FLIPPING THE FLOW: PROMOTING DIVERSITY IN FLOW MODELS VIA VELOCITY INVERSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-matching models such as FLUX achieve strong generative quality but often concentrate samples around a few modes, limiting diversity. Existing approaches address this through batch-interaction, optimization-based, or guidance-based methods. Batch-interaction and optimization-based methods improve diversity but require joint sample generation and often additional computation, precluding independent sampling. Guidance-based methods incur little computational overhead but typically rely on text-embedding perturbation or classifier-free guidance (CFG) modulation. In this work, we analyze how scaling the velocity field affects differential entropy, showing that negating a contractive velocity field increases entropy and drives trajectories toward low-density regions. Based on this insight, we discover that a surprisingly simple mechanism—reversing the velocity field during early sampling steps—serves as a remarkably effective tool for mode coverage. Our training-free method requires no model modification, text-embedding access, or CFG variation, enabling application across standard and distilled models, spanning image (FLUX, Z-Image Turbo), video (LTX), and 3D generation (TRELLIS). On GenEval, our approach exceeds the diversity range of noise-initialization methods, rivals batch-interaction methods without joint sampling, and outperforms guidance-based methods under worst-case quality.

1 INTRODUCTION

Flow-matching models Labs et al. (2025); Cai et al. (2025); Esser et al. (2024); Wu et al. (2025); Cao et al. (2025) have emerged as a powerful generative framework, matching or surpassing diffusion methods in sampling efficiency and synthesis quality Lipman et al. (2023); Liu et al. (2022); Esser et al. (2024). They learn a time-dependent velocity field that transports a source distribution to the data distribution, and generate samples by integrating the resulting deterministic ODE. Although linear interpolants yield straighter paths than diffusion Liu et al. (2022), each trajectory is fully determined by its initial noise. Unlike stochastic samplers, which inject noise at every step and can correct accumulated errors Karras et al. (2022), deterministic sampling cannot steer a trajectory away from the mode it is heading toward. As a result, flow-based sampling can concentrate in high-density regions, and flow-matching models have been observed to produce less diverse samples than diffusion models Schusterbauer et al. (2024).

Many standard sampling-time guidance strategies further reinforce this tendency toward mode collapse Malarz et al. (2025); Koulischer et al. (2024); Chang et al. (2024); Chung et al. (2025); Ho & Salimans (2022); Sadat et al. (2024b); Wang et al. (2026b); Saini et al. (2025). Classifier-Free Guidance (CFG) Ho & Salimans (2022) and its adaptive variants Malarz et al. (2025); Yehezkel et al. (2025); Sadat et al. (2024b) improve prompt adherence by amplifying conditional signals but at the cost of diversity, as they pull trajectories toward high-density modes. While Negative Guidance Zhao et al. (2024); AUTOMATIC1111 (2022); Chang et al. (2024); Chen et al. (2025) and Dynamic Negative Guidance (DNG) Koulischer et al. (2024) steer models away from undesired concepts, they primarily act as targeted corrections that mitigate high-frequency artifacts, rather than altering the global transport dynamics needed to explore the manifold or recover lost modes.

One approach to improving diversity uses repulsion between samples: Particle Guidance Corso et al. (2023) modifies the diffusion trajectory to push generated samples apart in pixel space but requires frequent interactions among samples, incurring substantial computational overhead. SPELL Kirch-

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.

hof et al. (2025) also applies pixel-space repulsion but intervenes sparsely rather than continuously perturbing the trajectory. In contrast, NegToMe Singh et al. (2024) applies repulsion between a target and a reference sample in token space, where semantic information is better preserved. However, relying on repulsion from only a single reference sample can be insufficient for effectively covering diverse modes of the data distribution.

Guidance-based methods such as Condition-Annealed Sampling (CADS) Sadat et al. (2024a) and Interval Guidance (IG) Kynkäänniemi et al. (2024) add little overhead but have specific limitations. CADS injects time-dependent noise into text embeddings, which at high diversity produces off-prompt or collapsed outliers that degrade visual quality and artificially inflate diversity metrics (Fig. 27). IG varies the CFG scale during sampling, making it less compatible with CFG-free distilled models. Temporal Score Rescaling (TSR) Xu et al. (2026) rescales the score over time to broaden or sharpen each mode while preserving relative mode weights; in our experiment (Fig. 18), it mainly adds high-frequency detail with little structural change. Opposite Guidance (OG) Jain et al. (2025) uses negative guidance to steer samples away from memorized training images but requires an unconditional branch. Its effect on diversity was not evaluated in the original work, and we find that raising its scale does not reliably increase diversity (Appendix G).

Building on the observation that the velocity field is typically contractive, we investigate an active intervention in the flow dynamics. This contraction continuously reduces differential entropy, concentrating probability mass into a few high-density modes. Leveraging the link between flow divergence and entropy rate, we propose a simple yet effective technique: velocity inversion, which selectively negates the velocity during early high-noise steps. As formalized in Section 4, this induces a positive entropy rate, transitioning the flow from contraction to expansion and pushing the distribution outward to encourage exploration of under-represented regions before the flow reverts to standard refinement. Importantly, by traversing along the learned velocity field, the motion is constrained to the flow lines and the manifold's local tangent space, permitting stronger inversion and greater diversity while delaying drift from the data manifold.

In practice, we scale the velocity with a simple exponential ease-out schedule that starts negative and approaches one by the end of sampling. Unlike prior methods, our approach requires only a single network evaluation per step, with no need to store historical samples or modify the architecture. It also avoids optimization, fine-tuning, and external networks such as CLIP Radford et al. (2021) or DINO Caron et al. (2021), making it readily applicable to any flow-based model with negligible overhead. Despite its simplicity, our method achieves a favorable diversity–quality trade-off on GenEval: it is competitive with guidance-based methods under mean quality, outperforms them under worst-case quality, and rivals batch-interaction methods with fewer constraints. It also maintains quality at diversity levels well beyond those of noise-initialization methods, which saturate early.

To summarize, our primary contributions are as follows:

- **Entropy-Based Analysis of Velocity Scaling:** Building on the entropy–divergence identity for continuous flows, we show that inverting a contractive velocity field makes it expansive and, unlike noise injection, empirically steers trajectories into distinct modes.
- **Parametric Exponential Schedule:** We propose a simple ease-out schedule for early exploration and late refinement that requires no training, architectural changes, or CFG access, and adds near-zero overhead.
- **Broad Generalization Across Modalities and Distilled Models:** Our method achieves a favorable diversity–quality trade-off on GenEval with FLUX and increases diversity with comparable or modestly reduced quality on distilled image (Z-Image Turbo), video (LTX-2.3), and 3D (TRELLIS) models.

2 Related Work

Existing approaches for promoting inference-time diversity offer various trade-offs between computational cost, optimization requirements, architectural modifications, and manifold stability:

Guidance-based methods weaken or modify guidance to mitigate mode collapse, often sacrificing prompt alignment. CADS Sadat et al. (2024a) injects time-dependent noise into text embeddings. In our experiment (Fig. 27), this produces anatomical artifacts and semantic drift that inflate diversity

scores, which we attribute to an orthogonal semantic shift that forces the model into a "semantic vacuum." Interval Guidance (IG) Kynkäänniemi et al. (2024) modulates the CFG scale during inference, which bounds its diversity by that of CFG $= 1$ and makes it incompatible with distilled models that use a fixed CFG scale. For TSR Xu et al. (2026), the rescaling factor approaches one at high noise levels and takes full effect only near the end of sampling, where fine details are resolved. However, on images, the method mainly controls high-frequency details rather than increasing diversity (Fig. 18). OG Jain et al. (2025) applies a negative guidance scale, pushing samples away from the conditional prediction, until a transition point, after which standard CFG is applied. The transition point is either fixed in advance (static transition point, STP) or detected per sample from the drop in the magnitude of the conditional guidance term (dynamic transition point, DTP). Yet increasing the negative guidance scale does not reliably increase diversity.

Batch interaction-based methods encourage diversity by modifying sampling trajectories to push samples apart. Techniques such as Particle Guidance (Corso et al., 2023) and Shielded Diffusion (SPELL) (Kirchhof et al., 2025) utilize intrabatch interactions or geometric constraints, while DiverseFlow Morshed & Boddeti (2025) steers the flow via a kernel matrix representing pairwise interactions. Negative Token Merging (NegToMe) (Singh et al., 2024) actively suppresses specific self-attention tokens associated with over-represented features during the forward pass. While effective, these strategies introduce computational and memory overhead, as they typically require maintaining historical sample sets, or relying on auxiliary pretrained models for feature extraction.

Noise initialization strategies include *post-hoc selection* methods, such as Scaling Group Inference (SGI) Parmar et al. (2026), which generates many candidate samples and prunes them to a diverse subset. This can be expensive due to costly pairwise feature comparisons and is limited by the initial pool's diversity. *Optimization-based* methods instead refine the initial noise: Never Too Late Harrington et al. (2026) relies on external pretrained models and memory-intensive backpropagation through the VAE decoder, while Contrastive Noise Optimization (CNO) Kim et al. (2025) avoids off-the-shelf models via a contrastive loss—pulling noise toward positive representations to maintain sample quality while pushing it away from negatives for diversity. CNO is not near-zero cost as it needs a few-step latent optimization before sampling, a large overhead when used with distilled models. A limitation shared by both selection-based and optimization-driven methods is that they require batch sampling, so samples cannot be generated in isolation, breaking inter-sample independence. Since these noise-centric techniques operate orthogonally to the ODE trajectory, they remain complementary to ours and can be combined to further boost diversity.

Prompt augmentation strategies improve generation quality by refining textual input rather than altering the underlying flow dynamics. Several approaches leverage Large Language Models (LLMs) in conjunction with generative models to provide more detailed or stylistically diverse descriptions (Palaev et al., 2025; Ashutosh et al., 2025; Ma et al., 2025; Wang et al., 2026a; Kou et al., 2026; Hao et al., 2023; Zhong et al., 2023; Fu et al., 2024; Kazimi et al., 2026; Dorfman et al., 2026). Our approach is orthogonal and complementary to these semantic interventions. While prompt engineering enriches the conditioning signal, many structural and geometric variations within the latent space are difficult to articulate through text alone.

3 Preliminaries

3.1 Flow Matching

Flow matching formulates generative modeling as learning a time-dependent velocity field that transports a simple distribution to a data distribution. Let $\mathbf{x}_t \in \mathbb{R}^d$ denote the state at time $t \in [0, 1]$. The dynamics are defined by the ordinary differential equation (ODE):

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{v}_\theta(\mathbf{x}_t, t), \tag{1}$$

where $\mathbf{v}_\theta(\cdot, t)$ is a parameterized velocity field, trained to match a target velocity field $\mathbf{v}^*(\mathbf{x}_t, t)$ via

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{x}_t}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t) - \mathbf{v}^*(\mathbf{x}_t, t)\|_2^2\right]. \tag{2}$$

Sampling is performed by integrating the learned velocity field from an initial noise sample $\mathbf{x}_0 \sim p_0$:

$$\mathbf{x}_1 = \mathbf{x}_0 + \int_0^1 \mathbf{v}_\theta(\mathbf{x}_t, t)\, dt. \tag{3}$$

4 Methodology: Adaptive Repulsive Flow Matching

Standard flow matching process is typically contractive, decreasing differential entropy $H(p_t)$ to concentrate mass into learned data modes. While this ensures high-fidelity samples, it tends to suppress variation and neglect low-density regions of the manifold. To enhance diversity, we propose Adaptive Repulsive Flow, which selectively inverts the velocity during inference.

4.1 Entropy Dynamics of Flow Matching

We first establish the relationship between the velocity field and the diversity of generated samples, characterized by differential entropy.

Definition 1 (Differential Entropy)**.** For a probability density $p_t(\mathbf{x})$ at time $t \in [0, 1]$, the differential entropy is defined as:

$$H(p_t) = -\int p_t(\mathbf{x}) \ln p_t(\mathbf{x})\, d\mathbf{x}. \tag{4}$$

Proposition 1 (Entropy Rate of Change)**.** *Given a probability density p_t evolving according to the continuity equation $\frac{\partial p_t}{\partial t} = -\nabla \cdot (p_t \mathbf{v})$, and assuming sufficient smoothness and decay at infinity such that boundary terms vanish, the rate of change of the differential entropy is:*

$$\frac{dH(p_t)}{dt} = \mathbb{E}_{\mathbf{x}\sim p_t}[\nabla \cdot \mathbf{v}(\mathbf{x})]. \tag{5}$$

Proof. Taking the time derivative of entropy:

$$\frac{dH(p_t)}{dt} = -\int \frac{\partial}{\partial t}[p_t \ln p_t]\, d\mathbf{x} = -\int (\ln p_t + 1)\frac{\partial p_t}{\partial t}\, d\mathbf{x}. \tag{6}$$

Substituting $\frac{\partial p_t}{\partial t} = -\nabla \cdot (p_t \mathbf{v})$ and applying integration by parts (boundary terms vanish):

$$\frac{dH(p_t)}{dt} = -\int p_t \mathbf{v} \cdot \nabla \ln p_t\, d\mathbf{x} = -\int \mathbf{v} \cdot \nabla p_t\, d\mathbf{x}. \tag{7}$$

Applying integration by parts again to shift the gradient to the velocity field:

$$\frac{dH(p_t)}{dt} = \int p_t(\nabla \cdot \mathbf{v}) d\mathbf{x} = \mathbb{E}_{\mathbf{x}\sim p_t}[\nabla \cdot \mathbf{v}(\mathbf{x})]. \tag{8}$$

□

Note that Eq. 5 is the entropy-level analogue of the pathwise log-density identity used in instantaneous change of variables for continuous normalizing flows Chen et al. (2018).

4.2 Adaptive Velocity Scaling

In standard inference, the model utilizes $\mathbf{v} = \mathbf{v}_\theta$. Since generative models exhibit negative divergence ($\nabla \cdot \mathbf{v}_\theta < 0$) during mode-seeking, $H(p_t)$ decreases. We introduce a time-varying scaling factor $\alpha(t)$ to balance exploration and exploitation:

$$\mathbf{v}_{\text{new}}(\mathbf{x}, t) = \alpha(t)\mathbf{v}_\theta(\mathbf{x}, t). \tag{9}$$

Under this formulation, the modified entropy dynamics become:

$$\frac{dH(p_t)}{dt} = \alpha(t) \cdot \mathbb{E}_{\mathbf{x}\sim p_t}[\nabla \cdot \mathbf{v}_\theta(\mathbf{x})]. \tag{10}$$

Crucially, we allow $\alpha(t)$ to be negative during the early stages of generation. When $\alpha(t) < 0$, the entropy rate $\frac{dH}{dt}$ becomes positive, transforming the naturally contractive model velocity into a repulsive field that actively pushes samples away from premature mode collapse.

