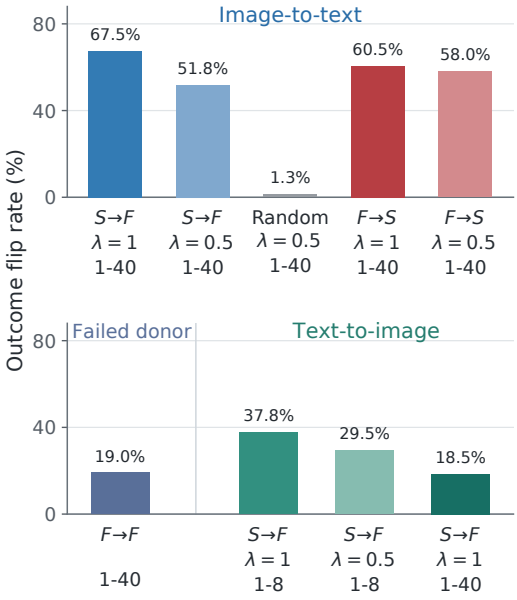


Figure 4: **Message transplantation and controls.** S/F denote successful/failed trajectories; arrows point from donor to recipient. Labels show donor mixing weight λ and transplant steps. Full results: Table 21.

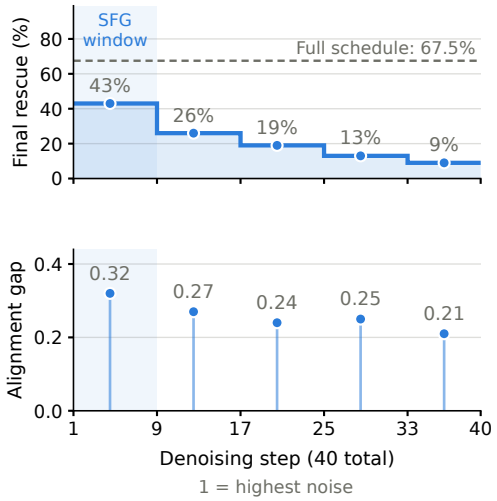


Figure 5: **Timing and alignment** on 400 pairs across SD3.5M and FLUX. Top: rescue by transplant window; steps 1–8 of 40 recover about two-thirds of the full-schedule effect. Bottom: cosine gap over random, averaged within pairs, then within models, with equal model weights. Shading marks steps 1–8.

GenEval `correct`, or T2I-CompBench scores ≥ 0.75 for success and ≤ 0.45 for failure; intermediate cases are excluded so that every pair has an unambiguous outcome label.

Message transplantation. We transplant the I2T message $B_{i\to t}$. For success-to-failure transplantation, we cache the message of the successful donor by step and block and insert it after softmax and before output projection:

$$B^{F\leftarrow S}_{t,\ell}(\lambda) = (1-\lambda)B^{F}_{t,\ell} + \lambda B^{S}_{t,\ell}, \qquad \lambda \in [0,1], \tag{2}$$

The donor mixing weight λ ranges from zero (retain) to one (fully replace). The prompt, seed, sampler, initial latent, and other activations of the recipient remain unchanged. We measure *rescue* among initially failed recipients and *disruption* among initially successful recipients: each counts a change to the opposite outcome. Norm-matched random messages rarely change the outcome (Figure 4).

We also test failed-donor and text-to-image transplants (Table 21); both yield lower rescue rates than successful-donor image-to-text transplantation in the reported controls.

Direction and timing. On a separate fixed cohort of 400 success/failure pairs, we multiply the I2T message by α in every joint block. At each unmodified failed state, J_t is the derivative of predicted velocity with respect to this multiplier at $\alpha = 1$; $-J_t$ is the local response to reducing I2T feedback. Successful-donor transplantation changes the velocity by R^S_t, while a norm-matched random message perturbation changes it by R^{rand}_t (Figure 3b). The *cosine gap over random* compares their cosine similarities with $-J_t$, subtracting the random-control value from the successful-donor value. This gap is positive in both models across all time intervals (Figure 5). Transplanting during steps 1–8 of 40 recovers about two-thirds of the full-schedule rescues; late transplants rarely change the outcome. Definitions and aggregation details are in App. F.2.

Appendix G uses object-omission probes in the internal text stream to test whether errors persist once detected during sampling. Warnings usually persist without control; feedback suppression reduces this persistence and rescues many failures, with less frequent disruption of initially successful outputs.

2.3 METHOD: SFG AND iSFG

SFG combines image-to-text feedback control (SFG-X) with text-to-image conditioning control (SFG-S). It implements both through an auxiliary prediction and an extrapolation step. The in-place variant, iSFG, applies the corresponding message scaling directly within conditional forward pass.

Control targets and insertion point. In every joint-attention block, we scale the I2T and T2I messages after softmax and value aggregation, before output projection. TACA and Attend-and-Excite strengthen T2I conditioning (Lv et al., 2025; Chefer et al., 2023); SFG-X controls the I2T message, which neither method modifies. For a fixed block input, scaling preserves attention logits, joint normalization, and within-modality terms. Model weights stay frozen, but each forward recomputes activations, so earlier edits can affect later blocks through the updated text stream.

SFG: auxiliary-branch guidance. For edit strength $u < 0$, the auxiliary branch increases image-to-text feedback and decreases text-to-image conditioning with tied magnitudes:

$$\begin{aligned} O_t^{\text{pert}}(u) &= A_{t\to t}V_t + (1-u)B_{i\to t}, \\ O_i^{\text{pert}}(u) &= A_{i\to i}V_i + (1-|u|)B_{t\to i}. \end{aligned} \tag{3}$$

The first edit defines SFG-X, and the second defines SFG-S; SFG applies both in the same auxiliary forward. To use either component alone, leave the other message at its original multiplier of one. Setting $u = 0$ removes both edits. Text-to-image attenuation uses $|u| \leq 1$; larger magnitudes reverse the sign of that message.

We call an unmodified conditional forward at the current latent the *clean branch*. At the same x_t, t, and c, its prediction is $v_{\text{clean}} = v_{\boldsymbol{\theta}}(x_t, t, c)$; v_{pert} is the auxiliary-branch prediction with the selected message scaling. With guidance strength $\omega \geq 0$ and CFG scale $s \geq 1$,

$$\begin{aligned} \hat{v} = v_{\text{clean}} &+ \omega\,(v_{\text{clean}} - v_{\text{pert}}) \\ &+ (s-1)\,[v_{\text{clean}} - v_{\boldsymbol{\theta}}(x_t, t, \varnothing)]\,. \end{aligned} \tag{4}$$

Here u sets the message multipliers, ω is the SFG guidance scale, and s is the CFG scale. SFG-X constructs its negative branch by strengthening I2T feedback ($1 - u > 1$), rather than degrading the model. This strengthening of a native pathway takes the opposite approach to guidance methods such as PAG (Ahn et al., 2024), SEG (Hong, 2024), and PG (Chen et al., 2026b), which construct auxiliary branches by degrading the model. Guiding away suppresses this feedback, as motivated by Section 2.2, while SFG-S strengthens conditioning by guiding away from its attenuation. The sampler advances one trajectory using $\hat{v}$. At $s = 1$, it omits the unconditional evaluation; otherwise, that prediction remains unmodified.

iSFG: in-place control. iSFG edits the ordinary conditional forward directly:

$$\begin{aligned} O_t^{\text{iSFG}} &= A_{t\to t}V_t + (1+\gamma)B_{i\to t}, \\ O_i^{\text{iSFG}} &= A_{i\to i}V_i + (1+\gamma_{\text{S}})B_{t\to i}. \end{aligned} \tag{5}$$

The I2T multiplier is $1 + \gamma$ with $\gamma < 0$; the T2I multiplier is $1 + \gamma_{\text{S}}$ with $\gamma_{\text{S}} > 0$. Thus, $u < 0$ strengthens feedback in the negative branch, so guiding away suppresses its influence; $\gamma < 0$ applies suppression in place. The component ablation attributes most of the gain of SFG to feedback control (Section 3.3, Figure 7a). Let v_{ip} be the resulting conditional prediction. Without CFG, the sampler uses v_{ip} directly; with CFG, it uses $sv_{\text{ip}} - (s-1)v_{\boldsymbol{\theta}}(x_t, t, \varnothing)$. Only the conditional branch is edited. These edit strengths are independent of ω; relations between matched branch and in-place controls are given in App. A.

Sampling schedule and computational cost. SFG uses approximately $1.2K$ model evaluations without CFG and $2.2K$ with CFG, activating its auxiliary branch during the first 20% of K image sampling steps. iSFG adds no model evaluations and has minimal runtime overhead (App. D).

3 EXPERIMENTS: EVALUATING MECHANISM-GROUNDED GUIDANCE

Tasks, datasets, and metrics. We evaluate text-to-image generation on SD3M, SD3.5M, and FLUX at 1024^2, and text-to-video and image-to-video generation on HunyuanVideo (Kong et al., 2024). For images, we evaluate compositional text-image alignment on GenEval (Ghosh et al., 2023) and

Table 1: **Text-to-image generation without CFG on SD3M and FLUX.** All methods use $\mathrm{cfg} = 1$. GenEval and T2I-CompBench++ are averaged over six and eight categories, respectively. FID uses 5k COCO images on SD3M; CFG (Ho and Salimans, 2022) at $s = 7.5$ gives 32.49 for reference. Dashes denote unavailable results.

	SD3M			FLUX	
Method	GenEval ↑	T2I-Comp Bench++ ↑	FID ↓	GenEval ↑	T2I-Comp Bench++ ↑
Base	0.3294	0.3230	41.76	0.3604	0.3254
S^2-Guidance (Chen et al., 2026a)	0.4866	0.3667	26.00	0.4990	0.3900
SEG (Hong, 2024)	0.5322	0.4331	28.19	0.5548	0.4505
PAG (Ahn et al., 2024)	0.5604	0.4278	28.53	0.4866	0.3843
SFG-S (ours)	0.4678	0.4249	40.19	0.4471	0.3961
SFG-X (ours)	<u>0.5621</u>	<u>0.4505</u>	<u>24.90</u>	<u>0.6277</u>	<u>0.4515</u>
SFG (ours)	**0.5811**	**0.4760**	**24.21**	**0.6853**	**0.4727**

T2I-CompBench++ (Huang et al., 2025). We report HPSv2.1 (Wu et al., 2023) and ImageReward (Xu et al., 2023) on the same 3,200 images: our evaluation set contains 800 prompts in total across four styles, with four seeds per prompt. We also report FID on COCO for SD3M. For video, we use VBench (Huang et al., 2024), evaluating seven text-to-video dimensions and eight image-to-video dimensions, and report the aggregate total score for each task.

Baselines. Image baselines include PAG (Ahn et al., 2024), SEG (Hong, 2024), S^2-Guidance (Chen et al., 2026a), and TACA (Lv et al., 2025), with comparisons both without and on top of CFG. TACA strengthens text-to-image conditioning, the reverse information-flow direction from SFG-X, making it a close attention-rebalancing control for SFG-S. We also include Prompt Reinjection, which restores shallow text features (Yao et al., 2026). Video baselines include CFG, S^2-Guidance, and CFG-Zero* (Fan et al., 2025).

Evaluation protocol. Within each configuration, methods share prompts and seeds. GenEval comparisons and guidance-scale sweeps use 553 prompts with six seeds per prompt, giving 3,318 images per configuration. For image benchmark comparisons, the SFG branch runs during the first 20% of sampling steps. Sample counts and generation settings appear in App. B, baseline weight sweeps in App. C, and single- and two-object probe diagnostics with separate intervention schedules in App. G. In Tables 1–5, higher scores are better except for FID; bold and underlined values mark the best and second-best results within each model.

3.1 Image generation

Guidance without CFG. At $\mathrm{cfg} = 1$, SFG improves GenEval and T2I-CompBench++ on both backbones (Table 1). On SD3M, the scores rise from 0.3294 to 0.5811 and from 0.3230 to 0.4760, respectively; on FLUX, they rise from 0.3604 to 0.6853 and from 0.3254 to 0.4727. GenEval gains over the strongest competing method are 0.0207 on SD3M (PAG) and 0.1305 on FLUX (SEG). Feedback control drives most of the improvement over the base model: SFG-X alone exceeds every baseline on all reported metrics and backbones, while SFG-S alone trails PAG and SEG on GenEval. SFG also reduces SD3M COCO FID from 41.76 to 24.21, below the best competing score of 26.00 and the 32.49 obtained with CFG at $s = 7.5$, without evaluating an unconditional branch.

Guidance on top of CFG. Adding SFG to CFG achieves the highest GenEval, HPSv2.1, and ImageReward scores on both backbones (Table 2). GenEval rises from 0.6890 to 0.7110 on SD3.5M and from 0.6661 to 0.7479 on FLUX, where counting rises from 0.6458 to 0.8167 and position from 0.1850 to 0.3033; HPSv2.1 increases by 0.85 and 0.70. In contrast, PAG, SEG, and S^2-Guidance do not raise overall GenEval above CFG on either backbone, and TACA, which only strengthens text-to-image conditioning, reaches 0.6960 on SD3.5M. No baseline isolates the I2T message.

Zero-extra-NFE guidance with iSFG. Without an auxiliary branch, iSFG improves CFG on all metrics (Table 3). GenEval rises from 0.6890 to 0.7047 on SD3.5M and from 0.6661 to 0.6868 on FLUX; HPSv2.1 increases by 1.33 and 0.32, and ImageReward by 0.0748 and 0.0532. Among reported categories, counting improves most on SD3.5M, rising from 0.6458 to 0.6813. Compared

Table 2: **Text-to-image generation with CFG on SD3.5M and FLUX.** Each additional guidance method is combined with CFG. Count., Pos., and C.-Attr. denote counting, position, and color-attribute GenEval scores, selected to show categories with larger SFG gains; Overall averages all six categories. HPSv2.1 averages the four style scores; ImageReward reports the mean over the same images, with identical prompts and seeds for every method.

		GenEval ↑				HPS ↑	ImageReward ↑
Model	Method	Count.	Pos.	C.-Attr.	Overall	Avg.	Mean
SD3.5M	CFG (Ho and Salimans, 2022)	0.6458	0.2150	0.5767	0.6890	31.65	1.1102
	CFG+PAG (Ahn et al., 2024)	0.6458	0.2017	0.5233	0.6807	31.51	1.1038
	CFG+SEG (Hong, 2024)	0.6646	0.2017	0.6017	0.6890	32.11	1.1180
	CFG+S^2 (Chen et al., 2026a)	0.6458	0.1967	0.5717	0.6811	31.71	1.1152
	CFG+TACA (Lv et al., 2025)	0.6771	0.1900	0.5950	0.6960	31.58	1.1298
	CFG+SFG (ours)	**0.6875**	**0.2250**	**0.6117**	**0.7110**	**32.50**	**1.1587**
FLUX	CFG (Ho and Salimans, 2022)	0.6458	0.1850	0.5233	0.6661	30.96	0.9837
	CFG+PAG (Ahn et al., 2024)	0.5792	0.1900	0.5300	0.6628	30.83	0.9765
	CFG+SEG (Hong, 2024)	0.6271	0.1900	0.4650	0.6500	31.10	1.0201
	CFG+S^2 (Chen et al., 2026a)	0.6458	0.1733	0.5117	0.6605	31.02	0.9912
	CFG+Prompt Reinjection (Yao et al., 2026)	0.7500	0.2500	0.5233	0.6975	31.25	1.0211
	CFG+SFG (ours)	**0.8167**	**0.3033**	**0.5933**	**0.7479**	**31.66**	**1.0436**

Table 3: **Improving CFG with iSFG at zero additional NFE on SD3.5M and FLUX.** iSFG edits the conditional forward pass without increasing model evaluations relative to CFG. Count., Pos., and C.-Attr. are selected GenEval categories with larger gains: counting, position, and color attribute; Overall averages six categories. HPSv2.1 averages the four style scores; ImageReward reports the mean over the same images.

		GenEval ↑				HPSv2.1 ↑	ImageReward ↑
Model	Method	Count.	Pos.	C.-Attr.	Overall	Avg.	Mean
SD3.5M	CFG (Ho and Salimans, 2022)	0.6458	0.2150	0.5767	0.6890	31.65	1.1102
	CFG+iSFG (ours)	**0.6813**	**0.2367**	**0.5950**	**0.7047**	**32.98**	**1.1850**
FLUX	CFG (Ho and Salimans, 2022)	0.6458	0.1850	0.5233	0.6661	30.96	0.9837
	CFG+iSFG (ours)	**0.6563**	**0.2200**	**0.5417**	**0.6868**	**31.28**	**1.0369**

with SFG, iSFG gives smaller overall GenEval gains but higher preference scores on SD3.5M (32.98 vs. 32.50 HPSv2.1 and 1.1850 vs. 1.1587 ImageReward), at zero extra NFE.

3.2 Transfer to video generation

Text-to-video generation. In both video tasks, SFG runs with $\mathrm{cfg} = 1$ and requires no training. For text-to-video, seven active steps of 25 require 32 branch predictions versus 50 for full-schedule CFG and 39 for CFG limited to its first 14 steps. SFG reaches 84.25 on seven VBench dimensions, exceeding all compared methods, including the base model (82.58; Table 4). Its $+1.67$ gain over the base model is more than four times that of S^2-Guidance ($+0.38$), the other self-guidance baseline. It leads in subject consistency and imaging quality (93.71 and 66.61, versus at most 93.08 and 64.94 under CFG). CFG retains higher background consistency and aesthetics; S^2-Guidance has the highest dynamic degree. Unlike CFG, which lowers temporal flickering and motion smoothness below the base model, SFG keeps both at or near base levels (98.18 and 99.36). It also raises aesthetic quality from 50.88 to 56.00, recovering most of the gain from CFG.

Image-to-video generation. For image-to-video, reference-image and text tokens jointly form the conditioning stream. SFG applies unchanged to this entire stream, suppressing feedback from the evolving video and strengthening conditioning, and thus generalizes from text-only to multimodal conditioning. SFG scores 90.04, versus 89.43 for CFG and 88.50 for the base model, and leads in eight of the nine reported columns (Table 5). Dynamic degree rises from 66.67 under CFG to 70.44, although it remains below the base model and S^2-Guidance. SFG also achieves the highest consistency with the reference image. CFG-Zero* falls below the base model (87.49).

Table 4: **Text-to-video generation with different guidance methods on HunyuanVideo.** We report seven VBench quality dimensions and the aggregate score (Total). SFG uses $\mathrm{cfg} = 1$ without an unconditional branch; the two CFG rows differ in their guidance window. Cons., Flicker, and Smooth. denote consistency, temporal flickering, and motion smoothness.

Method	Total Score	Subject Cons.	Background Cons.	Temporal Flicker	Motion Smooth.	Dynamic Degree	Aesthetic Quality	Imaging Quality
Base (no guidance)	82.58	92.11	91.98	98.01	**99.39**	97.92	50.88	63.65
CFG (all steps) (Ho and Salimans, 2022)	83.52	92.79	92.65	97.20	98.78	96.53	**58.97**	64.94
CFG (first 14 steps) (Ho and Salimans, 2022)	83.58	93.08	**92.67**	97.38	99.00	96.53	57.81	64.90
S^2-Guidance (Chen et al., 2026a)	82.96	92.35	92.43	97.49	99.35	**98.61**	52.69	64.62
CFG-Zero* (Fan et al., 2025)	83.41	92.74	92.38	96.97	98.86	97.92	58.41	64.86
SFG (ours)	**84.25**	**93.71**	92.37	**98.18**	99.36	97.92	56.00	**66.61**

Table 5: **Image-to-video generation with different guidance methods on HunyuanVideo.** We report eight VBench dimensions and the aggregate score (Total), covering reference-image consistency (I2V Subj./Bg.) and video quality (e.g., Subject/Background Cons.).

Method	Total Score	I2V Subj.	I2V Bg.	Subject Cons.	Background Cons.	Motion Smooth.	Dynamic Degree	Aesthetic Quality	Imaging Quality
Base (no guidance)	88.50	96.51	97.92	89.83	91.73	98.99	85.37	57.62	67.88
CFG (Ho and Salimans, 2022)	89.43	97.82	98.59	92.88	94.43	99.10	66.67	60.44	70.24
S^2-Guidance (Chen et al., 2026a)	88.66	96.63	97.98	90.01	91.74	98.93	**86.59**	58.12	67.90
CFG-Zero* (Fan et al., 2025)	87.49	95.79	97.19	86.59	90.61	98.87	85.77	57.23	67.74
SFG (ours)	**90.04**	**98.16**	**99.06**	**93.04**	**94.57**	**99.50**	70.44	**60.56**	**70.44**

3.3 Ablation studies

Guidance direction. Figure 6 tests the local response direction with finite edits on SD3.5M at $\mathrm{cfg} = 1$. We compare both signs at matched magnitudes $|u| \in \{0.20, 0.25, 0.30, 0.40\}$ and scales $\omega \in \{1, 3, 5\}$. All twelve suppression settings ($u < 0$) improve on the base GenEval score of 0.3481; all twelve opposite-sign settings reduce it. Larger scales widen the separation, with the strongest opposite-sign setting falling to roughly half the base score.

Guidance components. Both SFG-X and SFG-S improve generation independently. SFG-X gives larger gains, and their combination performs best (Table 1; Figure 7a). Combining them improves GenEval over SFG-X from 0.5621 to 0.5811 on SD3M and from 0.6277 to 0.6853 on FLUX. T2I-CompBench++ also rises from 0.4505 to 0.4760 and from 0.4515 to 0.4727, respectively. SFG-S thus adds gains within the same auxiliary evaluation.

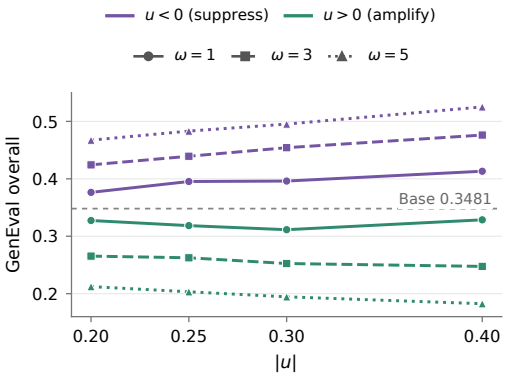


Figure 6: **Validation of the SFG-X guidance direction.** GenEval on SD3.5M ($\mathrm{cfg} = 1$), comparing SFG-X ($u < 0$) with the opposite-sign branch ($u > 0$) at matched $|u|$ and ω. Suppression/amplification labels denote the net guidance effect. Base marks the unguided score.