4.3 Ease-Out Scheduling for Diversity-Fidelity Balance

To achieve this behavior, one simple way is to parameterize $\alpha(t)$ with an exponential function that prioritizes high-noise steps:

$$\alpha(t) = 1 - \eta \cdot \exp\left(-\frac{t}{\sigma}\right), \tag{11}$$

where $\eta > 1$ is a diversity-strength hyperparameter. This schedule ensures a two-phase process:

1. Exploration Phase ($t \ll 1$): Since $\eta > 1$, $\alpha(0) = 1 - \eta$ is negative. This yields $\frac{dH(p_t)}{dt} > 0$, inducing a repulsive flow that expands the probability mass and prevents premature mode collapse.

2. Refinement Phase ($t \approx 1$): As $\alpha(t) \to 1$, the flow becomes contractive and refines details along the diverse trajectories set during exploration.

We provide ablation studies of various functional forms (Appendix E), visual intuition of flow dynamics on a 2D toy dataset (Appendix F), and a theoretical stability analysis (Appendix C).

5 Experiments

Implementation details. We tune the hyperparameter $\sigma \in [0.1, 0.3]$, leaving $\eta \in \{1.5, 2.0, 2.5\}$ as a diversity knob for our main experiments. Details on hyperparameter tuning are provided in Appendix I. Based on the validation, we fix $\sigma = 0.25$. Configuration details for each baseline are provided in Appendix A.

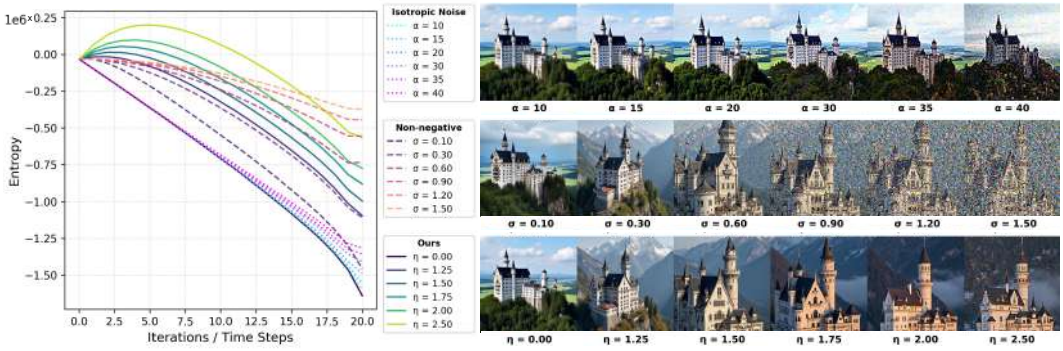


Figure 1: Entropy trajectories (20 total steps) and generated samples for three strategies: adding isotropic noise, non-negative velocity, and our velocity inversion. Our method successfully transitions the flow from contractive to expansive during early high-noise steps. While alternative strategies can similarly produce high final entropy, only our velocity inversion can escape the high-density basin of attraction, reaching a visibly different mode. Prompt: *A photo of Neuschwanstein Castle.*

5.1 Entropy Analysis

We track the evolution of $H(p_t)$ during generation. As shown in Section 4, the entropy rate equals the expected divergence of the velocity field. For high-dimensional models such as FLUX, we estimate the divergence with Hutchinson's trace estimator Hutchinson (1989), following Cheng et al. (2024); Avdeyev et al. (2023). Specifically, we approximate the divergence as $\nabla \cdot \mathbf{v} \approx \epsilon^\top \nabla_x \mathbf{v} \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ is a random noise vector. By averaging this estimator over multiple trajectories sampled from the probability path, we obtain a numerical estimate of the instantaneous entropy rate. Figure 1 shows that our schedule switches the flow from contractive ($\nabla \cdot \mathbf{v} < 0$) to expansive ($\nabla \cdot \mathbf{v} > 0$) during the early stages of generation.

A natural question arises: if the goal is simply to increase the final entropy at $t = 1$, could this be achieved by other means, and would it yield similar diversity? We examine two alternatives: slowing down the trajectory by scaling the velocity with a constant $\in (0, 1]$ (labeled *Non-negative* in Fig. 1) and adding isotropic noise during inference. While these strategies can be tuned to match or exceed

the high final entropy of our method, they fail to produce a corresponding increase in diversity. *Non-negative*'s increased entropy merely reflects a state of incomplete denoising and signal corruption (evident starting at $\sigma = 0.6$), as the sample remains trapped within the original mode's basin of attraction. Similarly, isotropic noise cannot steer the trajectory toward new semantic modes. (See qualitative analysis in Appendix K.) In contrast, at equivalent entropy values, our method reaches visibly different images from the same initial noise without corruption until $\eta = 2.5$.

5.2 Diversity-Fidelity Trade-off

We evaluate on FLUX.1-Dev Black Forest Labs (2024) and the 553 test prompts of the GenEval dataset Ghosh et al. (2023). We generate 10 images per prompt with the same seeds for all methods.

Baselines. We include 10 baselines, grouped into three categories:

- **Guidance-based** (6 baselines): Base (FLUX), CADS, IG, IG Marginal, OG+STP, and OG+DTP. IG Marginal is our extension of IG that uses marginal velocity to further enhance diversity, and OG+STP and OG+DTP are two variants of OG that use a static or dynamic transition point, respectively. We exclude TSR Xu et al. (2026) because it mainly adds high-frequency detail rather than structural variation when applied to images (Fig. 18).
- **Noise-initialization** (2 baselines): SGI and CNO. We exclude Never-Too-Late because it requires a one-step model.
- **Batch-interaction** (2 baselines): SPELL and NegToMe. We exclude DiverseFlow because it relies on an off-the-shelf model and exclude PG because SPELL shares its mechanism but performs better Kirchhof et al. (2025).

Detailed baseline configurations are described in Appendix A.

Metrics. Because diversity trades off against prompt adherence, we always report diversity against quality. We use four diversity metrics: i) DINO pairwise cosine distance Caron et al. (2021), ii) DreamSim Fu et al. (2023), iii) Vendi Score Friedman & Dieng (2022), and iv) Mean Pairwise Similarity (MSS) Pizzi et al. (2022); and three image quality metrics: i) HPSv2 Wu et al. (2023) for sample quality, and ii) CLIP Radford et al. (2021) and iii) PickScore Kirstain et al. (2023) for prompt alignment. Plotting these metrics yields 4x3 trade-off plots.

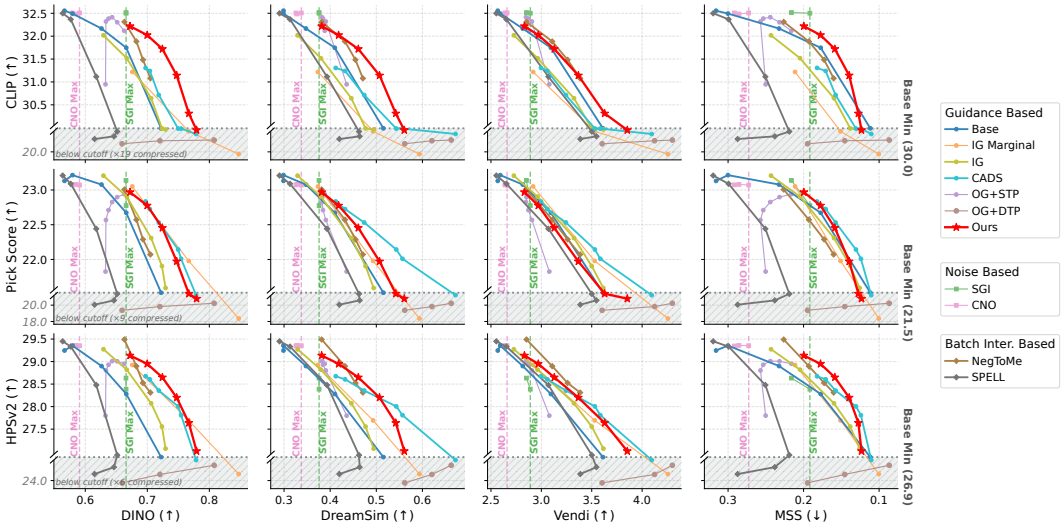
Crucially, we find that the standard practice of reporting the *mean* quality score per prompt in trade-off plots can be exploited to reach the Pareto front. Having a few samples with off-prompt outliers substantially increases pairwise diversity but only marginally lowers the mean quality score (Appendix B). We therefore also report the *minimum* quality score per prompt, which is high only if every sample is faithful to the prompt. Reference-based metrics like precision, recall, and density are reported only where the training data is known, such as unconditional generation (Appendix H).

Quality cut-offs. Since the parameters of any method can be swept to high diversity at the cost of severely degraded samples, we compare methods only within an acceptable quality range. For each quality metric, the cut-off is the score of the base model at CFG $= 1$, which samples from the model's unguided conditional distribution, and thus marks the quality floor a user already accepts when lowering the guidance scale to increase diversity. These cut-offs are CLIP 30.0, PickScore 21.5, and HPSv2 26.9 under mean quality, and 25.3, 20.0, and 25.1 under minimum quality.

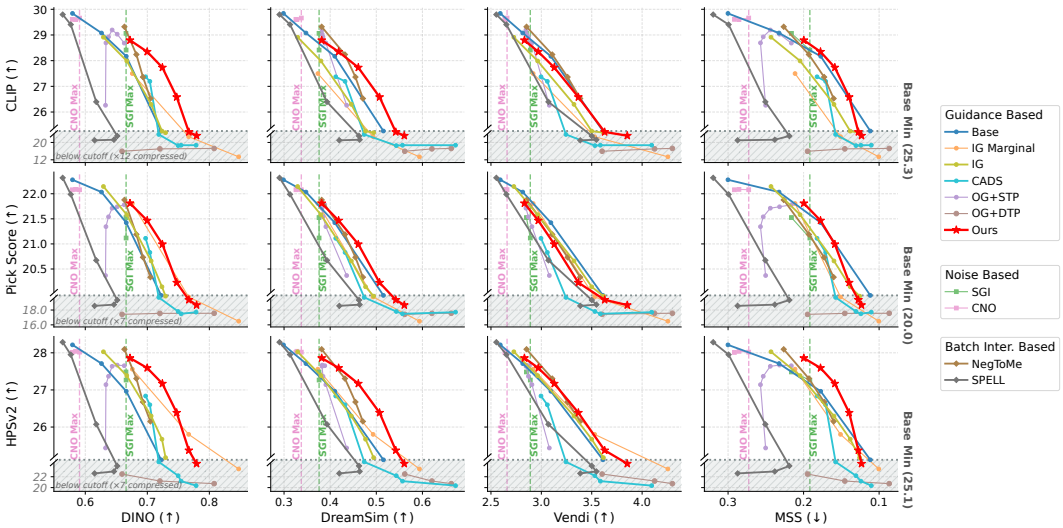
Results. Fig. 2 shows the diversity–quality trade-off. Several baselines have a limited diversity range. Noise-initialization methods saturate below DINO 0.67 (Fig. 20), and NegToMe tops out at 0.71. The base model and IG are capped by the diversity of CFG $= 1$ (Fig. 19), and OG+STP does not scale with s (Appendix G). In contrast, our method reaches DINO 0.747 and DreamSim 0.507 above all cut-offs. At matched diversity, it forms the Pareto front on DINO and DreamSim among guidance-based methods above the CLIP cut-off, under both mean and minimum quality. It also outperforms SPELL on all metric pairs and NegToMe on DINO and MSS (Figs. 17 and 26). On Vendi and MSS, it is competitive rather than dominant.

Comparing mean and minimum quality separates the methods further. Under mean PickScore and HPSv2, CADS is competitive at high diversity, but its advantage disappears under minimum quality (Fig. 27). OG+DTP reaches high diversity only below the cut-offs (Fig. 21). Part of the diversity gained by perturbing the conditioning signal therefore comes from a few off-prompt samples, which *mean quality* largely hides. Our method retains its advantage under minimum quality, which sug-

gests that velocity inversion better avoids off-manifold results. Its gains are also most consistent on DINO and DreamSim, which capture semantic and structural variation. Finally, as our method acts on the sampling trajectory, it is complementary to noise-initialization methods and further improves the diversity of SGI (Fig. 28).



(a) **Mean**: Mean quality (rows) vs. diversity (columns).



(b) **Min**: Worst-case per-prompt quality (rows) vs. diversity (columns).

Figure 2: **Quality–diversity trade-off** on GenEval ($N = 553$ prompts). Upper-right is better.

5.3 Other modalities and distilled model

In this section, we evaluate the generalizability of our proposed method on distilled image models, and other modalities, namely video and 3D generation. We report results using the parameters that achieve high diversity while maintaining text-alignment comparable to Base (vanilla sampling); complete details on our parameter search strategy are provided in Appendix N. We report the standard error (SE) and indicate statistically significant gaps using a paired t-test at $p < 0.05$. For fair comparison, we use the same set of seeds across all methods and prompts.

Metrics. For images, we use the same metrics as in Sec. 5.2. For videos, following Kazimi et al. (2026), we measure video diversity with TCE, TIE, and the Vendi Score, and video quality with VideoScore (He et al., 2024) and VBench (Huang et al., 2024). For 3D generation, we render a

rotation video of each asset, apply the video diversity metrics, and measure text alignment with VideoScore TVA using the prompt "A rotation video of {prompt}."

Results on distilled generation. On Z-Image Turbo (Table 1a), our method significantly increases all four diversity metrics over Base, with minimal change, or even a slight increase, in image quality. Note that SPELL's r of 110, 125, and 150 produces virtually identical outputs. Compared with SPELL and CNO, our method achieves the best scores in all diversity metrics (tie on DreamSim). Both baselines also require joint sampling of the batch, and CNO adds 3 to 5 epochs of noise optimization before generation, a substantial overhead relative to Z-Image Turbo's 8-step sampling. Qualitative results for Z-Image Turbo and preliminary results for FLUX.1-Schnell are shown in Fig. 24 and Fig. 25, respectively.

Results on video generation. On LTX-2.3 Distilled (Table 1b), the results on VBench metrics are mixed, with ours outperforming Base on Subject Consistency, Image Quality, and Temporal Flickering, but underperforming on Background Consistency, Aesthetic Quality, and Dynamic Degree. However, our method outperforms Base on all three diversity metrics and on every VideoScore metric except TVA, where the two tie.

Results on 3D generation. Despite lower prompt alignment score (TVA) compared to base (Table 1c), our method achieve higher diversity across the TCE, TIE, and Vendi metrics. To validate that the generated 3D assets maintain high visual quality. We demonstrate this by providing qualitative results for a subset of all evaluated prompts in the supplementary material.

Table 1: **Quantitative comparison across diverse generative modalities.** * indicates that the gap between each baseline and ours is statistically significant in a paired t-test at $p < 0.05$.

(a) Distilled Model Generation: Z-Image Turbo on GenEval (SE diversity / quality: N = 50 prompts)

Method	Quality			Diversity			
	CLIP ↑	PickScore ↑	HPSv2 ↑	DINO ↑	DreamSim ↑	Vendi ↑	MSS ↓
Base	32.77 ± 0.44	23.16 ± 0.13	29.65 ± 0.14	$0.43 \pm 0.01^*$	$0.19 \pm 0.01^*$	$2.05 \pm 0.06^*$	$0.47 \pm 0.02^*$
SPELL (Batch Inter.)							
$r = 110$	32.80 ± 0.45	23.14 ± 0.13	29.58 ± 0.14	$0.45 \pm 0.01^*$	$0.22 \pm 0.01^*$	$2.14 \pm 0.07^*$	$0.42 \pm 0.02^*$
$r = 125$	32.80 ± 0.45	23.14 ± 0.13	29.58 ± 0.14	$0.45 \pm 0.01^*$	$0.22 \pm 0.01^*$	$2.14 \pm 0.07^*$	$0.42 \pm 0.02^*$
$r = 150$	32.80 ± 0.45	23.14 ± 0.13	29.58 ± 0.14	$0.45 \pm 0.01^*$	$0.22 \pm 0.01^*$	$2.14 \pm 0.07^*$	$0.42 \pm 0.02^*$
$r = 175$	32.81 ± 0.45	23.12 ± 0.13	29.57 ± 0.14	$0.45 \pm 0.01^*$	$0.22 \pm 0.01^*$	$2.15 \pm 0.07^*$	$0.41 \pm 0.02^*$
$r = 200$	32.72 ± 0.45	23.06 ± 0.13	$29.49 \pm 0.15^*$	$0.47 \pm 0.01^*$	0.24 ± 0.01	2.21 ± 0.07	$0.39 \pm 0.02^*$
CNO (Noise Init.)							
3 Epochs, $\gamma = 1$	32.99 ± 0.43	$\mathbf{23.20 \pm 0.13}$	29.59 ± 0.14	$0.48 \pm 0.02^*$	0.23 ± 0.01	2.19 ± 0.07	$0.39 \pm 0.02^*$
5 Epochs, $\gamma = 1$	$\mathbf{33.03 \pm 0.43}$	23.14 ± 0.12	29.59 ± 0.14	$0.48 \pm 0.02^*$	$\mathbf{0.25 \pm 0.01}$	2.23 ± 0.07	0.38 ± 0.02
Ours (Guidance)	32.82 ± 0.48	23.11 ± 0.13	$\mathbf{29.66 \pm 0.13}$	$\mathbf{0.56 \pm 0.01}$	$\mathbf{0.25 \pm 0.01}$	$\mathbf{2.25 \pm 0.07}$	$\mathbf{0.35 \pm 0.01}$

(b) Video Generation: LTX 2.3 Distilled On Vbench (SE diversity: N = 60 prompts, quality: N = 600 videos)

Method	Diversity			VideoScore					VBench					
	TCE	TIE	Vendi	VQ	TC	DD	TVA	FC	SC	BC	AQ	IQ	TF	DD
Base	23.95^* ± 0.30	41.78^* ± 0.37	4.01^* ± 0.05	2.80^* ± 0.02	2.77^* ± 0.02	2.69^* ± 0.03	2.56 ± 0.03	2.69^* ± 0.02	0.955 ± 0.0015	$\mathbf{0.955}^*$ ± 0.0010	$\mathbf{0.568}^*$ ± 0.0030	70.77^* ± 0.30	0.9859^* ± 0.0003	$\mathbf{0.670}$ ± 0.0192
Ours	$\mathbf{24.92}$ ± 0.20	$\mathbf{42.51}$ ± 0.26	$\mathbf{4.19}$ ± 0.05	$\mathbf{2.87}$ ± 0.02	$\mathbf{2.82}$ ± 0.02	$\mathbf{2.78}$ ± 0.03	2.56 ± 0.03	$\mathbf{2.76}$ ± 0.02	$\mathbf{0.957}$ ± 0.0014	0.951 ± 0.0011	0.551 ± 0.0029	$\mathbf{72.66}$ ± 0.25	$\mathbf{0.9878}$ ± 0.0002	0.653 ± 0.0194

(c) 3D Generation: TRELLIS on synthetic prompts (SE diversity: N = 100 prompts)

Method	Diversity			VideoScore
	TCE	TIE	Vendi	TVA
Base	$21.27 \pm 0.32^*$	$41.50 \pm 0.27^*$	$5.69 \pm 0.12^*$	$\mathbf{2.35 \pm 0.02}^*$
Ours	$\mathbf{21.75 \pm 0.24}$	$\mathbf{42.15 \pm 0.22}$	$\mathbf{5.95 \pm 0.10}$	2.30 ± 0.02

5.4 Ablation Study: η and σ

We analyze the effects of varying η and σ in Equation 11 across T2I-CompBench Huang et al. (2023), with further details provided in Appendix I. As shown in Fig. 3, we observe that increasing

η consistently enhances sample diversity, manifesting as variations in viewpoint, style, and environment. Conversely, increasing σ induces more pronounced high-frequency patterns in the generated images. The impact of σ on 3D and video generation is visualized in Fig. 10.



Figure 3: Comparison of the effects of η and σ. Increasing η promotes greater diversity, while increasing σ introduces stronger high-frequency patterns.

6 Limitations and Discussion

While our method significantly enhances global semantic diversity by intervening in high-noise regimes, fine-grained attributes and high-frequency details often resolve in later time steps where our intervention is less pronounced. Furthermore, the degree of mode collapse may be inherently prompt-dependent. Following recent advances in adaptive guidance scaling for CFG-based methods Koulischer et al. (2024); Chung et al. (2025), our approach could be extended to incorporate sample-adaptive scaling in future work. Nevertheless, even with a single global schedule, our method achieves a favorable diversity-fidelity trade-off. We also observe that overly aggressive velocity scaling can degrade image quality when applied over extended sampling trajectories. However, as a test-time method with near-zero overhead, it can be efficiently combined with post-hoc selection mechanisms like Scaling Group Inference Parmar et al. (2026), which filter out degraded samples while preserving high diversity beyond the candidate noise distribution (Fig. 28).

7 Conclusion

We introduce a simple yet effective method for improving diversity in flow-matching models by modulating the velocity field with a time-dependent scaling schedule. Motivated by the link between velocity scaling and entropy dynamics, our approach applies early-stage velocity inversion to expand exploration in high-entropy regions, then transitions gradually to restore structural fidelity and text alignment. Unlike prior methods that rely on optimization, reward models, or batch-dependent interactions, our method operates entirely at test time, requires no architectural changes or additional supervision, and adds near-zero compute and memory overhead.

Broader Impacts. Greater diversity in generative models can benefit creative workflows by producing a broader range of plausible visual concepts from a single prompt (Fig. 22) and may help reduce representational bias by reducing models' tendency to default to dominant or stereotypical depictions of gender, race, culture, or social roles. However, it may also give adversarial users access to more harmful, biased, or policy-violating outputs and reduce predictability where strict instruction fidelity is required. This calls for careful evaluation of the trade-offs between variety and steerability.

REFERENCES

Kumar Ashutosh, Yossi Gandelsman, Xinlei Chen, Ishan Misra, and Rohit Girdhar. Llms can see and hear without any training. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Kendall E Atkinson. *An introduction to numerical analysis*. John wiley & sons, 2008.

AUTOMATIC1111. Stable diffusion webui: Negative prompt, 2022. URL https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt. GitHub repository wiki.

Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet diffusion score model for biological sequence generation. In *International Conference on Machine Learning*, pp. 1276–1301. PMLR, 2023.

Black Forest Labs. Flux.1. https://github.com/black-forest-labs/flux, 2024.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Lawrence Cayton et al. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

Jinho Chang, Hyungjin Chung, and Jong Chul Ye. Contrastive cfg: Improving cfg in diffusion models by contrasting positive and negative concepts. *arXiv preprint arXiv:2411.17077*, 2024.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized attention guidance: Universal negative guidance for diffusion models. *arXiv preprint arXiv:2505.21179*, 2025.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds. *Advances in Neural Information Processing Systems*, 37:54787–54819, 2024.

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.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102*, 2023.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023.

Sara Dorfman, Maya Vishnevsky, Omer Dahary, Or Patashnik, and Daniel Cohen-Or. Semantic browsing: Controllable diversity for image generation. In *European Conference on Computer Vision*, pp. 412–430. Springer, 2026.

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

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Yuchen Fu, Zhiwei Jiang, Yuliang Liu, Cong Wang, Zexuan Deng, Zhaoling Chen, and Qing Gu. Ap-adapter: Improving generalization of automatic prompts on unseen text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 37:98320–98346, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.

Anne Harrington, A Koepke, Shyamgopal Karthik, Trevor Darrell, and Alexei A Efros. It's never too late: Noise optimization for collapse recovery in trained diffusion models. 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024.

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

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tahira Kazimi, Connor Dunlop, and Pinar Yanardag. Diverse video generation with determinantal point process-guided policy optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12839–12848, 2026.

Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. *arXiv preprint arXiv:2510.03813*, 2025.

Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and Marco Cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23 of *Applications of Mathematics*. Springer, 1992. doi: 10.1007/978-3-662-12616-5.

Siqi Kou, Jiachun Jin, Zetong Zhou, Ye Ma, Yugang Wang, Quan Chen, Peng Jiang, Xiao Yang, Jun Zhu, Kai Yu, et al. Think-then-generate: Reasoning-aware text-to-image diffusion with llm encoders. *arXiv preprint arXiv:2601.10332*, 2026.

Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic negative guidance of diffusion models. *arXiv preprint arXiv:2410.14398*, 2024.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

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

Mashrur M Morshed and Vishnu Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23303–23312, 2025.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Andrey Palaev, Adil Khan, and Syed M Ahsan Kazmi. Llm-guided instance-level image manipulation with diffusion u-net cross-attention maps. *arXiv preprint arXiv:2501.14046*, 2025.

Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. Scaling group inference for diverse and high-quality generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14512–14522. IEEE, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations (ICLR)*, 2024a.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024b.

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

Johannes Schusterbauer, Ming Gui, Pingchuan Ma, Nick Stracke, Stefan A. Baumann, Vincent Tao Hu, and Björn Ommer. Boosting latent diffusion with flow matching. In *ECCV*, 2024.

Jaskirat Singh, Lindsey Li, Weijia Shi, Ranjay Krishna, Yejin Choi, Pang Wei Koh, Michael F Cohen, Stephen Gould, Liang Zheng, and Luke Zettlemoyer. Negative token merging: Image-based adversarial feature guidance. *arXiv preprint arXiv:2412.01339*, 2024.

Fu-Yun Wang, Han Zhang, Michael Gharbi, Hongsheng Li, and Taesung Park. Promptrl: Prompt matters in rl for flow-based image generation. *arXiv preprint arXiv:2602.01382*, 2026a.

Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. Cfg-ctrl: Control-based classifier-free diffusion guidance. *arXiv preprint arXiv:2603.03281*, 2026b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

Yanbo Xu, Yu Wu, Sungjae Park, Zhizhuo Zhou, and Shubham Tulsiani. Temporal score rescaling for temperature sampling in diffusion and flow models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Shai Yehezkel, Omer Dahary, Andrey Voynov, and Daniel Cohen-Or. Navigating with annealing guidance scale in diffusion space. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Barbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37:12293–12333, 2024.

Shanshan Zhong, Zhongzhan Huang, Weushao Wen, Jinghui Qin, and Liang Lin. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 567–578, 2023.

AI USE STATEMENT

In accordance with the ICLR 2027 policy on the use of large language models (LLMs), we disclose the following. We used ChatGPT (OpenAI) to generate synthetic text prompts for evaluation: the prompt set for 3D generation (Appendix M) and the ComplexPrompt set used for qualitative results (Fig. 22); we reviewed all prompts before use. We also used Claude (Anthropic) as an assistant for: (i) editing and drafting parts of the manuscript, including results descriptions, related-work paragraphs, and figure captions; (ii) improving figure presentation and writing auxiliary code for visualization, all of which we reviewed; Apart from the synthetic evaluation prompts, LLMs were not used to generate research ideas, design the proposed method. All claims and conclusions are the authors' own, and the authors take full responsibility for the content of this paper.

APPENDIX: FLIPPING THE FLOW: PROMOTING DIVERSITY IN FLOW MODELS VIA VELOCITY INVERSION

A BASELINE SETUPS

For FLUX.1-dev, all methods generate 512x512 images with 20 sampling steps.

Base. We sweep the classifier-free guidance (CFG) scale across the values $\{1, 1.5, 2, 3, 3.5\}$.

CADS. Although the original method is designed for diffusion models, following Scaling Group Inference, we explore threshold values $\tau_1 \in \{0.9, 0.8, 0.7, 0.6\}$. However, lower thresholds ($0.8, 0.7$, and 0.6) result in severe image degradation. Consequently, we set $\tau_1 = 0.9$ and $\tau_2 = 0.95$, tuning the noise strength over $s \in [0.01, 0.8]$ with a default CFG scale of 3.5. (Note: While the original paper defines sampling from $t = 1$ to 0 and our framework uses $t = 0$ to 1, all parameters reported here follow the original $1 \to 0$ convention.)

IG (CFG 1/Marginal). This method was originally proposed for diffusion models. We adopt the hyperparameters used in Scaling Group Inference, following its application to FLUX.1-DEV. We apply CFG = 3.5 (default) inside the guidance interval and CFG = 1 outside it, with intervals [0.95, 0.05], $[0.95, 0.05]$, $[0.9, 0.1]$, $[0.8, 0.2]$, $[0.7, 0.3]$, and $[0.6, 0.4]$. However, for the marginal cases, we found that images tend to deviate semantically; therefore, we selected $[0.95, 0.05]$, $[0.9, 0.1]$, and $[0.8, 0.2]$.

SPELL. We adopt the hyperparameters recommended in the original paper, using an overcompensation factor of 1.6 and a radius scale of 200 for SPELL. We further explore alternative radius scales of 275, 250,225, 175, and 150. We use the default CFG scale of 3.5.

NegToMe. We adopt the default parameters from the official implementation. In particular, we use the sampling configuration with merge parameter $\alpha = 0.9$, and additionally evaluate $\alpha \in \{2, 3, 4, 5\}$ to promote greater diversity in the generated images. We use the default CFG scale of 3.5.