Guidance windows. Longer SFG guidance windows do not consistently improve generation, whereas the tested image baselines improve as their guidance windows grow (Figure 7b). FLUX CFG+SFG scores 0.7479 over steps 1–6 versus 0.7441 over all 28 steps; SD3.5M CFG+SFG scores 0.7111 over steps 1–10 versus 0.6941 over all 40 steps. By comparison, extending guidance from 10 to 28 steps raises FLUX CFG from 0.5675 to 0.6487; extending it from 10 to 40 steps raises SD3M SEG from 0.4505 to 0.5323, with monotonic gains at the tested intermediate windows. Consistent with the temporal transplantation results, early feedback control matches or exceeds full-schedule SFG while requiring fewer auxiliary evaluations.

The video results show the same pattern for SFG (Figure 7c): activating it for the first seven of 25 steps scores 84.25, compared with 83.09 over all steps, while requiring 32 rather than 50 model

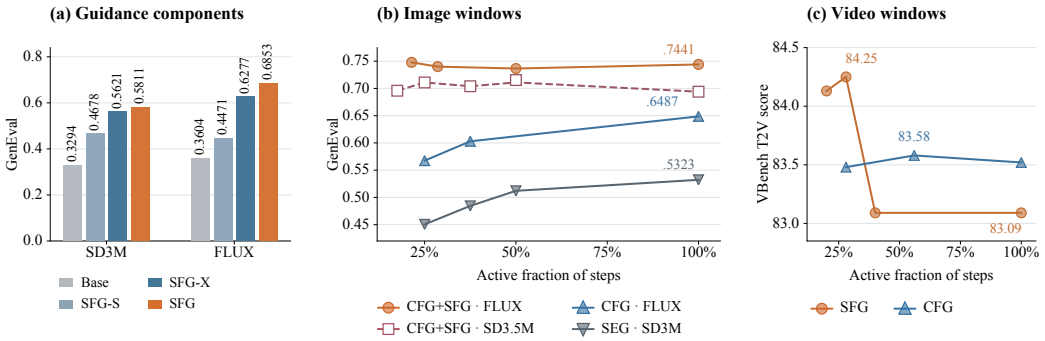


Figure 7: **Guidance components and guidance windows.** (a) GenEval at the Table 1 operating points. (b) Both FLUX curves use 28 steps; SD3.5M CFG+SFG and SD3M SEG use 40. FLUX CFG returns to scale 1 outside its guidance window. (c) HunyuanVideo T2V scores with 25 total sampling steps. SFG is active for the first 5, 7, 10, or 25 steps; CFG is active for the first 7, 14, or 25 steps. Guidance is disabled after the selected prefix.

evaluations. CFG changes little across its tested windows (83.48–83.58), remaining below the early SFG window. The useful SFG window therefore varies by task: the strongest tested T2V setting uses seven of 25 steps, rather than the first 20% used in the image benchmark comparisons. These results support concentrating feedback control early in sampling; extending it to the full trajectory can reduce quality while adding computation.

4 Related Work

Negative branches and attention strengthening. CFG contrasts conditional and unconditional predictions; training-free self-guidance extrapolates away from a perturbed model prediction (Ho and Salimans, 2022; Chung et al., 2025; Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Chen et al., 2026a; Zhang et al., 2026; Chen et al., 2026b). SFG-X constructs this reference by strengthening I2T feedback instead of degrading the model through blurred attention, identity maps, or skipped blocks, using the measured response to select the sign. Limited-interval guidance restricts CFG to an empirically chosen interval (Kynkäänniemi et al., 2024); our transplantation experiment measures when feedback most affects the outcome. Attend-and-Excite amplifies neglected subject-token activations (Chefer et al., 2023). TACA uses a timestep-aware temperature to rebalance cross-modal attention and counter text-token imbalance (Lv et al., 2025). Both strengthen T2I conditioning, as SFG-S does through guidance, but neither modifies how the image writes back into text. SFG-X instead controls the I2T direction, where our intervention sweeps support suppressing feedback.

The two streams of an MMDiT. Studies of MMDiTs examine padding registers, multimodal circuits, prompt reinjection, and attention editing (Toker et al., 2025; Mazur et al., 2026; Ge et al., 2026; Zhou et al., 2025; Shen et al., 2025; Yao et al., 2026). Toker et al. (2025) find image content in padding tokens; Li et al. (2026) find template tokens maintaining object identity; Shin et al. (2025) identify I2T as the weaker cross-modal direction and build on T2I. We test how the stored content affects generation and find that I2T feedback can preserve errors.

5 Conclusion

We show that image-to-text feedback in MMDiTs can preserve image content that conflicts with the prompt. Message transplantation, local response analysis, and matched interventions identify feedback suppression as a useful control direction. Building on these findings, SFG guides away from an auxiliary branch with stronger feedback and weaker text-to-image conditioning (about $1.2\times$ base NFE without CFG), while iSFG applies the corresponding controls directly with zero extra NFE. Experiments across three image backbones demonstrate gains in alignment and preference scores, both without and alongside CFG. The same control strategy improves the evaluated VBench scores in text-to-video and image-to-video generation.

AI Use Statement

AI tools assisted with code implementation under human instructions and with language polishing of the manuscript. The authors remain responsible for the code, analyses, and final content of the paper.

Reproducibility statement

App. B–E report the settings, the exact insertion point and pseudocode, and paired seed/metric protocols for the three public checkpoints. The complete code for this project is provided in the supplementary material.

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 *European Conference on Computer Vision (ECCV)*, 2024.

Kanghyun Baek, Jaihyun Lew, Chaehun Shin, Jungbeom Lee, and Sungroh Yoon. Diagnosing and correcting concept omission in multimodal diffusion transformers. In *International Conference on Machine Learning (ICML)*, 2026.

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

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4), 2023.

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2026a.

Zehua Chen, Yucheng Yang, Binjie Yuan, Kaiwen Zheng, Jun S. Liu, and Jun Zhu. GuidedBridge: Training-freely improving bridge models with prior guidance. In *International Conference on Machine Learning (ICML)*, 2026b.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 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.

Chris Ge, Rohit Gandikota, Antonio Torralba, and Tamar Rott Shaham. Vision-language binding in in-context image generation. *arXiv preprint arXiv:2605.24624*, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

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

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Maohua Li, Qirui Li, Yanke Zhou, Yiduo Li, Zhaosheng Chi, Chao Xu, Cuifeng Shen, Yixuan Xu, Hanlin Tang, Kan Liu, Tao Lan, Lin Qu, and Shao-Qun Zhang. Text template tokens are implicit semantic registers in diffusion transformers. *arXiv preprint arXiv:2607.19139*, 2026.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024.

Zhengyao Lv, Tianlin Pan, Chenyang Si, Zhaoxi Chen, Wangmeng Zuo, Ziwei Liu, and Kwan-Yee K. Wong. Rethinking cross-modal interaction in multimodal diffusion transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Artyom Mazur, Nina Konovalova, and Aibek Alanov. DifFRACT: Diffusion feature reconstruction and attribution for circuit tracing. *arXiv preprint arXiv:2606.15796*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Tiancheng Shen, Zilong Huang, Xiangtai Li, Zhijie Lin, Jiyang Liu, Yitong Wang, Jiashi Feng, Ming-Hsuan Yang, and Jun Hao Liew. QK-Edit: Revisiting attention-based injection in MM-DiT for image and video editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Joonghyuk Shin, Alchan Hwang, Yujin Kim, Daneul Kim, and Jaesik Park. Exploring multimodal diffusion transformers for enhanced prompt-based image editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Bingda Tang, Boyang Zheng, Sayak Paul, and Saining Xie. Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik, and Yonatan Belinkov. Padding tone: A mechanistic analysis of padding tokens in T2I models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, 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.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuxuan Yao, Yuxuan Chen, Hui Li, Kaihui Cheng, Qipeng Guo, Yuwei Sun, Zilong Dong, Jingdong Wang, and Siyu Zhu. Prompt reinjection: Alleviating prompt forgetting in multimodal diffusion transformers. In *International Conference on Machine Learning (ICML)*, 2026.

Weijia Zhang, Yuehao Liu, Shanyan Guan, Wu Ran, Yanhao Ge, Wei Li, and Chao Ma. Guiding a diffusion model by swapping its tokens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3DIS-FLUX: Simple and efficient multi-instance generation with DiT rendering. *arXiv preprint arXiv:2501.05131*, 2025.

A Theoretical analysis of path interventions

Guidance conventions.

CFG uses condition dropout during training and extrapolates between conditional and unconditional predictions; AG instead contrasts well-trained and under-trained models (Ho and Salimans, 2022; Karras et al., 2024). Training-free self-guidance perturbs one pretrained model:

$$v_{\text{CFG}} = v_{\text{uncond}} + s(v_{\text{clean}} - v_{\text{uncond}}), \tag{6}$$

$$v_{\text{guide}} = v_{\text{clean}} + \omega(v_{\text{clean}} - v_{\text{pert}}). \tag{7}$$

Here v_{clean} is the clean-branch prediction, v_{uncond} is the unconditional prediction, s is the CFG scale, and v_{pert} supplies the negative branch. SAG blurs image regions, PAG replaces attention with the identity, SEG smooths attention, S^2-Guidance drops blocks, and SSG swaps extreme text tokens (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Chen et al., 2026a; Zhang et al., 2026). We choose the perturbation by measuring a native feedback path.

Local relation between SFG and iSFG. Let $\mathbf{a} = (a_X, a_S)$ denote the I2T and T2I message multipliers: $(1-u, 1-|u|)$ in the SFG auxiliary branch, $(1+\gamma, 1+\gamma_S)$ in the iSFG conditional branch, and $(\alpha, 1)$ in the local-response experiment. At fixed latent, time, and prompt, write the prediction as $v(\mathbf{a})$. The clean branch uses $\mathbf{a}_0 = (1,1)$, and the auxiliary branch uses $\mathbf{a}_0 + \mathbf{h}$, where $\mathbf{h} = (d_X, -d_S)$ and $d_X, d_S \geq 0$. For a smooth prediction, the first-order correction is

$$\omega[v(\mathbf{a}_0) - v(\mathbf{a}_0 + \mathbf{h})] = -\omega d_X J_X + \omega d_S J_S + O(\omega\|\mathbf{h}\|^2), \tag{8}$$

where J_X and J_S are the derivatives with respect to the two message multipliers at $\mathbf{a}_0$. SFG-X uses $d_S = 0$, and full SFG uses $d_X = d_S = |u|$. Setting the message multipliers to $\mathbf{a}_0 - \omega\mathbf{h}$ gives the same first-order correction in place. With CFG scale $s > 0$ and an unchanged unconditional branch, the matching conditional-branch multipliers are $\mathbf{a}_0 - (\omega/s)\mathbf{h}$. This local relation motivates iSFG; the separately tuned methods need not agree at finite strengths.

Connection to the measured response. The image-to-text derivative J_X corresponds to J in the alignment experiment. Appendix F.2 gives its centered-difference estimate and the successful-donor velocity response R^S. Reducing feedback moves locally along $-J$. A positive cosine gap over random indicates better relative alignment; it does not establish positive absolute cosine or improved final quality. Final outcomes and intervention windows are evaluated by transplantation and the sign ablation, rather than inferred from the local expansion. Probe definitions and their aggregation are given with the probe results in App. G.

$$\begin{aligned} J_t &= \partial_\alpha v(x_t^F, t, c; \alpha)\Big|_{\alpha=1}, \\ R_t^S &= v(x_t^F, t, c; B_t \leftarrow B_t^S) - v(x_t^F, t, c; B_t \leftarrow B_t^F), \\ \cos(-J_t, R_t^S) &= \frac{\langle -J_t, R_t^S \rangle}{\|J_t\|_2 \, \|R_t^S\|_2}, \\ \Delta\cos_t &= \cos(-J_t, R_t^S) - \cos(-J_t, R_t^{\text{rand}}). \end{aligned} \tag{9}$$

Control-direction measurements. Across 256 prompts and 40 sampling steps, SFG-X and SFG-S have low mutual cosine similarity; their similarities with CFG also remain well below one (Figure 8). These distinct correction directions are consistent with the gains from combining the controls.

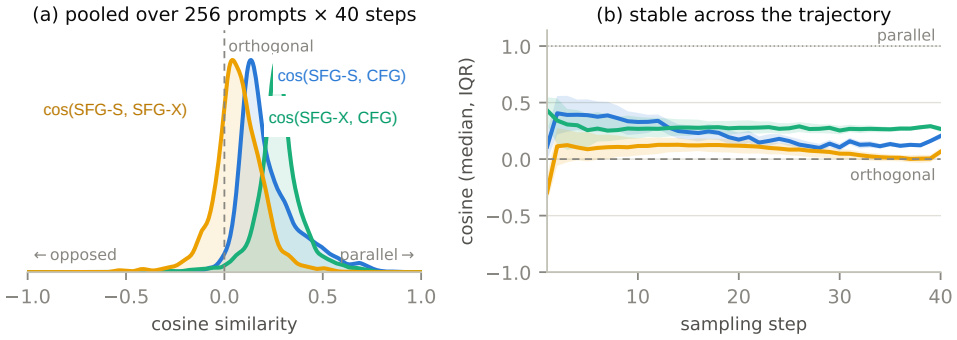


Figure 8: **Complementary directions.** SFG-X and SFG-S corrections have low cosine similarity with CFG and with each other, both over prompts and through sampling.

B Benchmark evaluation and parameter settings

Within each model and evaluation configuration, all methods share the prompts, seed set, and initial noise. Table 6 lists distinct prompt counts and generation seeds separately. Each prompt–seed pair produces one sample; for I2V, each prompt also has a fixed reference image. ImageReward scores the same generated images as HPSv2.1.

Table 6: **Evaluation sample counts per model and method configuration.** Seeds/prompt gives the number of generation seeds for each distinct prompt, or prompt–reference-image pair for I2V. The 800 HPSv2.1 prompts are the total across four styles; four seeds per prompt yield 3,200 images. ImageReward reuses these images without additional generation.

Benchmark	Prompts	Seeds/prompt	Total outputs
GenEval	553	6	3,318
T2I-CompBench++	300	8	2,400
HPSv2.1	800	4	3,200
ImageReward	800	4	3,200
VBench (T2V)	326	3	978
VBench (I2V)	355	3	1,065

Generation settings. Images use 1024^2 resolution. Image benchmarks and guidance-window ablations use 28 steps for FLUX and 40 for SD3M/SD3.5M; the transplantation experiments in App. F.1 use 40 steps for both models. HunyuanVideo T2V/I2V use 848×480, 121 frames, 24 fps, 25 steps, Flow shift 5.0, and BF16; I2V uses matched 16:9 reference crops. Their $(|u|, \omega)$ settings are $(0.25, 4)$ and $(0.35, 6)$, respectively. Explicit SFG/SFG-X acts in all joint-attention layers over a guidance window. Image benchmark comparisons use the first 20% of steps; the video window is set for the task, with seven of 25 steps used for the main T2V result. iSFG acts at every step.

SFG guidance-scale sweeps. We report recorded evaluations across guidance scales ω. Curves connect the recorded points in order of guidance scale. Bold entries and plot annotations mark the highest recorded score in each table. Each point uses the seed set specified in Table 6; GenEval uses six seeds per prompt.

Table 7: **SFG scale sweep on SD3M without CFG** (Ho and Salimans, 2022). GenEval versus guidance scale ω; bold marks the best tested score.

ω	GenEval ↑
5	**0.5811**
6	0.5787
7	0.5751
8	0.5730

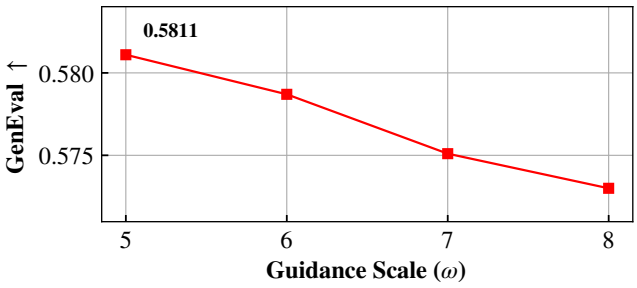


Figure 9: SD3M without CFG / GenEval guidance-scale sweep.

Table 8: **SFG scale sweep on FLUX without CFG** (Ho and Salimans, 2022). GenEval versus ω, with built-in guidance set to 1; bold marks the best tested score.

ω	GenEval ↑
7	0.6065
9	0.6233
10	0.6475
11	0.6655
12	0.6739
13	0.6779
14	**0.6853**

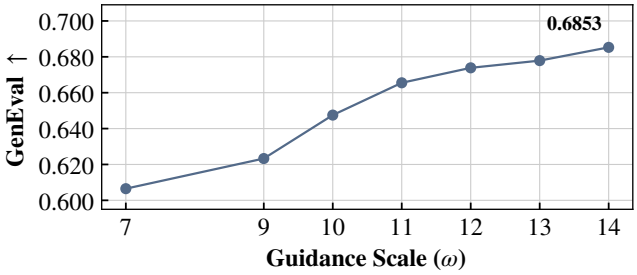


Figure 10: FLUX without CFG / GenEval guidance-scale sweep.

The non-CFG records use FLUX.1-dev with built-in guidance 1. **FLUX stacked on CFG.** We report GenEval and HPSv2.1 separately. The GenEval records use de-distilled FLUX with explicit CFG scale 3.5.

Table 9: **SFG scale sweep on FLUX with CFG** (Ho and Salimans, 2022). GenEval versus ω at CFG scale 3.5; bold marks the best tested score.

ω	GenEval ↑
7	0.7421
7.5	**0.7479**
8	0.7432
8.5	0.7442

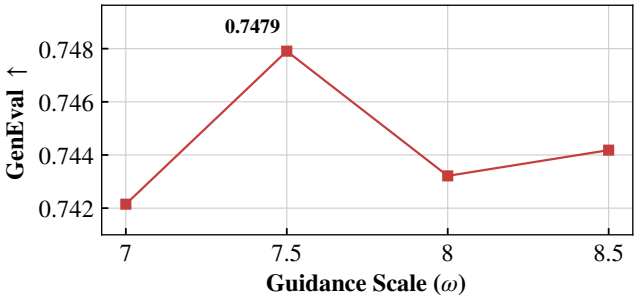


Figure 11: FLUX with CFG / GenEval guidance-scale sweep.

Table 10: **SFG scale sweep on FLUX with CFG** (Ho and Salimans, 2022). HPSv2.1 versus ω; bold marks the best tested score.

ω	HPSv2.1 ↑
5	31.54
5.5	31.54
6	31.58
6.5	31.57
7.5	**31.66**

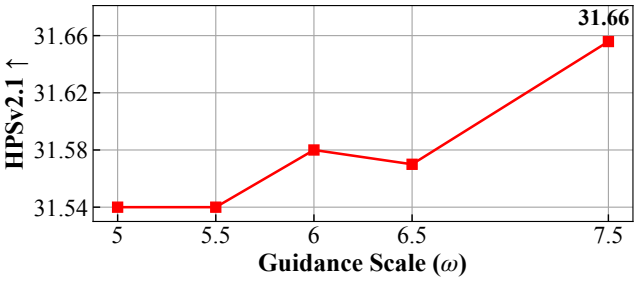


Figure 12: FLUX with CFG / HPSv2.1 guidance-scale sweep.

C Baseline hyperparameter sweeps

Each GenEval sweep point uses 553 prompts with six seeds per prompt (3,318 images), matching App. B, at 1024×1024 resolution with the official Mask2Former evaluator and 40 steps for SD3M/SD3.5M or 28 for FLUX. **Negative-branch operators:** PAG replaces selected attention maps by identity ($A \leftarrow I$), removing cross-token mixing (Ahn et al., 2024); SEG Gaussian-blurs spatial queries ($Q \leftarrow G_\sigma * Q$), smoothing attention logits before softmax (Hong, 2024); S^2 stochastically skips Transformer blocks (Chen et al., 2026a). Our sweeps fix PAG at block 8, SEG at blocks 7/8/9 with $\sigma = 9{,}999{,}999$, and S^2 to the all-block range. Only w varies within each curve. Five-point displays retain the tested endpoints and observed maxima; SD3.5M PAG/SEG instead show six/three points. Bold entries mark observed maxima; source records accompany the manuscript. The updated main-table GenEval rows use the best confirmed sweep points.

C.1 SD3M without CFG

The CFG scale is 1.0 and the matched GenEval reference is 0.3294. S^2 uses drop 1 during steps 5–36. All three methods improve on the reference; PAG reaches 0.5604 at $w = 3$ among the completed fixed-configuration points.

Table 11: **PAG on SD3M without CFG.** (Ahn et al., 2024) GenEval versus guidance weight w; bold marks the best tested setting ($w = 3$).

Scale (w)	GenEval ↑
0.5	0.4696
1.25	0.5521
2.5	0.5580
3	**0.5604**
4	0.5563

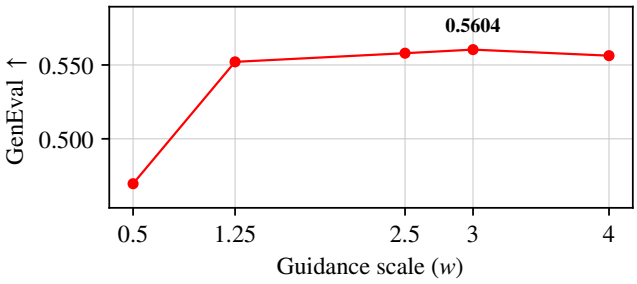


Figure 13: **PAG.** Five-point weight sweep.

Table 12: **SEG on SD3M without CFG.** (Hong, 2024) GenEval versus guidance weight w; bold marks the best tested setting ($w = 1$).

Scale (w)	GenEval ↑
0.2	0.4264
0.5	0.4684
1	**0.5322**
1.25	0.5303
3	0.5212

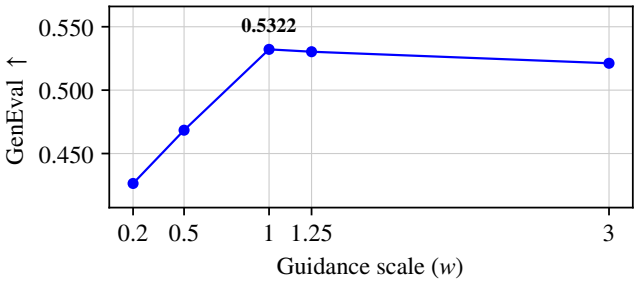


Figure 14: **SEG.** Five-point weight sweep.