SGI.. We use $M = 64$ candidates and $\lambda = 0.9$, within the parameter range recommended by the original implementation. We also tested $M = 128$, but excluded it because it substantially increases computational cost and roughly doubles VRAM usage relative to the other baselines. To verify that SGI's diversity plateaus, we additionally evaluate $\lambda = 20$. We use the default CFG scale of 3.5.

CNO. We perform a hyperparameter sweep for the γ parameter over the set $\{0.85, 1.0, 100\}$. For all configurations, the optimization process is conducted over 5 epochs. We use the default CFG scale of 3.5.

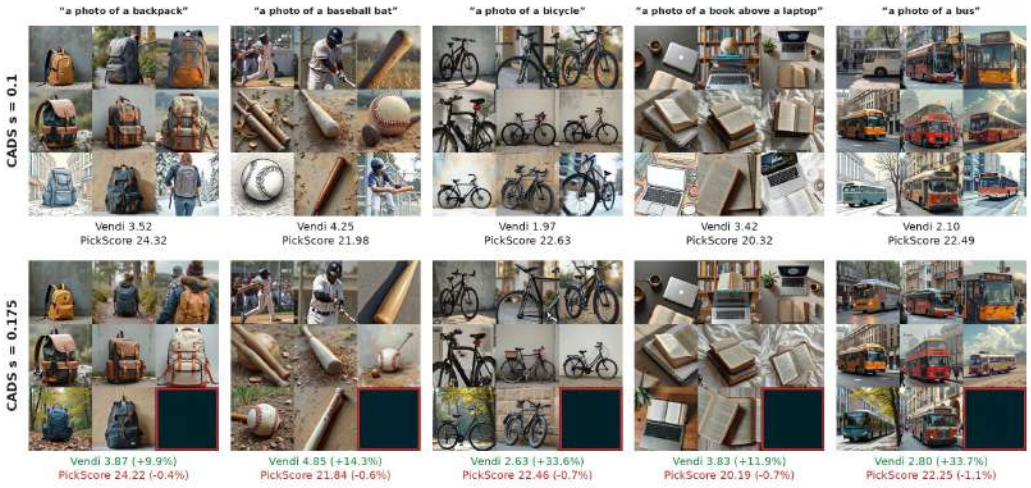


Figure 4: **CADS can inflate diversity scores.** Labels show (Vendi/PickScore) for each set of samples. Increasing the CADS noise scale from $s = 0.1$ to $s = 0.175$ raises Vendi by 20.7% on average over the five prompts, while mean PickScore drops by only 0.7%. The added diversity comes largely from a few off-prompt or collapsed samples, which *mean quality* barely penalizes.

OG with STP and DTP. For the OG + STP baseline, we evaluate the scaling parameter $s \in \{1, 1.25, 1.5, 1.75, 2, 3\}$, whereas for OG + DTP, we test $s \in \{1, 2, 3\}$. Note that $s = 1$ corresponds to the minimum scaling factor evaluated in the original OG paper. Additionally, for STP, we set the flipping timestep to $t = 0.2$ (where $t = 0$ denotes Gaussian noise, and $t = 1$ denotes real image).

Ours. We fix the variance parameter to 0.25 and explore scaling factors $\{1.25, 1.5, 1.75, 2.0, 2.25, 2.5\}$. We use the default CFG scale of 3.5.

B Mean Quality Inflation

Diversity–quality trade-off plots typically summarize each prompt by the mean quality of its samples. This summary can be exploited: if a method produces a few off-prompt or severely degraded samples, these outliers are very different from the remaining samples and therefore increase pairwise diversity substantially, while lowering the mean quality only slightly. For example, with $n = 10$ samples per prompt, replacing m faithful samples with off-prompt ones lowers the mean quality by only a fraction m/n of the per-sample quality gap, whereas the minimum quality drops by the full gap as soon as $m \geq 1$.

Fig. 4 illustrates this effect with CADS on five prompts. Increasing the noise scale from $s = 0.1$ to $s = 0.175$ raises Vendi by 20.7% on average, while the mean PickScore drops by only 0.7%. Judged by mean quality alone, the larger noise scale therefore appears to offer a much better diversity–quality trade-off. The samples, however, show that a substantial part of the added diversity comes from off-prompt or collapsed images, which the mean largely hides. This motivates reporting the minimum quality per prompt in addition to the mean (Fig. 2): the minimum is high only if every sample remains faithful to the prompt, so a method cannot improve its trade-off by adding a few outliers.

C Stability and Semantic Precision Analysis

This section provides a geometric and information-theoretic analysis of why Adaptive Repulsive Flow maintains high fidelity and semantic grounding where alternative strategies fail.

C.1 Geometric Stability via Manifold Tangency

The primary risk of any diversity-promoting intervention is "manifold departure," where the sampling trajectory enters regions of the latent space that cannot be semantically mapped to valid images. To justify the stability, we analyze the behavior of the learned vector field $\mathbf{v}_\theta$ and consider the Manifold Hypothesis Cayton et al. (2005), which posits that natural images lie on a low m-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$. We argue that our method is inherently more stable than stochastic alternatives for two reasons.

Directional Alignment with Learned Priors: In high-noise time steps where samples are far from $\mathcal{M}$, $\mathbf{v}_\theta$ encodes the model's global transport structure. Our repulsive displacement, $\Delta\mathbf{x} = (\alpha(t)\mathbf{v}_\theta)\Delta t$, keeps the trajectory aligned with these learned flow lines even when $\alpha(t) < 0$. In contrast, isotropic noise $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is nearly orthogonal to any fixed direction in high dimensions, and thus is typically nearly orthogonal to $\mathbf{v}_\theta$.

As samples approach $\mathcal{M}$, the learned velocity $\mathbf{v}_\theta(\mathbf{x}, t)$ is expected to align with the local tangent space $T_\mathbf{x}\mathcal{M}$, so the scaled update remains within $T_\mathbf{x}\mathcal{M}$ for any $\alpha(t) \in \mathbb{R}$ by construction. In contrast, isotropic noise has only $m/D \ll 1$ fraction of its energy to $T_\mathbf{x}\mathcal{M}$, with the majority lying in normal directions. As a result, consistent with the misalignment in high-noise regimes, such perturbation introduces dominant components orthogonal to $T_\mathbf{x}\mathcal{M}$, which can drive trajectories off-manifold.

Geometric Error Scaling: While discrete steps following a negative velocity or tangent space may drift off a curved manifold, the geometric deviation for our velocity scaling is $\mathcal{O}(\Delta t^2)$—matching the local truncation order of the ODE sampler itself Atkinson (2008). In contrast, the drift from isotropic noise is significantly larger, with an expected magnitude of $\mathcal{O}(\sqrt{\Delta t})$ Kloeden & Platen (1992). Under mild smoothness assumptions on the flow (e.g., Lipschitz continuity), our $\mathcal{O}(\Delta t^2)$ perturbations do not amplify and can be mitigated by subsequent steps, keeping trajectories near $\mathcal{M}$.

C.2 Semantic Precision: Signal Corruption vs. Velocity Inversion

Beyond geometric stability, we must ensure that increasing diversity does not lead to semantic drift (i.e., a loss of prompt alignment). We analyze our method against two common alternatives: Marginal Attraction Kynkäänniemi et al. (2024) (pulling toward $p(\mathbf{x})$) and Stochastic Condition Perturbation (pulling toward $p(\mathbf{x}|y + \epsilon)$, e.g., as in CADS Sadat et al. (2024a)).

Justification: Consider the score decomposition: $\nabla_\mathbf{x} \log p(\mathbf{x}|y) = \nabla_\mathbf{x} \log p(\mathbf{x}) + \nabla_\mathbf{x} \log p(y|\mathbf{x})$.

1. Marginal Attraction: Sampling via $\nabla_\mathbf{x} \log p(\mathbf{x})$ (unconditional flow) reduces mutual information $I(\mathbf{x}; y)$, effectively "erasing" the prompt and steering toward global data averages.

2. Stochastic Condition Perturbation:

Perturbing the embedding $y' = y + \epsilon$ introduces variations unlikely to align with the model's semantic directions, shifting the sampling distribution from $p(\mathbf{x}|y)$ to $p(\mathbf{x}|y')$. As noise grows to increase diversity, the model is pushed into latent regions increasingly misaligned with the prompt-conditioned manifold $\mathcal{M}_y \subset \mathcal{M}$, leading to semantic drift, anatomical artifacts, and signal corruption.

3. Velocity Inversion (Ours): By setting $\alpha(t) < 0$, the flow follows $-\nabla_\mathbf{x} \log p(\mathbf{x}|y)$, repelling samples from high-density regions of the conditional distribution. This does not alter the conditioning variable or induced manifold $\mathcal{M}_y$, but only the trajectory within the fixed score field of $p(\mathbf{x}|y)$. As a result, it explores low-density tails of $p(\mathbf{x}|y)$ and promotes higher-entropy states within the same conditional distribution.

D Qualitative Analysis of Marginal Velocity Attraction and Negative Velocity

We qualitatively compare marginal velocity attraction against our velocity negation using z-image and FLUX-dev (Figure 5). Marginal guidance leads to inconsistent text adherence across different noise initializations, as its success depends on whether the unconditional noise coincidentally aligns

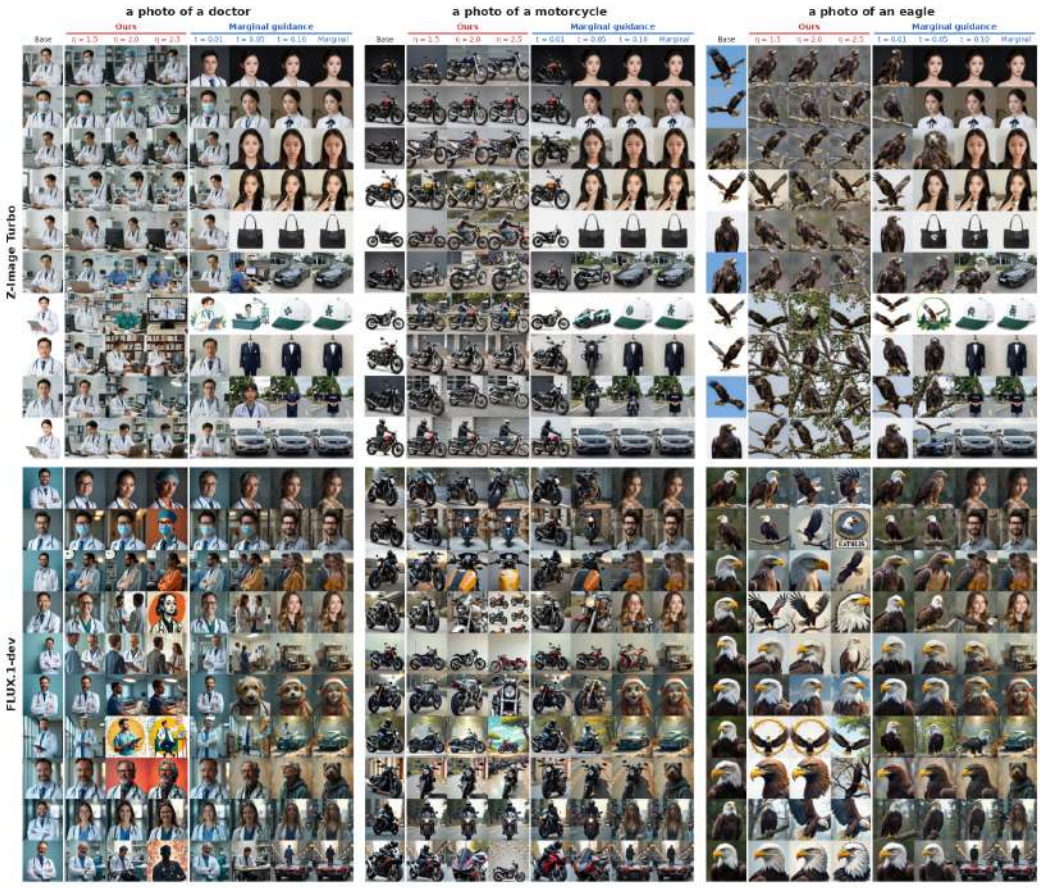


Figure 5: Qualitative comparison on z-image and FLUX-dev (rows denote fixed noise seeds). Marginal guidance (rightmost columns) has inconsistent prompt adherence, forcing identical seeds to collapse into prompt-ignoring concepts (e.g., handbag, baseball cap, dog creture) if the initial noise does not naturally align with the target. In contrast, properly scaled velocity negation (η) increases intra-class diversity while maintaining consistent prompt alignment.

with the target concept. For example, given the exact same "doctor" prompt, one noise seed might successfully yield a woman doctor, while other seeds completely ignore the conditioning and collapse into unrelated concepts like a dog, a handbag, or a car. Furthermore, these failure modes are tied to the seed itself: an unaligned noise seed will collapse into the exact same unrelated object (e.g., a handbag) regardless of whether the prompt specifies a doctor, a motorcycle, or an eagle.

In contrast, our velocity negation mechanism ($\eta \in \{1.5, 2.0, 2.5\}$) consistently preserves the target semantics across all noise initializations while significantly enhancing intra-class diversity (e.g., yielding varied poses, demographics, and backgrounds).

E Ablation on scheduler function

To further explore alternative schedule functions, we evaluate smooth, continuous functions that start at a negative value during early timesteps (to promote initial diversity) and increase to 1 (to ensure prompt alignment).

To systematically ablate the schedule function, we model the schedule as a continuous function $s(t)$ over the normalized generation process $t \in [0, 1]$, where $t = 0$ corresponds to pure noise and $t = 1$ to the final generated sample. We evaluate initial values $S_{\text{start}} \in \{-0.25, -0.5, -0.75\}$ with 50 prompts randomly sampled from the GenEval. An example profile of the scheduler function is illustrated in Fig. 6.

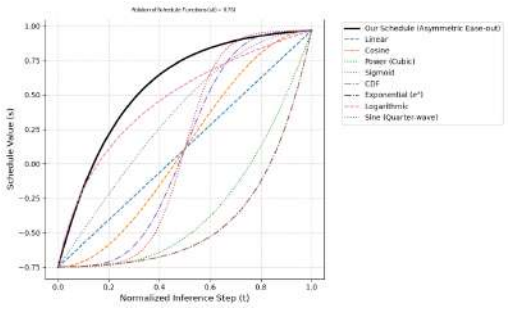


Figure 6: Illustration of the scheduler function, initialized at $S_{\text{start}} = -0.75$.