Table 13: **S^2-Guidance on SD3M without CFG.** (Chen et al., 2026a) GenEval versus guidance weight w; bold marks the best tested setting ($w = 2$).

Scale (w)	GenEval ↑
0.2	0.3814
0.5	0.4415
1	0.4646
2	**0.4866**
3	0.4754

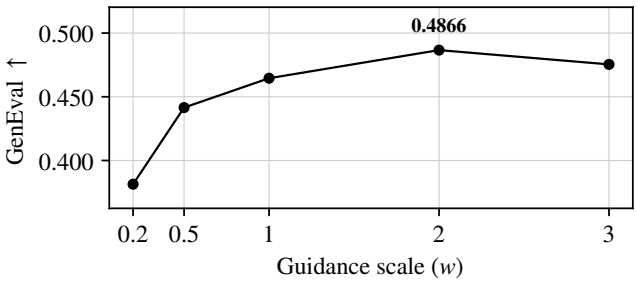


Figure 15: **S^2.** Five-point weight sweep.

C.2 SD3.5M WITH CFG

The CFG scale is 7.5 and the matched GenEval reference is 0.6890. S^2 uses drop 3 during steps 5–36. PAG peaks at 0.6807 for $w = 0.1$, and the displayed S^2 maximum is 0.6811; both remain below the reference. SEG reaches 0.6890 at $w = 0.2$, close to the reference.

Table 14: **PAG on SD3.5M with CFG.** (Ahn et al., 2024) GenEval versus guidance weight w; bold marks the best tested setting ($w = 0.1$).

Scale (w)	GenEval ↑
0.05	0.6729
0.1	**0.6807**
0.2	0.6669
0.75	0.6650
1	0.6568
1.5	0.6499

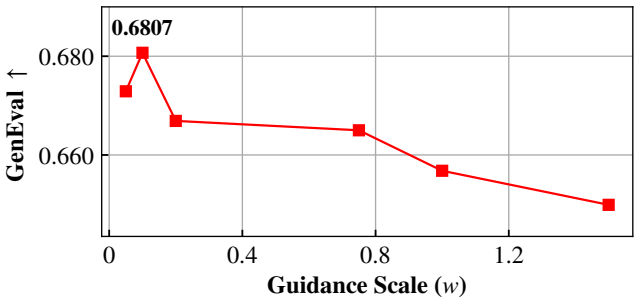


Figure 16: **PAG.** Six-point weight sweep.

Table 15: **SEG on SD3.5M with CFG.** (Hong, 2024) GenEval versus guidance weight w; bold marks the best tested setting ($w = 0.2$).

Scale (w)	GenEval ↑
0.05	0.6729
0.2	**0.6890**
0.5	0.6741

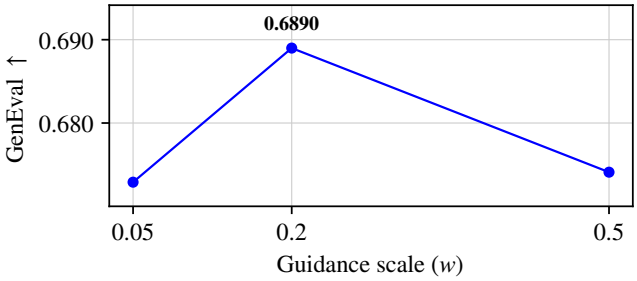


Figure 17: **SEG.** Three-point weight sweep.

Table 16: **S^2-Guidance on SD3.5M with CFG.** (Chen et al., 2026a) GenEval versus guidance weight w; bold marks the best tested setting ($w = 0.25$).

Scale (w)	GenEval ↑
0.25	**0.6811**
1.5	0.6274
2	0.5846
3	0.4977
4.5	0.4227

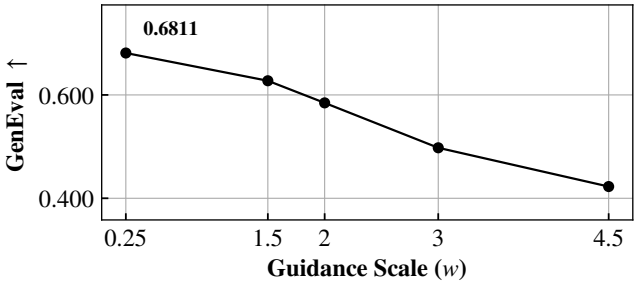


Figure 18: **S^2.** Five-point weight sweep.

C.3 Original FLUX.1-dev (g = 1.0)

The original FLUX.1-dev model uses $g = 1.0$, with a matched GenEval reference of 0.3604. S^2 uses drop 5 during steps 4–25. All three methods improve on the reference; SEG reaches 0.5548 at $w = 3$. These points use the original model and are separate from de-distilled FLUX.

Table 17: **PAG on FLUX without CFG.** (Ahn et al., 2024) GenEval versus guidance weight w; bold marks the best tested setting ($w = 2$).

Scale (w)	GenEval ↑
1	0.4626
1.5	0.4728
2	**0.4866**
2.5	0.4706
3	0.4757

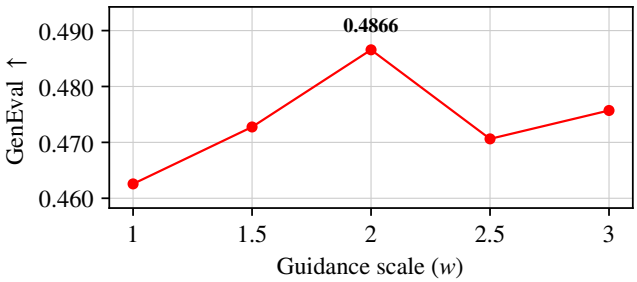


Figure 19: **PAG.** Five-point weight sweep.

Table 18: **SEG on FLUX without CFG.** (Hong, 2024) GenEval versus guidance weight w; bold marks the best tested setting ($w = 3$).

Scale (w)	GenEval ↑
2	0.5315
2.5	0.5295
3	**0.5548**
3.5	0.5544
4	0.5525

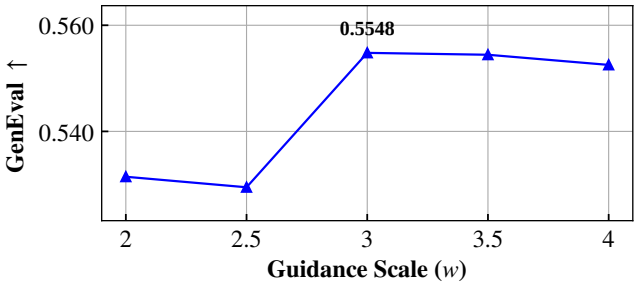


Figure 20: **SEG.** Five-point weight sweep.

Table 19: **S^2-Guidance on FLUX without CFG.** (Chen et al., 2026a) GenEval versus guidance weight w; bold marks the best tested setting ($w = 2$).

Scale (w)	GenEval ↑
1	0.4876
2	**0.4990**
2.5	0.4709
3	0.4548
3.5	0.4347

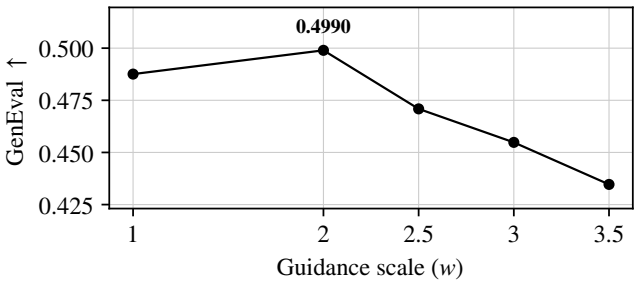


Figure 21: **S^2.** Five-point weight sweep.

D Implementation details and pseudocode

All experiments use one attention-processor patch on the public `diffusers` implementations of SD3-Medium, SD3.5-Medium, and FLUX. The patch touches nothing but the fused scaled-dot-product attention (SDPA) call of each joint-attention block:

```
# One MMDiT joint-attention block. X: image tokens, C: text tokens.
Qx, Kx, Vx = qkv_image(X)              # modality-specific projections
Qc, Kc, Vc = qkv_text(C)
Q, K, V    = concat(Qx,Qc), concat(Kx,Kc), concat(Vx,Vc)
O          = SDPA(Q, K, V)             # one joint softmax

# Cross-modal term of a chosen direction, at the SAME softmax weights:
# zero every value row outside the source modality, rerun SDPA on the
# target-query slice only.
B_i2t = SDPA(Q[text],  K, V * mask(image))  # = A_{i->t} V_i
B_t2i = SDPA(Q[image], K, V * mask(text))   # = A_{t->i} V_t

O[text]  += (a_X - 1) * B_i2t  # I2T message multiplier a_X
O[image] += (a_S - 1) * B_t2i  # T2I message multiplier a_S
# ... followed by the block's usual output projections.
```

Because the mask only zeroes value rows, the softmax weights are exactly those of the unmodified forward, and the added term is exactly $A_{\text{source}\to\text{target}}V_{\text{source}}$. Adding $(a_X-1)B_{i\to t}$ or $(a_S-1)B_{t\to i}$ sets the corresponding message multiplier to a_X or a_S. The sampler-level combination is then

```
# Explicit SFG: use the guidance window selected for the task.
# K_act: e.g. 8/40 image steps or 7/25 T2V steps.
if step < K_act:
    v_clean = v_theta(x_t, t, c)
    # Full SFG (u < 0); SFG-X alone sets a_S = 1
    v_pert  = v_theta(x_t, t, c, a_X=1-u, a_S=1-abs(u))
    v_hat   = v_clean + omega * (v_clean - v_pert)  # + CFG if stacked
else:
    v_hat   = v_theta(x_t, t, c)       # ordinary forward, no extra NFE
# iSFG: gamma < 0 for I2T, gamma_S > 0 for T2I
v_ip = v_theta(x_t, t, c, a_X=1+gamma, a_S=1+gamma_S)
# Use v_ip directly, or combine with the unedited unconditional
# prediction through CFG; no auxiliary branch is evaluated.
```

The path edit is applied in every joint-attention block (`layers=all`). For explicit SFG, the negative branch is evaluated only during the selected guidance window, batched with the clean branch in a single fused forward; the remaining steps run the ordinary forward with no extra evaluation. The separate probe diagnostics in App. G use their explicitly specified windows, including full-schedule guidance. The in-place iSFG edit runs at every step at zero extra NFE. A norm-clipping safeguard on the residual ($\text{clip} = \min(1, \tau\|v_{\boldsymbol{\theta}}(x_t, t, c; \mathbf{1})\|/\|\Delta\|)$) exists in the implementation but is disabled (τ=999) in the benchmark comparisons.

D.1 Low-overhead iSFG implementation

The optimized implementation folds the message multipliers into value aggregation, without a separate auxiliary Transformer branch or recomputation of the cross-modal messages above. Image queries use the original image values and text values scaled by 1.20; text queries use the original text values and image values scaled by 0.80. Both query groups attend to all keys with the original joint-softmax normalization, and their outputs are concatenated in token order. The disjoint query groups cover the original joint attention without duplicating its attention-score computation. Under CFG, only the conditional branch is edited. Remaining overhead comes from value copies, scaling, and kernel dispatch.

We time SD3.5M on one A800 80GB GPU with BF16, 1024^2 resolution, 40 steps, CFG scale 7.5, and batch size 1. iSFG uses $(\gamma_S, \gamma) = (+0.20, -0.20)$ in all joint blocks at all steps. After one warm-up per method, we alternate three paired runs with matched prompts and seeds. Timing includes text encoding, denoising, and VAE decoding, excluding model loading and image writes. Both methods make 40 batched conditional/unconditional Transformer calls. iSFG adds about 0.31 seconds per image (2.86%), with zero extra NFE.

Table 20: **iSFG runtime.** Mean $\pm$ standard deviation over three paired runs.

Method	Extra NFE	Time per image (s)	Relative time
CFG	0	10.800 ± 0.020	$1.000\times$
CFG+iSFG (ours)	0	11.108 ± 0.018	$1.029\times$

E Paired sampling protocol

For each benchmark and backbone, all baselines and our methods use the same prompts and seed set. Each prompt is generated with multiple seeds before scoring; every generated sample is evaluated, and scores are aggregated according to the benchmark protocol. GenEval uses 553 prompts with six seeds per prompt (3,318 images per configuration); the counts for other benchmarks are listed in Table 6. Methods share the initial noise for each prompt–seed pair, and I2V also fixes the reference image. For each method on a given backbone, we use the same hyperparameters across all applicable benchmarks, with no benchmark-specific tuning. Every hyperparameter-sweep point is evaluated on the full test sets of all applicable benchmarks using the same seed set. Mechanism diagnostics retain their separately stated sampling settings.

Seed variability. The reported standard deviations across seeds are 0.031 for VBench and 0.00332 for GenEval, in the same units as their reported overall scores.

F Full-model intervention and sign checks

F.1 I2T message and temporal transplantation

The transplantation experiment uses 400 same-prompt success/failure pairs across SD3.5M and FLUX, at 1024^2, $\mathrm{cfg} = 1$, and 40 steps, editing all joint-attention blocks. Main-text success criteria are rechecked during replay; conditions share the recipient prompt, initial latent, and sampler. Let k, t_k, ℓ denote sampling step, scheduler timestep, and block, and d_h the head dimension. With image tokens preceding text tokens,

$$K = [K_i; K_t], \qquad V = [V_i; V_t], \qquad A_t = \mathrm{softmax}\Big(Q_t K^\top / \sqrt{d_h}\Big). \tag{10}$$

$$B_{i\to t} = A_t[V_i; 0], \qquad O_t = B_{i\to t} + A_t[0; V_t]. \tag{11}$$

$$\widetilde{O}_t = O_t - B + \widetilde{B}. \tag{12}$$

B retains the full joint-softmax denominator. Only its image-to-text contribution is replaced after value aggregation and before text output projection. Cache $B^S_{k,\ell}$ on the native successful rollout and align by step/block; do not recompute it with recipient queries. For enabled steps/blocks $\mathcal{W}, \mathcal{L}$ and current recipient message $B^{\mathcal{W}}_{k,\ell}$,

$$\widetilde{B}^{\mathcal{W}}_{k,\ell} = B^{\mathcal{W}}_{k,\ell} + \lambda\, \mathbf{1}\{k \in \mathcal{W}, \ell \in \mathcal{L}\}(B^S_{k,\ell} - B^{\mathcal{W}}_{k,\ell}). \tag{13}$$

$$x^{\mathcal{W}}_{k+1} = \mathrm{Step}\big(x^{\mathcal{W}}_k, t_k, v(x^{\mathcal{W}}_k, t_k, c; \widetilde{B}^{\mathcal{W}}_k)\big). \tag{14}$$

$\lambda = 1$ fully transplants and $\lambda = 0$ gives the baseline. Restart each condition from the same failed initial state; outside the window use the current recipient message. Steps run from noise (1) to image (40). Full, window, prefix, and suffix sets are $[1, K]$, $[a, b]$, $[1, b]$, and $[a, K]$; Figure 5 uses eight-step windows. With benchmark success criterion E_i, evaluate each window separately:

$$\mathrm{Rescue}(\mathcal{W}) = \frac{\sum_i \mathbf{1}\{E_i(I_i^F) = 0,\ E_i(I_i^{\mathcal{W}}) = 1\}}{\sum_i \mathbf{1}\{E_i(I_i^F) = 0\}}. \tag{15}$$

Donor and path controls. For the failed-donor control, we form 200 failure/failure pairs and transplant $B_{i\to t}$ from one failed trajectory into the other. This rescues $38/200$ recipients (19.0%) over steps 1–40. For the path control, we retain the same 400 success/failure pairs and successful-donor/failed-recipient assignments, replacing the text-to-image message:

$$A_i = \mathrm{softmax}\Big(Q_i K^\top / \sqrt{d_h}\Big), \qquad B_{t\to i} = A_i[0; V_t]. \tag{16}$$

For this text-to-image (image-read-text) control, full replacement ($\lambda = 1$) over steps 1–40 rescues $74/400$ recipients (18.5%). Over steps 1–8, rescue is $151/400$ (37.8%) at $\lambda = 1$ and $118/400$ (29.5%) at $\lambda = 0.5$. Table 21 reports the window and outcome for each condition.

Table 21: **Message transplantation and controls on paired generation trajectories.** S and F denote successful and failed trajectories; arrows run from donor to recipient. Image-to-text and text-to-image transplantation use the same 400 success/failure pairs, while failed-to-failed transplantation uses 200 failure/failure pairs. λ is the message mixing weight. Steps specify the transplant window. Counts and percentages report rescued failures or disrupted successes.

Message path	Donor → recipient	Steps	Outcome	Count (rate)
Image-to-text	$S \to F$, $\lambda = 1$	1–40	Rescue	270/400 (67.5%)
Image-to-text	$S \to F$, $\lambda = 0.5$	1–40	Rescue	207/400 (51.8%)
Image-to-text	Random, $\lambda = 0.5$	1–40	Rescue	5/400 (1.3%)
Image-to-text	$F \to F$	1–40	Rescue	38/200 (19.0%)
Text-to-image	$S \to F$, $\lambda = 1$	1–40	Rescue	74/400 (18.5%)
Text-to-image	$S \to F$, $\lambda = 1$	1–8	Rescue	151/400 (37.8%)
Text-to-image	$S \to F$, $\lambda = 0.5$	1–8	Rescue	118/400 (29.5%)
Image-to-text	$F \to S$, $\lambda = 1$	1–40	Disruption	242/400 (60.5%)
Image-to-text	$F \to S$, $\lambda = 0.5$	1–40	Disruption	232/400 (58.0%)

F.2 Local response alignment and aggregation

All response measurements share the native failed latent x_k^F, timestep, embeddings, and weights. Each block scales its current message as $O_t(\alpha) = O_t - B + \alpha B$. J differentiates the full prediction with respect to I2T multiplier α; its clean-branch value is one. The finite-difference step size is $\delta = 0.05$ in Figure 5:

$$\begin{aligned} v_k^0 &= v(x_k^F, t_k, c; 1), \quad v_k^\pm = v(x_k^F, t_k, c; 1 \pm \delta), \\ J_k^{(\delta)} &= \frac{v_k^+ - v_k^-}{2\delta}, \qquad J_k = \partial_\alpha v(x_k^F, t_k, c; \alpha)\big|_{\alpha=1}. \end{aligned} \tag{17}$$

$$x_{k+1}^F = \mathrm{Step}(x_k^F, t_k, v_k^0). \tag{18}$$

Here v_k^0 is the clean-branch prediction at the failed latent. Only this prediction advances the original trajectory; the other forwards measure local responses. Anchor the fixed-seed random control to B^F recorded in a separate baseline forward and match message-space norms per step/block:

$$\begin{aligned} D_{k,\ell} &= B_{k,\ell}^S - B_{k,\ell}^F, \qquad U_{k,\ell} \sim \mathcal{N}(0, I), \\ \xi_{k,\ell} &= U_{k,\ell} \frac{\|D_{k,\ell}\|_F}{\max(\|U_{k,\ell}\|_F, \epsilon)}, \qquad B_{k,\ell}^{\mathrm{rand}} = B_{k,\ell}^F + \xi_{k,\ell}. \end{aligned} \tag{19}$$

$$\begin{aligned} R_k^S &= v(x_k^F, t_k, c; B \leftarrow B_k^S) - v_k^0, \\ R_k^{\mathrm{rand}} &= v(x_k^F, t_k, c; B \leftarrow B_k^{\mathrm{rand}}) - v_k^0. \end{aligned} \tag{20}$$

$$c_k^q = \frac{\langle -J_k^{(\delta)}, R_k^q \rangle}{\|J_k^{(\delta)}\|_2 \, \|R_k^q\|_2}, \quad q \in \{S, \mathrm{rand}\}, \qquad \Delta\cos_k = c_k^S - c_k^{\mathrm{rand}}. \tag{21}$$

Both corrections replace all tested blocks at the same x_k^F; cosines use flattened predictions and the reduced-feedback direction $-J$. Reverse checks differentiate at native successful states, anchor the random control to B^S, and match $\|B^F - B^S\|_F$, without post-hoc sign reversal:

$$R_k^F = v(x_k^S, t_k, c; B \leftarrow B_k^F) - v(x_k^S, t_k, c; 1). \tag{22}$$

$$\overline{\Delta\cos}_i(\mathcal{W}) = \frac{1}{|\mathcal{W}|} \sum_{k \in \mathcal{W}} \Delta\cos_{i,k}, \qquad \overline{\Delta\cos}_m(\mathcal{W}) = \frac{1}{N_m} \sum_{i \in \mathcal{I}_m} \overline{\Delta\cos}_i(\mathcal{W}). \tag{23}$$

$\mathcal{I}_m$ contains N_m pairs for model m. Average steps within pairs, then pairs within models; Figure 5 weights models equally. Bootstrap and sign tests use pairs. For post-start analysis, obtain the start only from independent suffix transplants:

$$a_i^* = \max\{a : E_i(I_i^{\{a,\ldots,K\}}) = 1\}, \tag{24}$$

Use $k \geq a_i^*$; no successful suffix means no start. Check baseline equivalence for recording, $\lambda = 0$, successful self-injection, and $\alpha = 1$, plus cache alignment, unchanged image-target outputs, matched norms, and finite values.

F.3 Two-sided sign check

Figure 6 in the main text tests whether the same-state sign survives finite interventions and changes final outcomes. Both directions are run at matched strengths on SD3.5M (cfg $= 1$; GenEval: 553 prompts with six seeds per prompt, totaling 3,318 images; base 0.3481), using SFG-X in all joint-attention blocks and with the first-20% schedule. All twelve feedback-suppression settings ($u < 0$ in the auxiliary branch) beat the base; the opposite-sign settings ($u > 0$) amplify feedback in the net guidance effect and fall below the base, with lower GenEval scores as ω grows.