Figure 7: Qualitative results illustrating representative functions across different categories. ($S_{\text{start}} = -0.25$)

Under these unified constraints, we benchmark **9 alternative functional forms** across four structural categories:

1. **Ease-Out / Concave** (*ours, logarithmic, sine*): Rapid early transition to positive values.
2. **Linear / Constant Rate** (*linear*): Steady, constant increase over time.
3. **S-Curve / Sigmoidal** (*sigmoid, cdf, cosine*): Gradual acceleration followed by deceleration.
4. **Ease-In / Convex** (*power, exponential*): Delayed transition, staying negative longer.

As shown in Fig. 6, our proposed schedule achieves a stable tradeoff between quality (CLIP, PickScore, HPSv2) and diversity (DINO, DreamSim, Vendi, MSS) metrics, managing the gradual degradation of quality as diversity increases:

Function Trajectory Comparison Across Categories

- **Ease-Out / Concave (Ours, Logarithmic, Sine):** Rapidly transitions from negative values to positive alignment territory during early sampling timesteps. This allows the model to explore early structural variations while reserving most of the sampling trajectory for strong prompt conditioning, yielding the best overall results.
- **S-Curve / Sigmoidal / Constant Rate (Sigmoid, CDF, Cosine, Linear):** Spends too much time accelerating in the mid-range timesteps. While diversity remains decent, visual quality drops sharply compared to Ease-Out schedules.
- **Ease-In / Convex (Power, Exponential):** Stays negative for too long in early-to-mid timesteps. This severely degrades image generation quality (CLIP drops down to ~ 19.0) and leads to mode collapse or severe artifacts (very low DreamSim and Vendi scores).

F Qualitative Behavior on 2D Toy Distributions

In this section, we provide a qualitative analysis of our method using a 2D toy dataset. The target distribution is a 10-mode Gaussian mixture with highly imbalanced mixture weights ($p = [0.60, 0.20, 0.08, 0.04, 0.03, 0.02, 0.01, 0.01, 0.005, 0.005]$), designed to test whether the method can recover rare modes.

By varying the parameter η, we observe a direct impact on state-space exploration (see Figure 9). Increasing η drives the model to explore low-probability regions, thereby enhancing sample diversity and recovering rare Gaussian modes. However, this introduces a strict quality-diversity trade-off: if η is set too high, the trajectories overshoot the target modes and exit the learned data manifold, ultimately degrading sample validity.

G Empirical and Entropy Analysis of Opposite Guidance

While both Opposite Guidance (OG) and our method utilize a negative scaling component, their motivations and theoretical behaviors differ significantly. OG applies negative guidance primarily

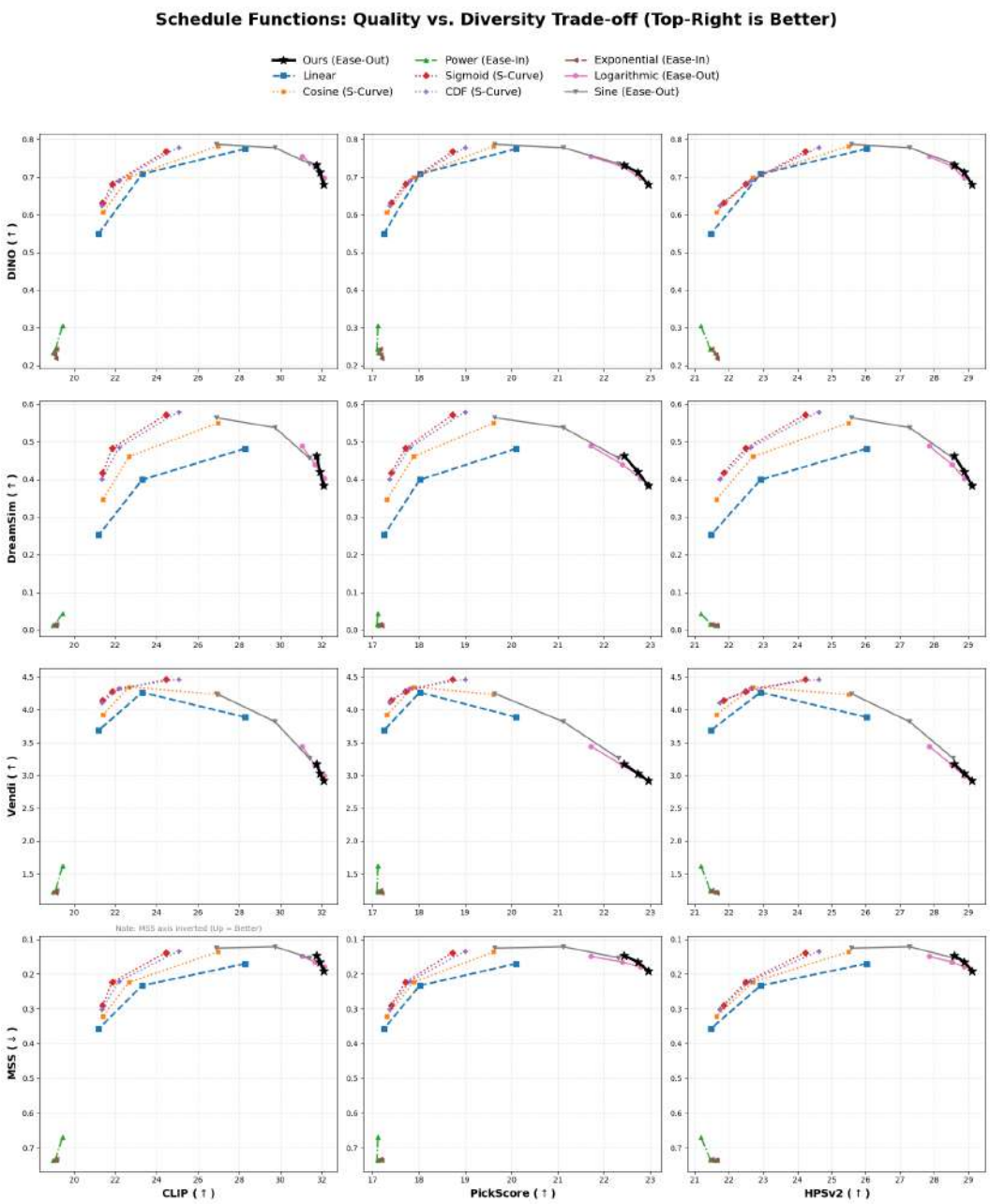


Figure 8: Quality versus diversity trade-offs across different scheduling functions.

to escape memorization basins. However, avoiding memorization does not inherently guarantee increased sample diversity. Here, we analyze OG through the lens of differential entropy to explain this limitation.

Theoretical Guarantee of Entropy Increase. Generative baseline velocity fields are contractive in expectation, mapping a high-entropy prior to a structured, lower-entropy data distribution (i.e., $\mathbb{E}[\nabla \cdot \mathbf{v}_\emptyset] < 0$ and $\mathbb{E}[\nabla \cdot \mathbf{v}_p] < 0$). Substituting the OG velocity $\mathbf{v}_{OG} = \mathbf{v}_\emptyset - s(\mathbf{v}_p - \mathbf{v}_\emptyset)$ into the continuity equation yields its entropy rate:

$$\frac{dH_{OG}(p_t)}{dt} = (1+s)\mathbb{E}_{\mathbf{x}}[\nabla \cdot \mathbf{v}_\emptyset] - s\mathbb{E}_{\mathbf{x}}[\nabla \cdot \mathbf{v}_p]$$

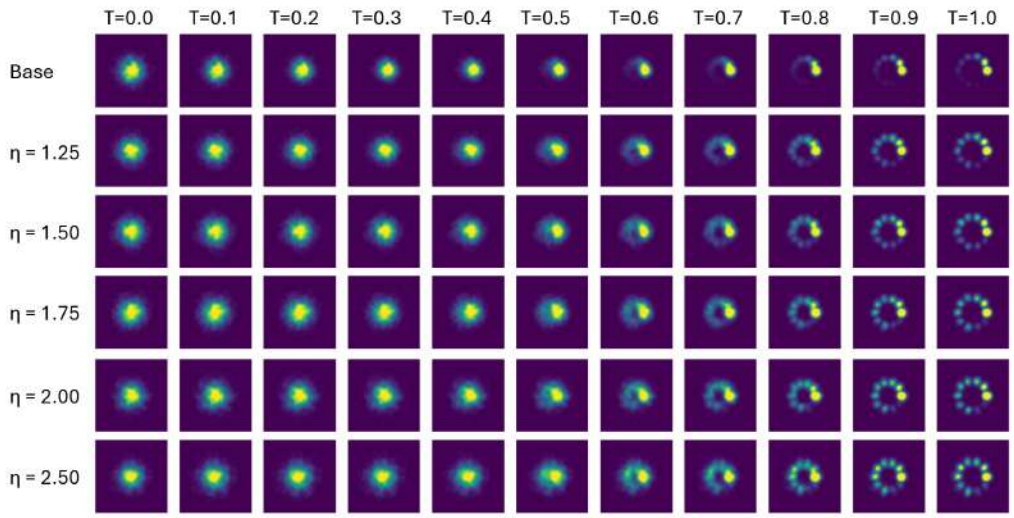


Figure 9: Experiments on a 2D toy dataset containing 10 different modes with varying probability densities demonstrate that increasing η improves the coverage of lower-probability modes.

Because both divergence terms are negative, increasing the scale parameter s creates a dynamic where the two terms are effectively fighting each other: the conditional term ($-s\mathbb{E}_{\mathbf{x}}[\nabla \cdot \mathbf{v}_p]$) acts to increase entropy, while the heavily weighted unconditional term ($(1+s)\mathbb{E}_{\mathbf{x}}[\nabla \cdot \mathbf{v}_\emptyset]$) acts to reduce it. Because this entropy-reducing term can easily overwhelm the entropy-increasing term, OG lacks a mathematical guarantee of positive differential entropy. In contrast, our method explicitly ensures increased sample diversity by scaling the negative velocity relative to the model's final velocity.

Empirical Evaluation. This theoretical limitation is directly reflected in our empirical results. As demonstrated in the table below, scaling OG's negative guidance parameter does not guarantee a monotonic increase in sample diversity. Specifically, increasing s from 1.0 to 2.0 actually harms diversity: the DINO score decreases from 0.6628 to 0.6333, and MSS worsens from 0.2162 to 0.2571 in mean quality metrics. see in Fig 2

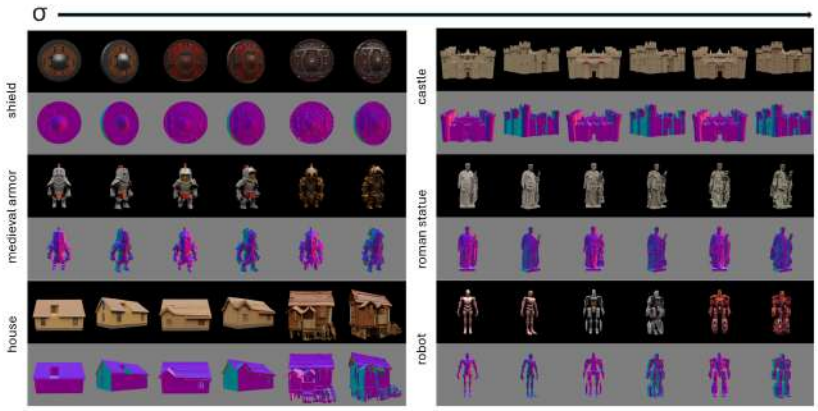
To further explain this phenomenon, we tracked the empirical entropy over time during generation using the same setup as Fig.1:

This empirical tracking (Fig. 11) aligns with our theoretical proof. Unlike our approach (η), which explicitly increases differential entropy over time to promote structural variation, OG cannot guarantee an entropy increase. This explains the observed reduction in sample diversity when scaling the parameter s.

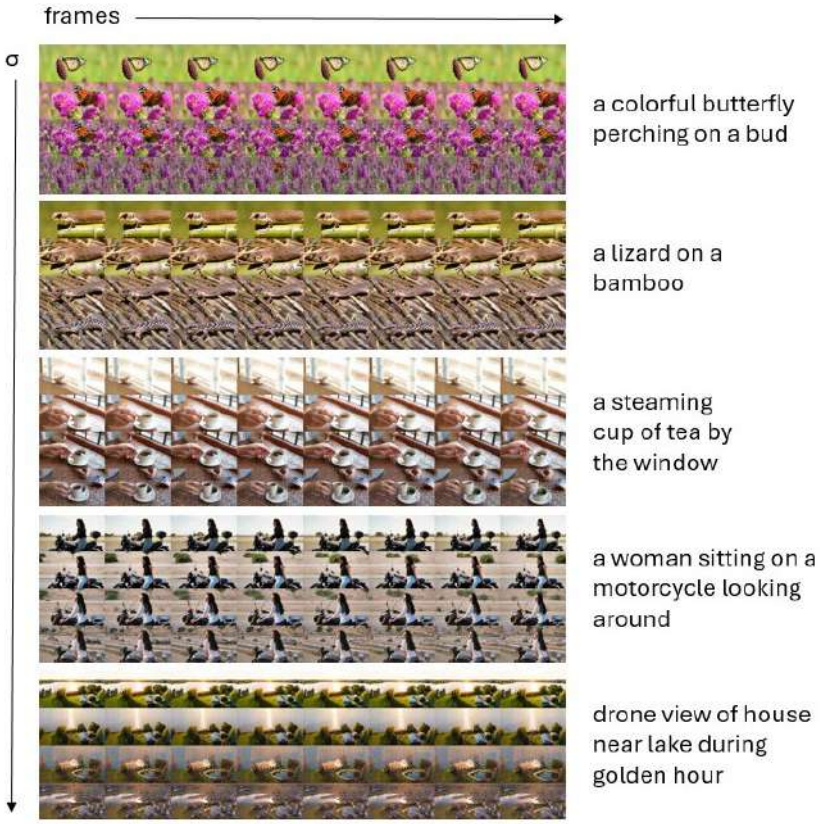
H Evaluating Unconditional Generation on Real Images

Reference-based metrics such as precision and recall require access to the model's training distribution, which is unavailable for large text-to-image models such as FLUX, whose training data are not disclosed. To verify that the diversity gained by velocity inversion does not come at the expense of fidelity to real data, we therefore evaluate it in a setting where the training data is known: unconditional generation with a latent flow matching (LFM) model (Dao et al., 2023) trained on LSUN Bedroom. This setting cannot be used for CADS, IG, or OG, which perturb the text embedding or modulate classifier-free guidance and therefore require a conditioning signal. In contrast, velocity inversion only rescales the model's velocity prediction and applies directly to unconditional models.

We generate 2,000 samples with 100 sampling steps for both the base model and our method ($\eta = 1.25$), and evaluate them against real LSUN Bedroom images using precision and recall (Kynkäänniemi et al., 2019) and density and coverage (Naeem et al., 2020). As shown in Table 2, our method improves all four metrics beyond the base model.