G Single- and two-object omission probes

This appendix tests whether image-to-text feedback preserves existing errors and whether suppressing it helps generation recover. We use object-omission probes on single- and two-object prompts to track error signals in the evolving text stream and compare them with final-image outcomes. Without control, warnings tend to persist; with SFG-X, their persistence falls and many warned failures recover. Suppression can also disrupt initially successful outputs, so we measure both effects in the paired cohort: 72.0% of failures are rescued, whereas 16.7% of successes are disrupted.

G.1 Probe protocol and paired outcomes

We test whether object omissions persist and how I2T feedback suppression changes probe readouts and final images. We reuse OSI's released SD3.5M per-head linear classifiers(Baek et al., 2026), without retraining: read normalized, projected text keys for noun T5 tokens after the first 77 CLIP tokens, average noun subtokens, then take the median omission probability over the top 100 heads by released accuracy. They span 14 layers (zero-based 9–20, 22, 23). Track each concept separately and probe guided paths on the clean branch at their current latent.

All runs use SD3.5M, CFG 1.0, 40 steps, 1024×1024, and SFG-X, with $u_X = -0.4, \omega = 8$ in all joint blocks. The 500-pair run enables steps 1–40 and uses official GenEval Mask2Former evaluation. Its 80 prompts yield 233 single-/267 two-object success/failure pairs and 233/306 concept trajectories (539 total); both seeds receive baseline and guided rollouts. These are offline mechanism diagnostics with fixed schedules, not warning-triggered guidance.

A strict symptom exceeds the step-specific 75th percentile of baseline-success readouts and the paired success readout by at least 0.05. Three consecutive symptomatic steps confirm a warning; onset/confirmation are the first/third step, with medians 8/10. By step 20, onset covers 315/539 (58.4%), but confirmation covers 303/539 (56.2%). This is paired coverage, not held-out detection accuracy.

Table 22: **Omission warnings, generation outcomes, and probe states on 500 SD3.5M trajectory pairs.** Single and Two denote single- and two-object prompts; All combines both. Guided uses SFG-X throughout 40 sampling steps at $\mathrm{cfg} = 1$. Count-based rows report events over the eligible concepts or paths, with percentages in parentheses. Symptom persistence and terminal non-recovery are percentages averaged with equal prompt weights.

Metric	Condition	Single	Two	All
Warning coverage (concepts)	Any step	166/233 (71.2%)	224/306 (73.2%)	390/539 (72.4%)
	By step 20	118/233 (50.6%)	185/306 (60.5%)	303/539 (56.2%)
Failure rescue (paths)	Guided	188/223 (84.3%)	161/262 (61.5%)	349/485 (72.0%)
Success disruption	Guided	13/228 (5.7%)	68/258 (26.4%)	81/486 (16.7%)
Final failure after warning	Base	158/166 (95.2%)	201/205 (98.0%)	359/371 (96.8%)
	Guided	31/166 (18.7%)	81/205 (39.5%)	112/371 (30.2%)
Warned-failure rescue	Guided	127/158 (80.4%)	120/201 (59.7%)	247/359 (68.8%)
Symptom persistence	Base	80.3	71.2	75.4
	Guided	24.2	30.4	27.6
No terminal recovery	Base	76.9	59.5	67.4
	Guided	37.5	32.6	34.8

Final outcomes weight paths equally: re-evaluation retains 485 failures and 486 successes (single/two denominators 223/262 and 228/258). The 390 warned concepts deduplicate to 371 paths (166/205); warned-failure rescue uses only their 359 failures. Success retention is 405/486 (83.3%), comprising 215/228 (94.3%) single and 190/258 (73.6%) two-object paths. Among warnings confirmed by step 20, baseline failure is 280/290 (96.6%), rescue 188/280 (67.1%), and guided failure 92/290 (31.7%).

Probe metrics weight prompts equally: persistence is the symptomatic-step fraction from baseline onset through step 40, including the trigger; terminal recovery requires at least three symptom-free steps after confirmation and no recurrence. Persistence drops 47.8 points (prompt-cluster bootstrap 95% interval for guided minus baseline: $[-52.9, -42.7]$). This supports persistence and controllable reduction of object omissions, not monotonic amplification or irreversibility.

G.2 Guidance windows: outcomes and probe states

The separate 85-pair cohort uses 50 prompts: 59 single-/26 two-object pairs and 90 concept trajectories. All windows share model, parameters, baseline labels, and probe records; all 85 failures and 85 successes are verified. Compare baseline with inclusive windows 1–10, 1–20, 10–25, and 1–40. Lines connect categorical settings, not a continuous duration sweep. S/T/all denominators are 59/26/85 paths for rescue/disruption and 41/25/66 for warned-failure rescue. Bold marks each column's best guided setting.

Table 23: **SFG-X window effects on final images in the 85-pair SD3.5M cohort.** Windows specify inclusive sampling steps. S/T denote single-/two-object prompts; All combines both. Values are percentages. Arrows indicate the preferred direction; bold marks the best guided window per column and metric.

Window	S	T	All
Failure rescue ↑			
Base	0.0	0.0	0.0
1–10	81.4	42.3	69.4
1–20	83.1	57.7	75.3
10–25	67.8	42.3	60.0
1–40	**86.4**	**69.2**	**81.2**
Success disruption ↓			
Base	0.0	0.0	0.0
1–10	11.9	26.9	16.5
1–20	**3.4**	26.9	**10.6**
10–25	10.2	**19.2**	12.9
1–40	5.1	26.9	11.8
Warned-failure rescue ↑			
Base	0.0	0.0	0.0
1–10	78.0	44.0	65.2
1–20	80.5	60.0	72.7
10–25	65.9	44.0	57.6
1–40	**85.4**	**72.0**	**80.3**

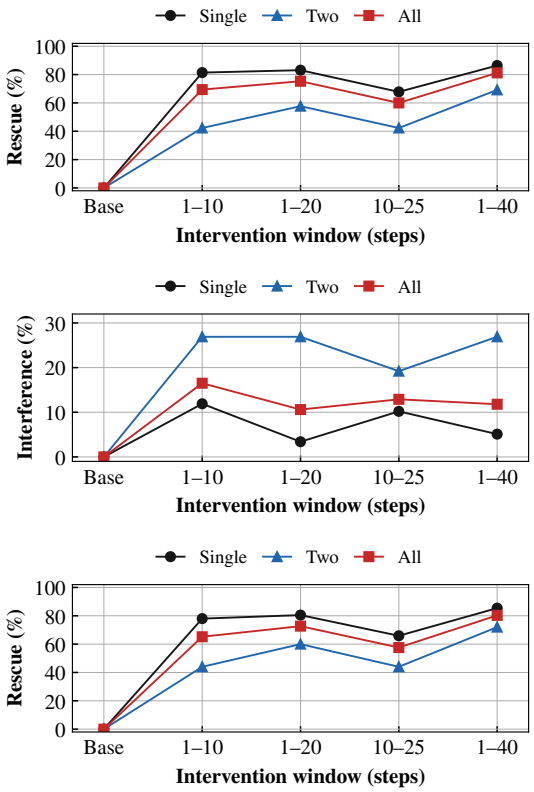


Figure 22: Top to bottom: failure rescue, success disruption, and warned-failure rescue.

Full-window rescue is highest (69/85); 1–20 causes the fewest disruptions (9/85). In the above window order, success retention is 71/85 (83.5%), 76/85 (89.4%), 74/85 (87.1%), 75/85 (88.2%); final failure on the 66 warned paths is 34.8%, 27.3%, 42.4%, 19.7%. Rescue on the 53 paths confirmed by step 20 is 34/53 (64.2%), 37/53 (69.8%), 28/53 (52.8%), 42/53 (79.2%). This ordering has not been tested on the 500-pair cohort.

Probe-state comparisons fix the same 70 baseline-warned concept trajectories (66 deduplicated failed paths) in the 85-pair cohort, with the same criterion and prompt weighting. Probe recovery and final rescue can rank windows differently: two-object terminal non-recovery is lowest for 1–10, whereas image rescue peaks at 1–40.

Table 24: **SFG-X window effects on omission-probe states in the 85-pair SD3.5M cohort.** S/T denote single-/two-object prompts. Percentages use equal prompt weights on fixed baseline-warned concepts. Lower is better; bold marks the best guided window per column and metric.

Window	S	T	All
Symptom persistence ↓			
Base	75.7	59.6	69.1
1–10	26.8	37.9	31.4
1–20	24.0	**25.5**	24.6
10–25	44.0	34.5	40.1
1–40	**21.3**	28.2	**24.1**
No terminal recovery ↓			
Base	72.8	61.1	68.0
1–10	48.1	**29.2**	40.3
1–20	48.4	37.5	43.9
10–25	51.3	35.2	44.7
1–40	**32.7**	45.8	**38.1**

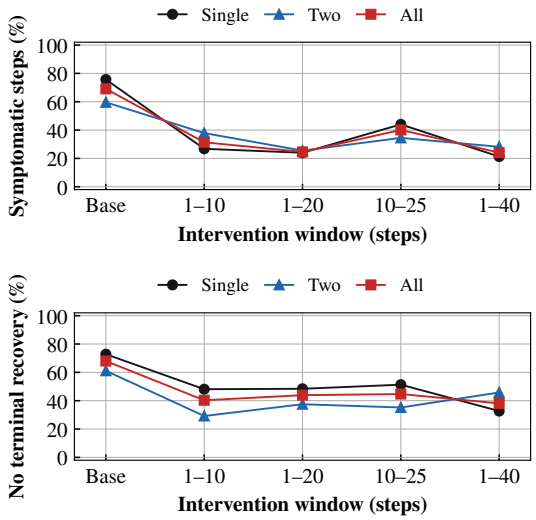


Figure 23: Top to bottom: symptom persistence and no terminal recovery.

G.3 Separate held-out threshold check

Use the temporal median over steps 9–24, calibrate on seeds 42/10042, and test seed 20042 with the same released classifiers. The test set has 22 successes and 22 failures for single-object prompts, and 10 each for two-object prompts. q90/q95 are calibration-success percentiles; false alarms flag successful baseline paths and differ from guidance disruption and paired warning coverage.

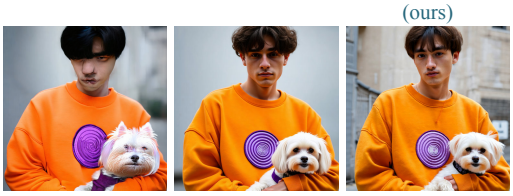
Table 25: **Omission-probe threshold evaluation on a held-out SD3.5M seed.** S/T denote single-/two-object prompts. q90/q95 are the 90th/95th percentiles of probe scores on successful calibration trajectories. Detect is the percentage of failed test trajectories flagged by the probe; False alarm is the percentage of successful test trajectories flagged.

Class	Threshold	Detect	False alarm
S	q90	40.9	18.2
S	q95	27.3	9.1
T	q90	50.0	20.0
T	q95	20.0	10.0

H Qualitative results

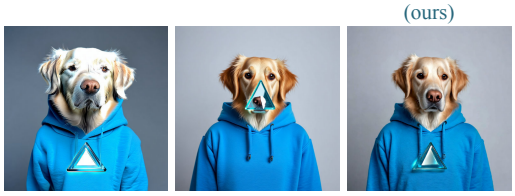
... a young man holding a small white dog, ... a plain orange sweatshirt with one large purple spiral-shaped patch centered on the chest ...