(a) **3D Assets:** Progression as σ increases from left to right.



(b) **Video Generation:** Progression as σ increases from top to bottom.

Figure 10: **Impact of σ on high-frequency details across modalities. (a)** At lower σ values, the 3D models prioritize coarse, smooth geometries. As σ increases, the model synthesizes intricate micro-details and complex surface topologies (e.g., detailed wooden planks on the *house*, metallic rivets on the *shield*, and layered plating on the *medieval armor*). This is corroborated by the corresponding normal maps, which shift from smooth gradients to dense, varied normals. **(b)** For video generation, increasing σ similarly enhances the high-frequency details in the background.

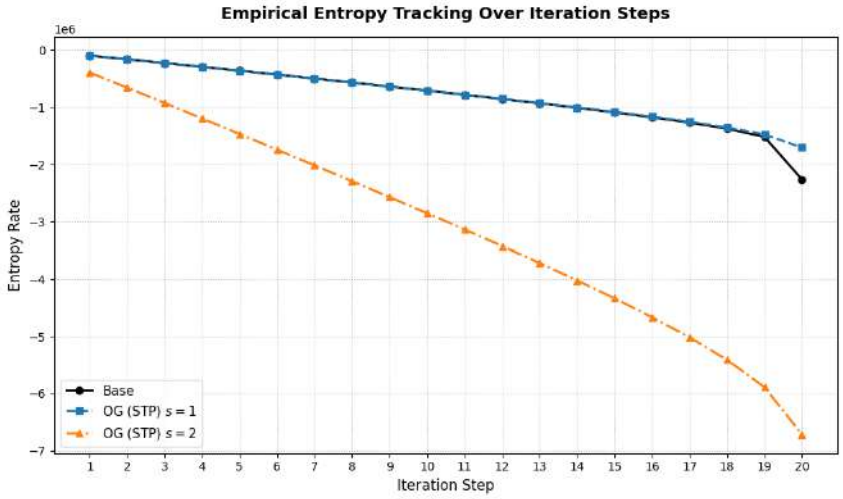


Figure 11: **Entropy of OG over sampling steps.** Estimated entropy for the base model and OG+STP with $s \in 1, 2$, computed as in Fig.1. With $s = 1$, OG closely follows the base model; increasing the scale to $s = 2$ lowers entropy throughout sampling instead of raising it, consistent with the decrease in diversity observed in Fig. 2.

Table 2: Quantitative comparison of unconditional generation using LFM on the LSUN Bedroom dataset. Evaluation is based on 2,000 samples generated with 100 sampling steps.

Model	**Precision**	**Recall**	**Density**	**Coverage**
Base	0.5950	0.4590	0.8852	0.7850
Ours ($\eta = 1.25$)	**0.6210**	**0.4645**	**0.9543**	**0.8225**

I Hyperparameter Selection

We randomly sample 50 prompts from T2I-CompBench Huang et al. (2023). We explore $\sigma \in [0.1, 0.3]$ and $\eta \in \{1.5, 2.0, 2.5\}$, evaluating the Vendi score and PickScore as shown in Fig. 12. For our main results, we select $\sigma = 0.25$, as this combination achieves the highest Vendi score while maintaining a PickScore comparable to the baseline at $\eta = 1.5$.

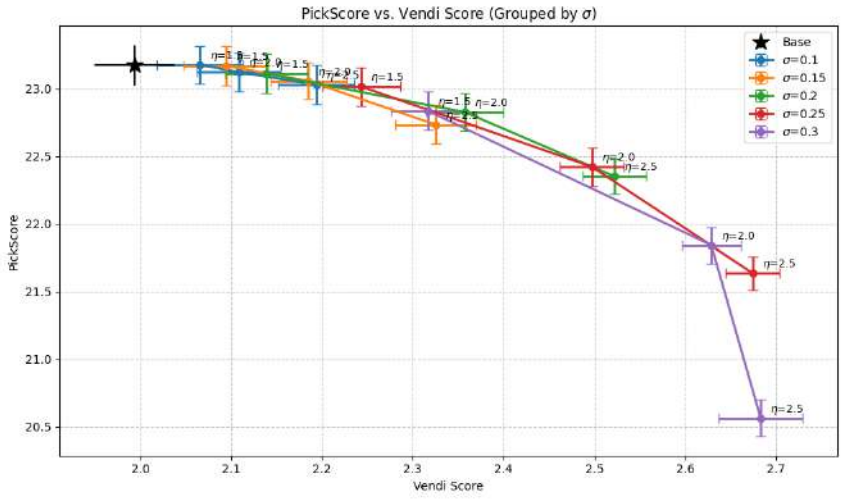


Figure 12: Diversity–fidelity trade-off on T2Bench (diversity/quality SE $N = 50$ prompts).



Figure 13: Alternative strategy for increasing diversity: adding small noise to $\mathbf{x}_t$ during sampling. Each row displays samples at *the end* of the generation process, with noise magnitude increasing from top (no noise) to bottom (highest noise). Columns represent samples generated from different initial noise seeds. Prompt: *a photo of a bench.*



Figure 14: Alternative strategy for increasing diversity: using non-negative velocity $\in (0, 1]$. Each row displays samples at *the end* of the generation process, with σ increasing from top (I.I.D. samples) to bottom (highest σ), and columns show different samples generated from distinct initial noise seeds. Prompt: *a photo of a bench.*

J Qualitative Results

We present qualitative results across several prompts from both the GenEval (Fig. 23) and Complex-Prompt (Fig. 22) . Additionally, we include results on other architectures, including FLUX.1-schnell (Fig. 25) and Z-Image (Fig. 24). These results demonstrate that our method effectively improves diversity across prompts and architectures while preserving image quality.

K Negative Velocity vs. Non-negative Velocity vs. Isotropic Noise

We explore alternative approaches to increasing velocity as discussed in Section 5.1. Although adding noise can theoretically increase entropy over time, it often introduces artifacts and fails to improve diversity without degrading image quality. We therefore weight the noise using α and a $(1-t)^{10}$ schedule to restrict it to high-noise regions. Despite this, the generated samples collapse before achieving high diversity as seen in Fig. 13.

Using a non-negative velocity schedule to delay convergence similarly fails to escape the initial basin of attraction; while it slows the contractive process, the trajectories ultimately settle into nearby high-density modes, resulting in incomplete denoising rather than increased diversity as seen in Fig. 14

L Compute Resources

All experiments were conducted on a workstation with an AMD Ryzen Threadripper 2950X (16-core) CPU and an NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPU with 96 GB VRAM.

M 3D synthesis prompt

Fantasy & Medieval Props: A wooden treasure chest with rusted iron bands; A glowing blue health potion in a spherical glass flask; A medieval steel broadsword with a leather-wrapped hilt;

A weathered wooden tavern barrel; A worn leather spellbook with a glowing magical rune on the cover; An ornate silver goblet inlaid with red rubies; A sturdy wooden archery target with arrows stuck in it; A medieval kite shield painted with a golden lion crest; A glowing magical staff topped with a floating green crystal; A heavy iron blacksmith's anvil; A rustic wooden cart wheel; A small leather pouch overflowing with gold coins; A rusted iron dungeon key; A burning wooden torch with a wrought iron handle; A stone medieval water well with a wooden bucket; A bronze dwarven battleaxe with geometric engravings; A hanging iron cage; A large, open clamshell revealing a glowing pearl; A wooden apothecary shelf stocked with small colorful vials; A golden royal crown with velvet lining.

Sci-Fi & Cyberpunk Props: A metallic sci-fi cargo crate with glowing blue LED strips; A futuristic energy pistol with a sleek carbon-fiber body; A holographic computer terminal with a floating green interface; A rusted steampunk brass pressure valve; A futuristic medical pod with a translucent glass lid; A spherical floating robot drone with a single glowing red eye; A heavy sci-fi blast door made of reinforced steel; A glowing neon cyberpunk street sign shaped like a noodle bowl; A futuristic motorcycle helmet with a black visor and neon accents; A portable sci-fi fusion battery glowing with orange energy; A sleek white spacecraft engine thruster; A cyberpunk cybernetic arm with exposed hydraulic wiring; A high-tech sniper rifle with a digital scope; A metallic communication relay dish; A futuristic hoverboard with anti-gravity thrusters on the bottom; A small glass terrarium containing a glowing alien plant; A heavy-duty yellow construction mech arm; A cylindrical sci-fi stasis tube filled with bubbling green liquid; A metallic data drive shaped like a hexagonal crystal; A futuristic command chair with built-in control panels.

Urban & Everyday Objects: A classic red fire hydrant with chipped paint; A green metal street trash can; A vintage 1970s rotary telephone; A worn brown leather Chesterfield armchair; A modern ceramic coffee mug resting on a wooden coaster; A detailed mechanical wrist watch; A bright yellow plastic wet floor sign; A standard wooden shipping pallet; A rusty urban chain-link fence gate; A vintage brass desk lamp; A detailed acoustic guitar with a polished wood finish; A cardboard moving box sealed with packing tape; A modern black flatscreen television on a stand; A stainless steel kitchen refrigerator; A red brick with mortar residue; A classic yellow school bus tire; A glass Coca-Cola bottle; A retro boombox radio from the 1980s; A ceramic potted fern plant; A shiny red fire extinguisher.

Stylized & Casual Game Assets: A stylized low-poly wooden pine tree; A cute cartoonish red mushroom with white spots; A low-poly floating dirt island with green grass on top; A stylized fluffy white cloud; A low-poly campfire with glowing orange geometric flames; A stylized oversized wooden hammer; A cute low-poly yellow submarine; A stylized chunky stone pillar covered in moss; A cartoonish shiny gold coin with a star emblem; A stylized glowing purple magic crystal; A low-poly stylized wooden windmill; A cute cartoonish mailbox with a red flag; A stylized pirate cannon; A low-poly snowy mountain peak; A stylized chunky wooden bridge; A cartoonish bomb with a burning fuse; A stylized open book with blank pages; A low-poly sandy beach with a single palm tree; A cute stylized beehive hanging from a branch; A stylized green potion bottle with a cork stopper.

Simple Characters & Avatars: A simple wooden artist mannequin; A low-poly futuristic space marine in white armor; A cute chibi-style knight in shining armor; A stylized robot mascot character with wheels for feet; A basic untextured humanoid base mesh; A stylized fantasy goblin warrior holding a club; A classic 8-bit style voxel character; A stylized pirate captain with an eyepatch; A cute cartoonish astronaut in a round helmet; A stylized ninja wearing a black gi and mask; A low-poly medieval peasant character; A stylized zombie character with green skin; A simple crash test dummy; A cute stylized alien with large black eyes; A low-poly police officer character; A stylized wizard with a tall pointed hat; A basic anatomical skeletal mesh; A cute cartoonish farmer character holding a pitchfork; A stylized cyberpunk street thug character; A low-poly female archer character in green tunic.

N Experimental Setup and Justification for Additional Modalities/Distilled Model

Baseline exclusions. We exclude IG and OG, since these methods require an unconditional/marginal velocity estimate in the same way CFG does, and all guidance-distilled models in our evaluation (LTX-2.3 Distilled, Z-Image Turbo, FLUX.1-schnell) collapse this branch during distillation, leaving nothing for these methods to guide against. We omit SGI throughout as it depends on an image-specific reward model, and no equivalent off-the-shelf reward model exists for video or 3D. We exclude NegToMe from the video and 3D settings because its per-step token-matching mechanism assumes stable spatial correspondence between source and reference tokens, which is not well-defined for temporally-varying video tokens.

Distilled Image Models. We evaluate our guidance-based approach against the base model and representative methods from two distinct categories: SPELL (batch interaction-based) and CNO (noise initialization-based). For our method, we fix the hyperparameters to a single setting ($\eta = 1.5$, $\sigma = 0.1$), selected for achieving the highest Vendi score on the validation set while maintaining a PickScore comparable to the base model. We report final performance on the test set, comparing our single configuration against all swept configurations of the baselines. Both the validation and test sets are sampled from GenEval and consist of 50 prompts each.

Video and 3D Models. For our experimental setup, we evaluate our models using the VBench framework to assess both generation diversity and sample quality. We omit noise initialization and batch interaction baselines for the video and 3D modalities due to three primary factors: the lack of off-the-shelf reward models for these domains, the prohibitive memory overhead required to store and batch large samples, and the architectural complexity of adapting these methods (e.g., the audio-video cross-modality in LTX-Video and the dual flow matching models in TRELLIS).

To identify the optimal settings, we conduct a hyperparameter grid search (detailed in Figure 15). For our method, we evaluate parameters within the ranges of $\eta \in [1.25, 4.0]$ and $\sigma \in [0.01, 0.2]$. For the CADS baseline, we fix $\tau_2 = 0.95$ and perform a search across the intervals $\tau_1 \in [0.6, 0.9]$ and $s \in [0.15, 0.4]$.

We restricted the CADS search space to the recommendations from the original paper (i.e., high τ_1, high τ_2, and low s) to prevent sample collapse. However, during our search on LTX-Video, CADS failed to yield high-quality samples. Even when using very low scaling parameters to keep predictions close to the base model, every tested CADS combination fell below an acceptable Text-to-Video Alignment (TVA) score (qualitative results are provided in the supplement). While a functional configuration for CADS may exist, optimizing across its three-dimensional parameter space (τ_1, τ_2, and s) is computationally demanding. We provide an example in the supplementary website.

Our core criterion for final parameter selection is to maximize generation diversity while maintaining text alignment comparable to the base model. Therefore, we set $\eta = 2.0$ and $\sigma = 0.15$ for LTX-Video.

We do not tune our method separately on TRELLIS and instead reuse values from our image experiments: $\sigma = 0.1$, as selected for Z-Image Turbo, and $\eta = 2.0$, the midpoint of the range evaluated for images ($\eta \in \{1.5, 2.0, 2.5\}$). The base model uses its default configuration. Tuning on 3D generation could further improve the trade-off.

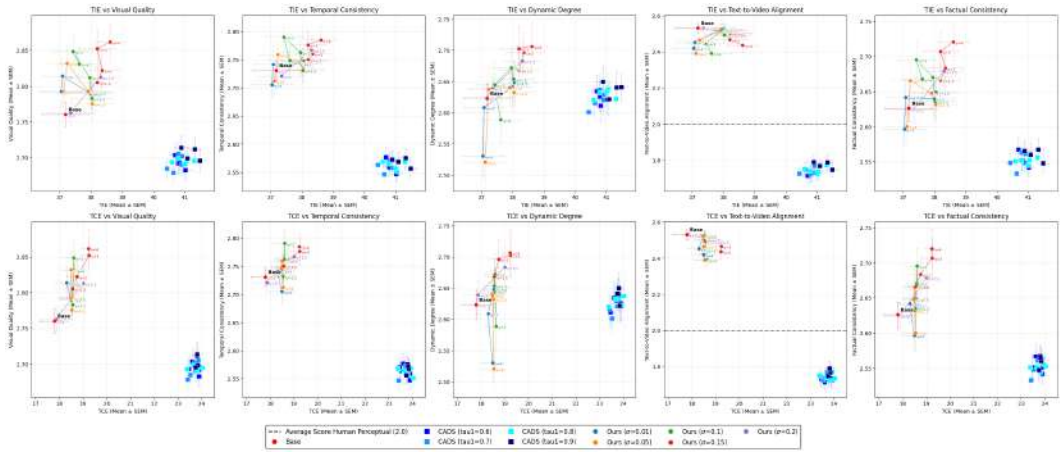


Figure 15: Diversity–quality trade-off on VBench (diversity $N = 40$ prompts and quality $N = 200$ videos, $\pm$ SE, 5 samples/prompt).

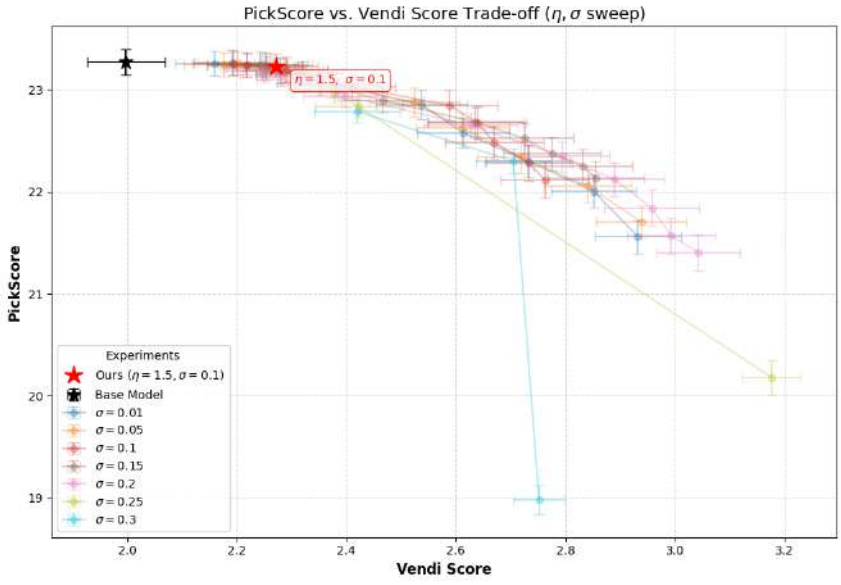


Figure 16: Diversity–quality trade-off on GenEval (diversity/quality $N = 50$ prompts, $\pm$ SE, 5 samples/prompt).



Figure 17: SPELL guarantees geometric separation but risks manifold departure, resulting in visual artifacts. Our method effectively increases sample variance within the available manifold range, ensuring both diversity and high-quality generation. Prompt: A photo of an eiffel tower

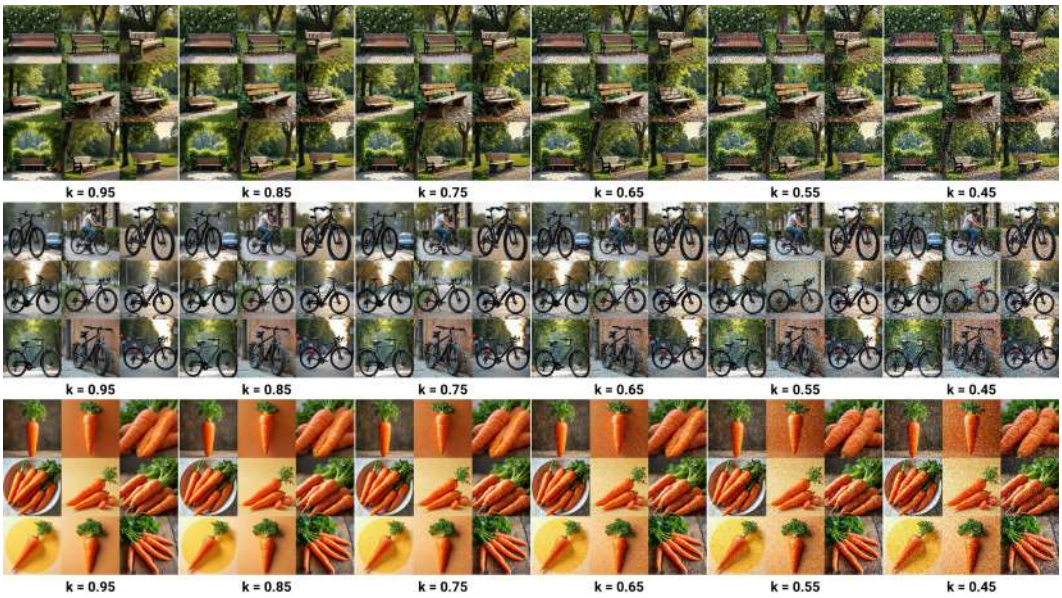


Figure 18: **TSR on FLUX.1-dev.** Using the official text-to-image setting ($\sigma = 1$), we sweep $k \in 0.95, 0.85, 0.75, 0.65, 0.55, 0.45$. Decreasing k mainly adds high-frequency detail without changing the global structure of the samples, and thus does not increase structural diversity.

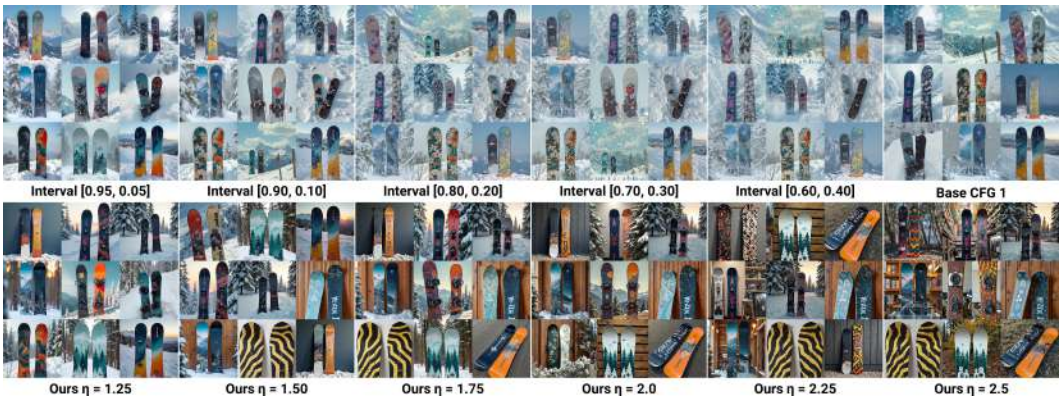


Figure 19: **Comparison of Diversity with IG.** While **IG CFG1** limits generation diversity to the base CFG=1 level, **Ours** is not bound by this restriction. As demonstrated in these qualitative samples, our approach at $\eta = 1.75$ achieves greater diversity than both the base CFG=1 and IG [0.60, 0.40], all while preserving image quality.



(a) Prompt: a photo of a clock

(b) Prompt: a photo of a backpack

Figure 20: **Comparison of Diversity with Noise-Initialization-Based Methods.** While generation diversity for **CNO** and **SGI** plateaus at $\gamma = 1.0$ and $\lambda = 0.9$, respectively, **our method** is not bound by this restriction. As demonstrated in these qualitative samples, our approach at $\eta = 1.7$ achieves greater diversity than both CNO and SGI while maintaining competitive image quality.



(a) Prompt: a photo of a train

(b) Prompt: a photo of a backpack

Figure 21: **Comparison of Diversity with OG.** For the OG method, it is ambiguous whether increasing s increases or decreases diversity. Additionally, the marginal velocity effect produces inconsistent image quality across different prompt concepts. In contrast, increasing η in our method consistently increases generation diversity.

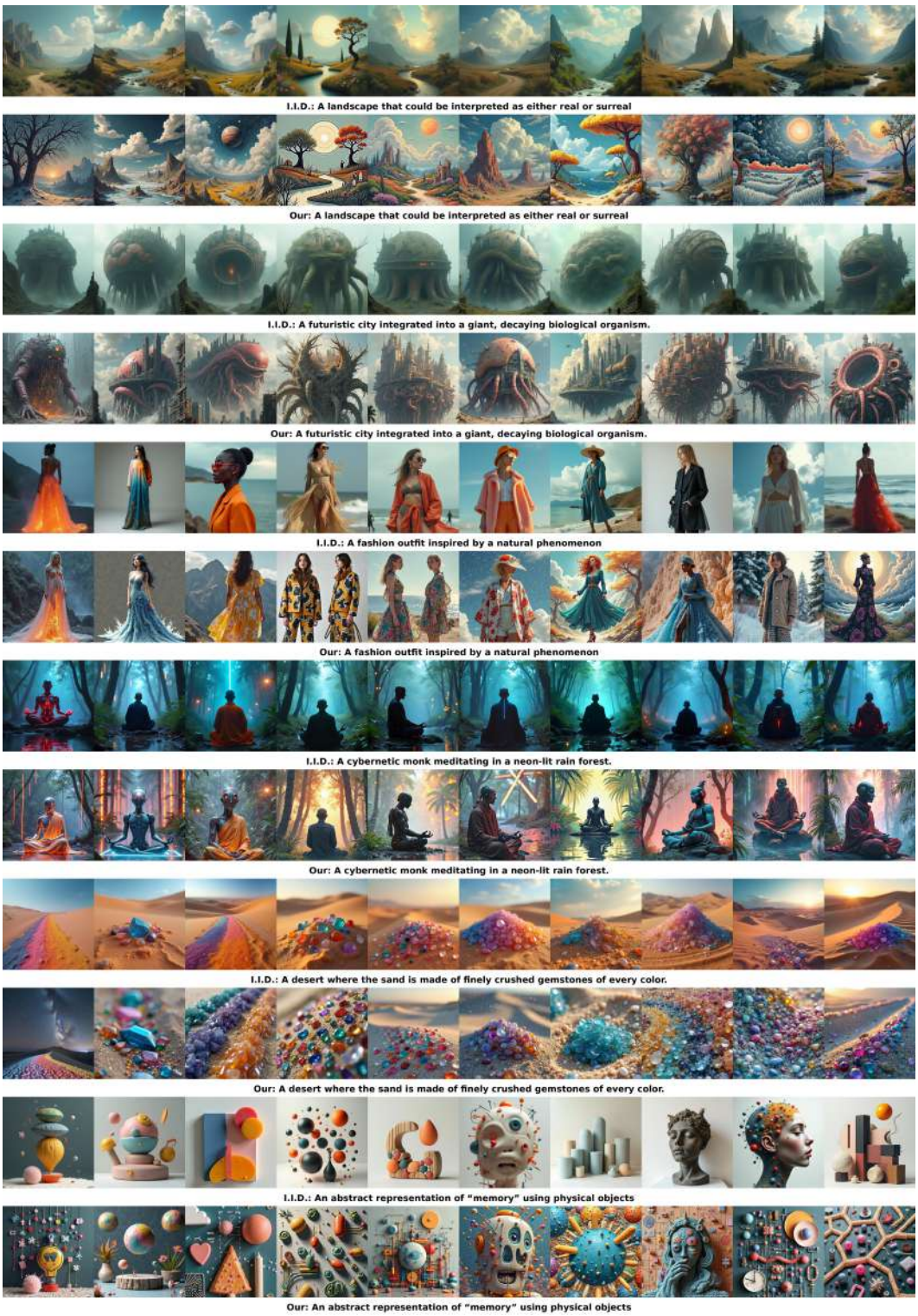


Figure 22: Flux-Dev results on the ComplexPrompt evaluation set (ChatGPT-synthesized prompts).
.

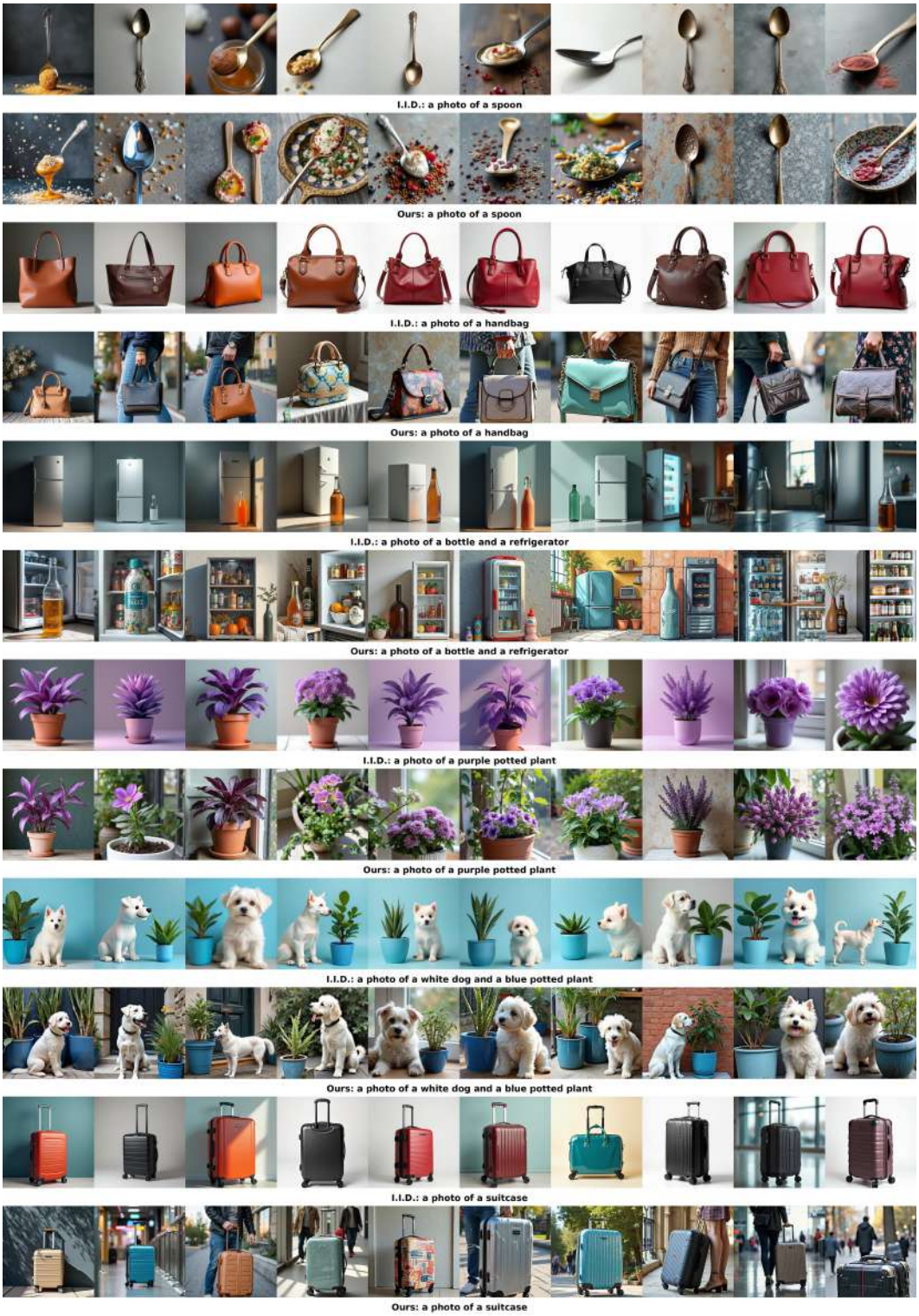


Figure 23: Flux-Dev results on the GenEval benchmark.