CFG + PAG | CFG + SEG | CFG + SFG (ours)



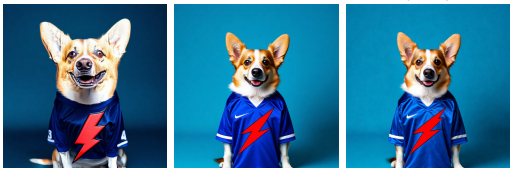
... a golden retriever wearing a plain blue hoodie with one large translucent glass triangle badge centered on the chest ...

CFG + PAG | CFG + SEG | CFG + SFG (ours)



... a corgi wearing a plain blue sports jersey with one large red jagged lightning-bolt shaped logo centered on the jersey ...

CFG + PAG | CFG + SEG | CFG + SFG (ours)



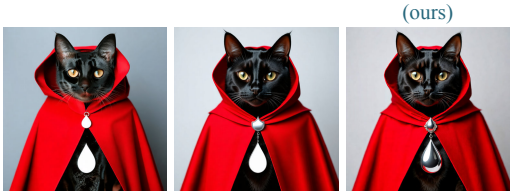
... a person in a plain gray hoodie holding a single orange tabby cat. A simple cyan heart-shaped patch is centered on the chest of the hoodie ...

CFG + PAG | CFG + SEG | CFG + SFG (ours)



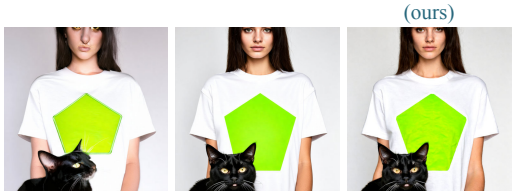
... a black cat wearing a red cape with one large silver teardrop-shaped badge centered on the cape ...

CFG + PAG | CFG + SEG | CFG + SFG (ours)



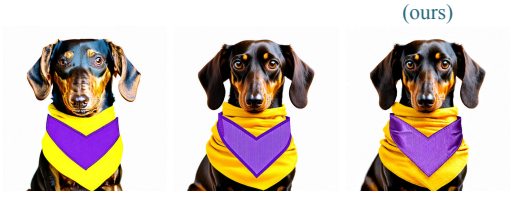
... a woman holding a black cat, ... a plain white t-shirt with one large neon green five-sided pentagonal road-sign patch centered on the chest ...

CFG + PAG | CFG + SEG | CFG + SFG (ours)



... a dachshund wearing a yellow bandana with one large purple ribbon-shaped zigzag patch centered on the bandana ...

CFG + PAG | CFG + SEG | CFG + SFG (ours)



... a fluffy white cat wearing a teal collar with one oversized red hollow ring-shaped charm hanging at the center ...

CFG + PAG | CFG + SEG | CFG + SFG (ours)

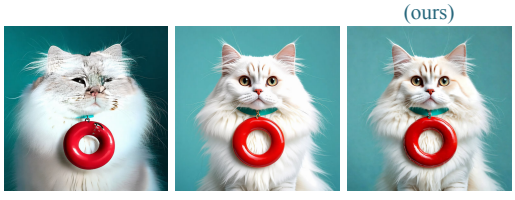


Figure 24: **Qualitative comparisons on SD3.5M.** Each triplet compares CFG+PAG, CFG+SEG, and CFG+SFG (ours) with the same prompt and seed. Prompt excerpts highlight the requested subjects and visual attributes.

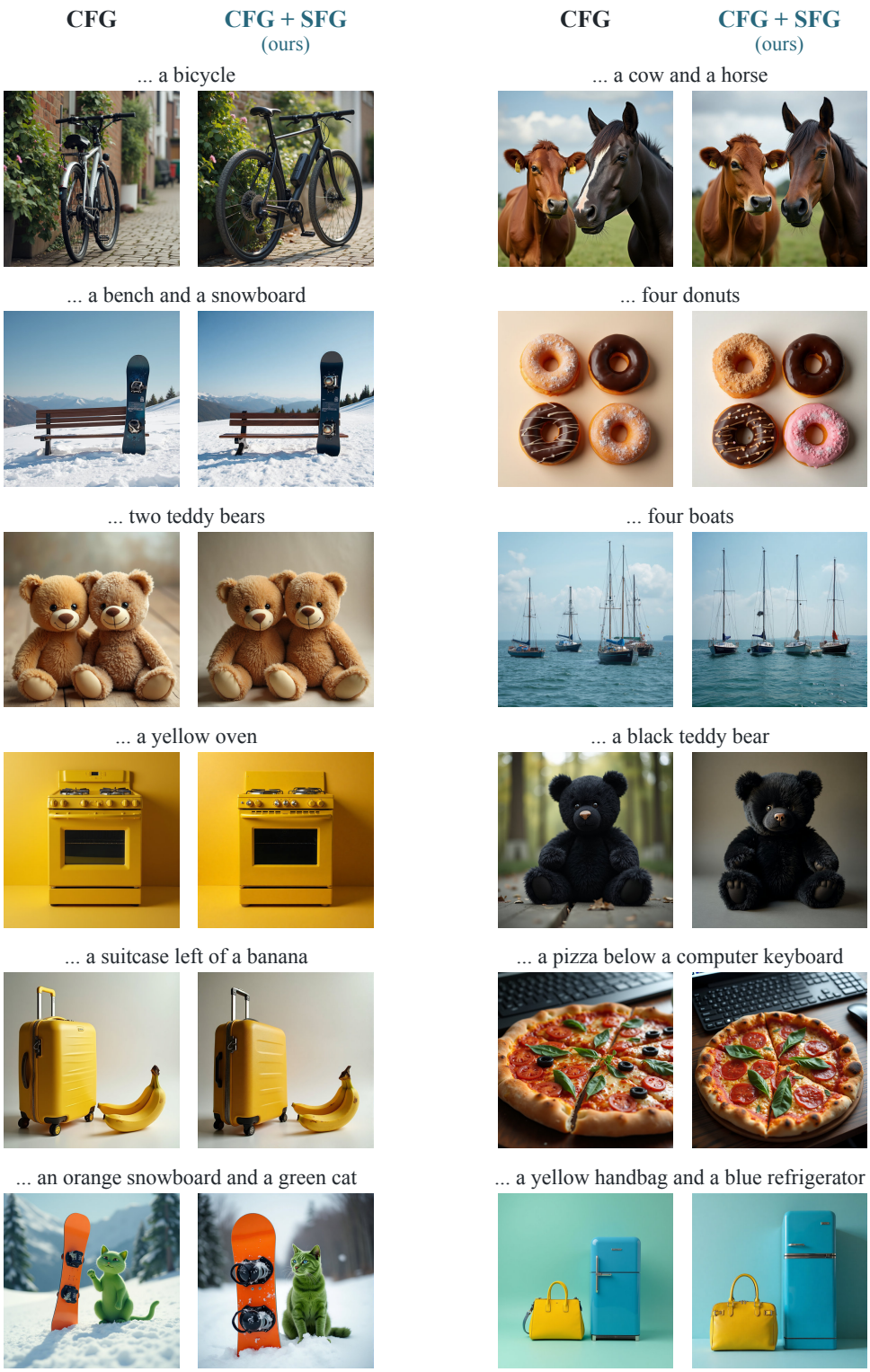


Figure 25: **Compositional comparisons on FLUX.** Each pair compares CFG and CFG+SFG (ours) with the same prompt and seed. Prompt excerpts appear above the pairs.

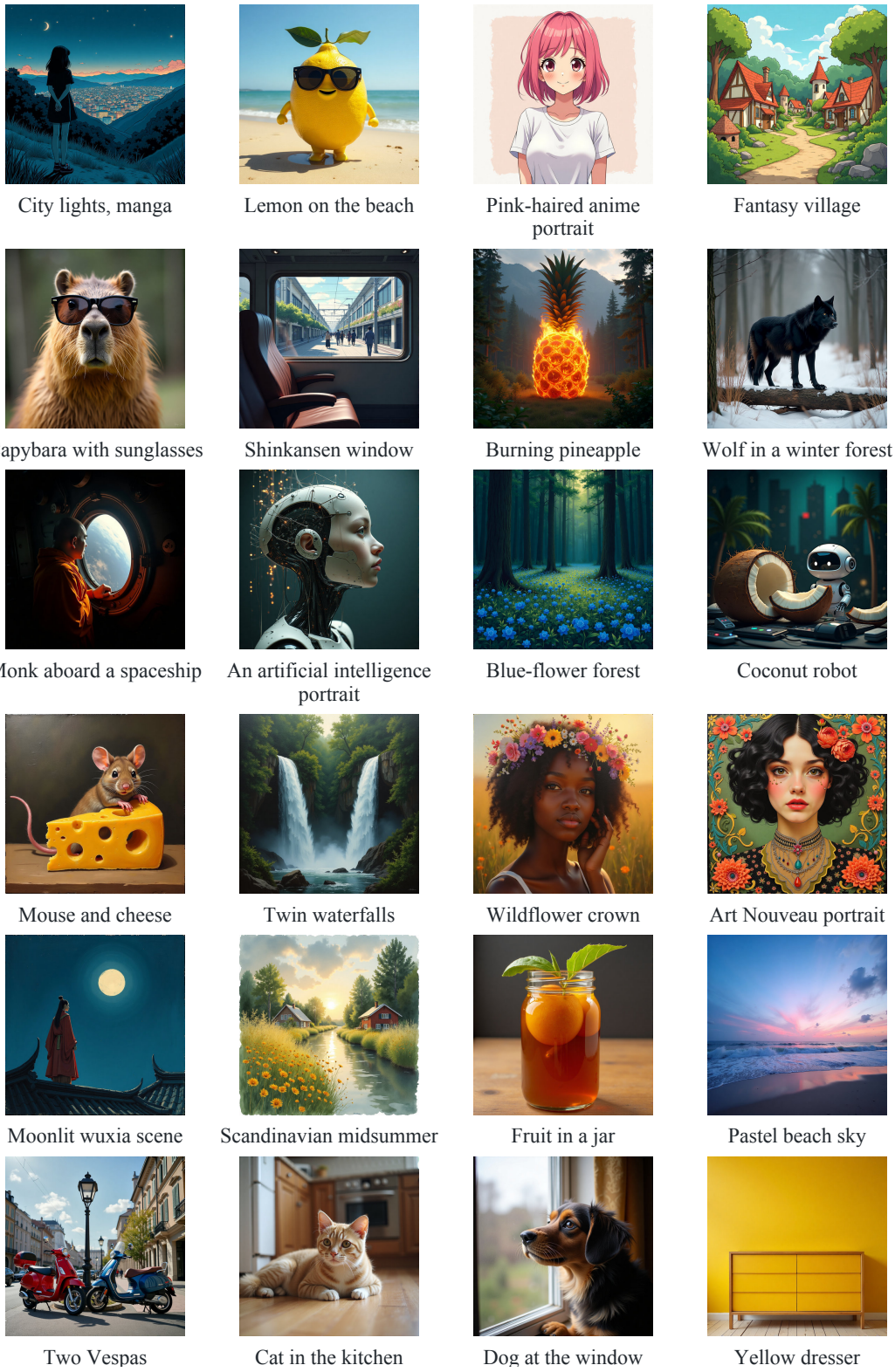


Figure 26: **FLUX with CFG+SFG (ours) across styles and scenes.** Selected outputs from our 800-prompt HPSv2.1 evaluation set; labels describe the images and are not the original prompts.