Figure 24: Z-Image results on the GenEval benchmark.

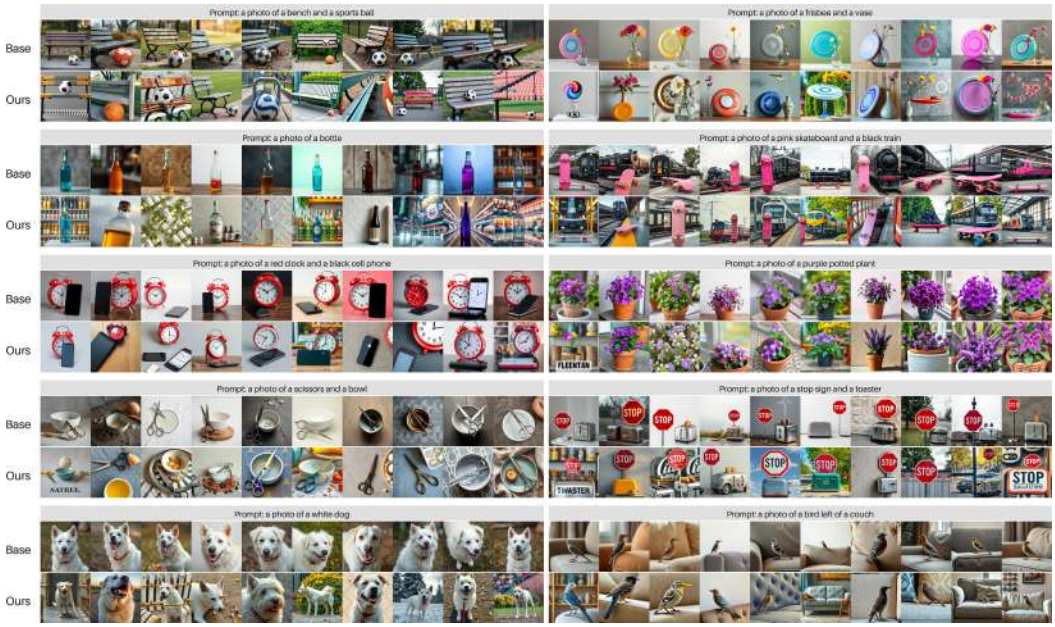


Figure 25: Flux Schnell results on the GenEval benchmark.

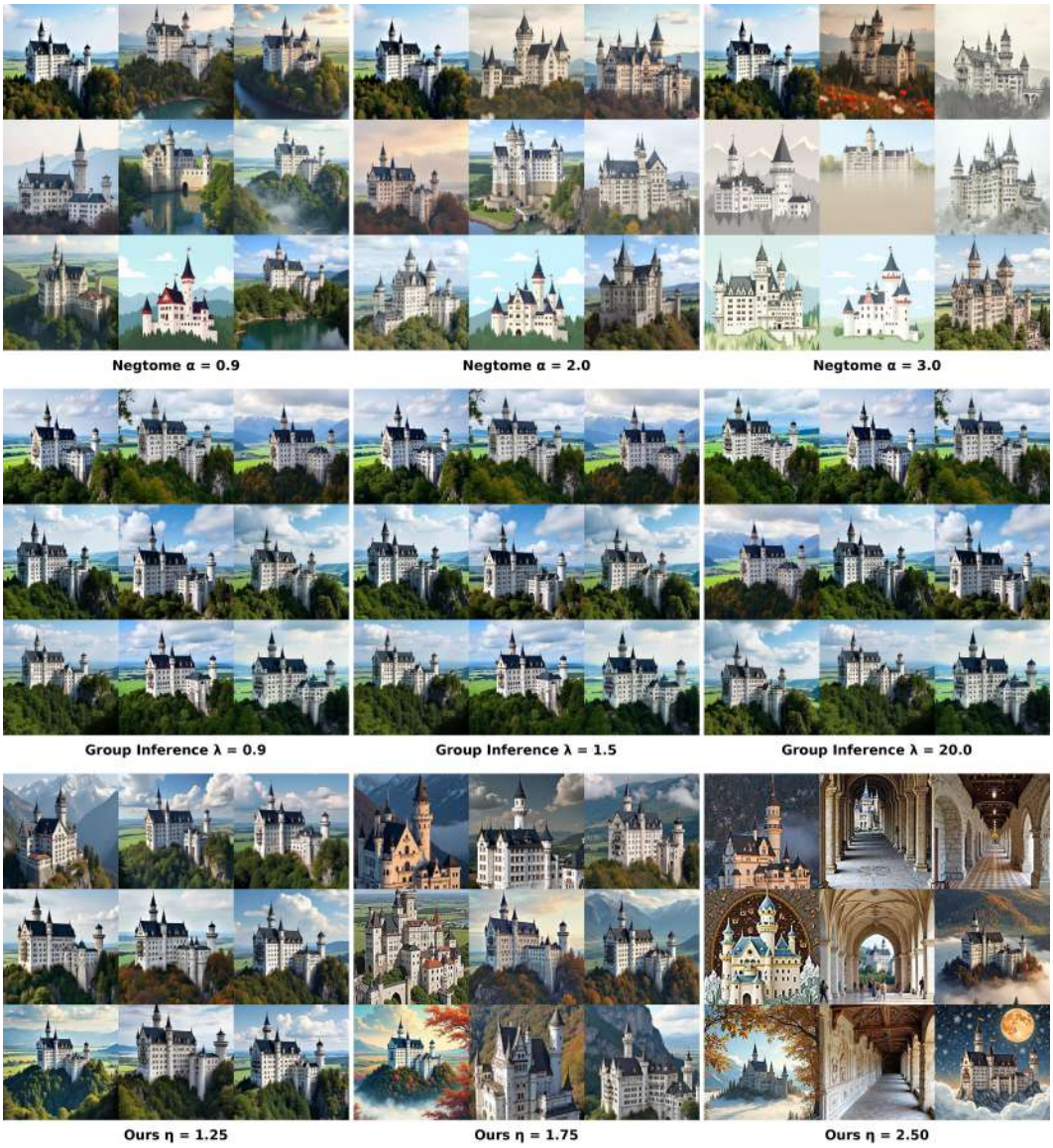


Figure 26: **Comparison of diversity between batch interaction methods.** While Group Inference suffers from mode collapse and Negtome exhibits limited diversity by merely shifting away from a single reference, our proposed method simultaneously achieves high sample variance and superior image quality. Prompt: A photo of Neuschwenstein Castle

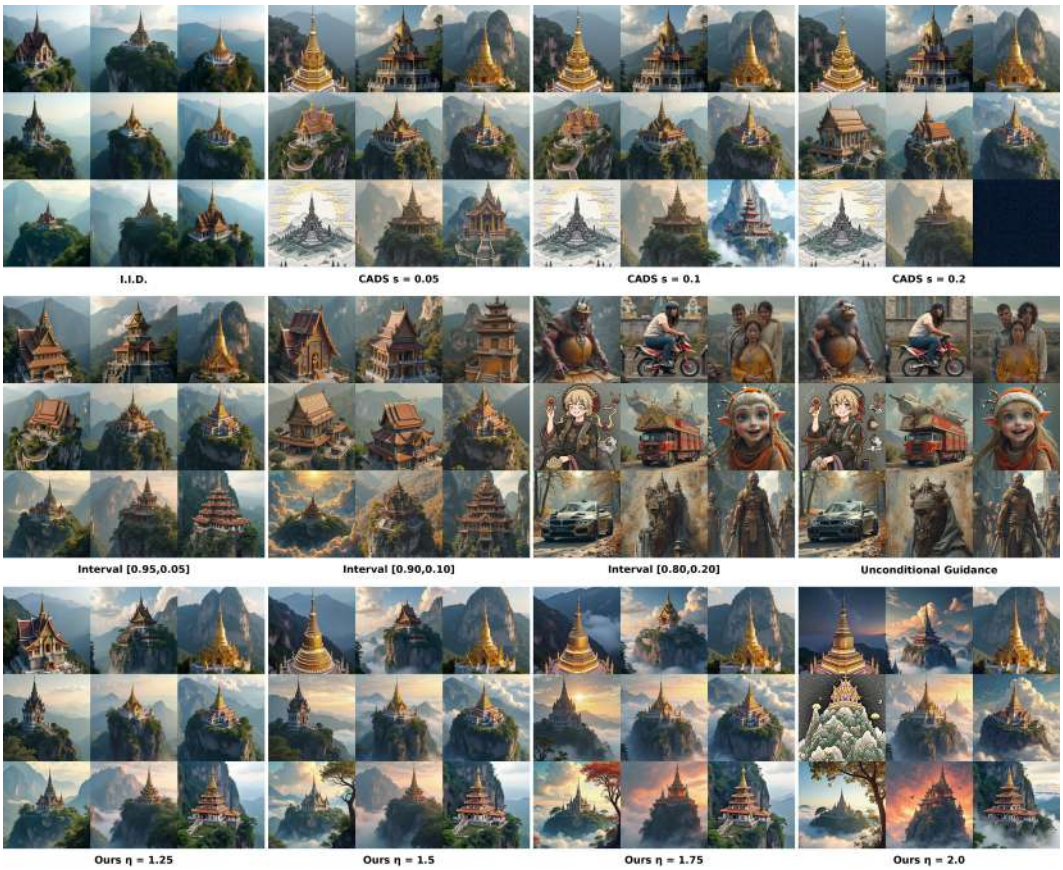


Figure 27: **Comparison of Diversity with Guidance Base. CADS:** High guidance weights in Condition-Annealed Sampling can push images outside the learned manifold, resulting in degraded image quality. **IG marginal:** Restricting guidance to a specific interval pulls samples toward the marginal distribution, which may reduce prompt alignment as the high noise levels are unguided. **Ours:** By focusing on increasing entropy over time, our approach generates high-quality samples that maintain strong prompt alignment while significantly enhancing sample diversity. Prompt: A photo of Thai temple on a mountain

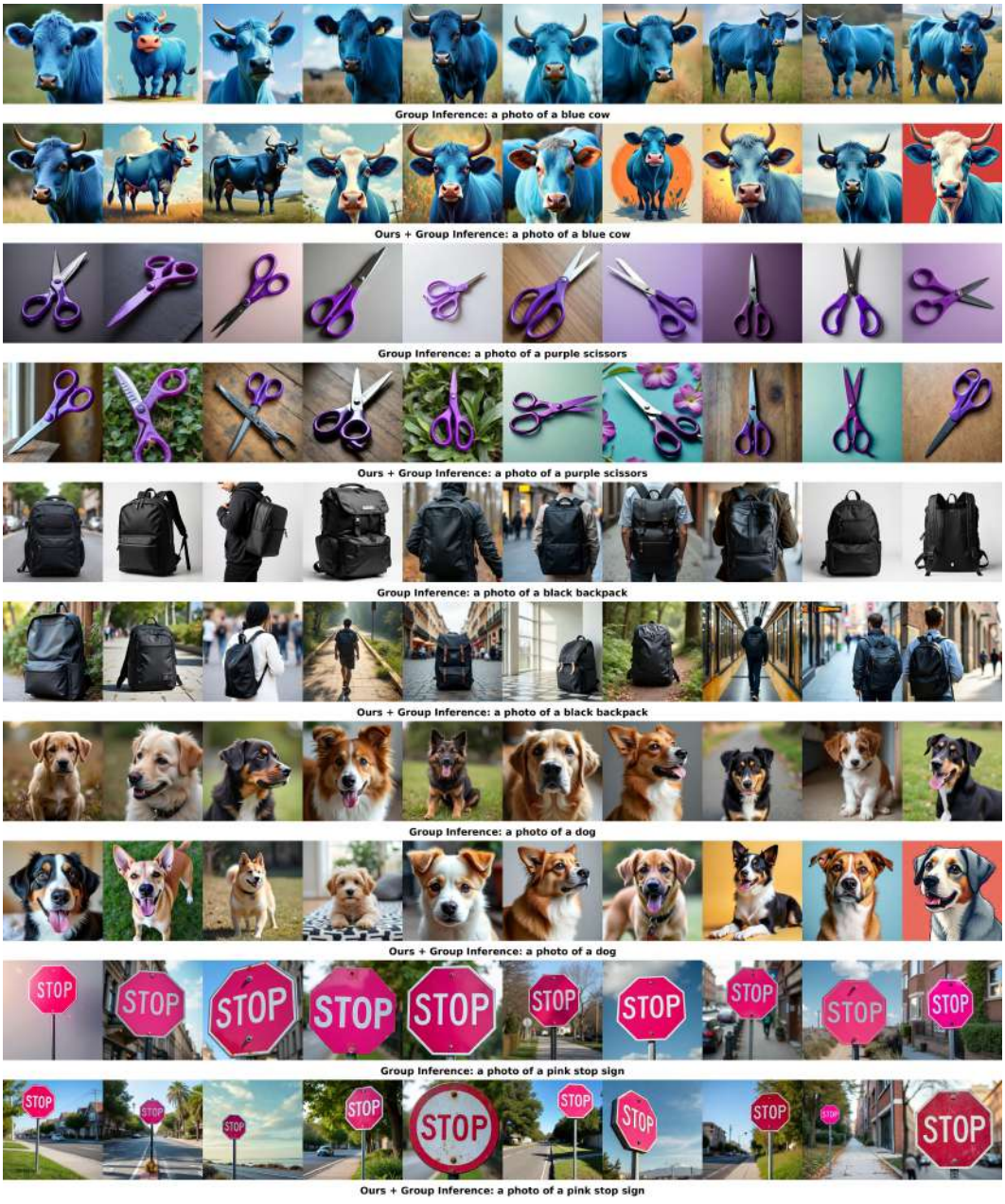


Figure 28: Our method increases the diversity of group inference by modifying the velocity , and SGI selects the noise . Using $\eta = 2.0$, the approach enables varied sample generation while maintaining image quality.