
Diversify Diffusion with Temperature Sampling and Variance-Corrective Time Shifting

Peizhuo Li
ETH Zurich

Emre Aksan
Google

Alexandru-Eugen Ichim
Google

Thabo Beeler
Google

Olga Sorkine-Hornung
ETH Zurich

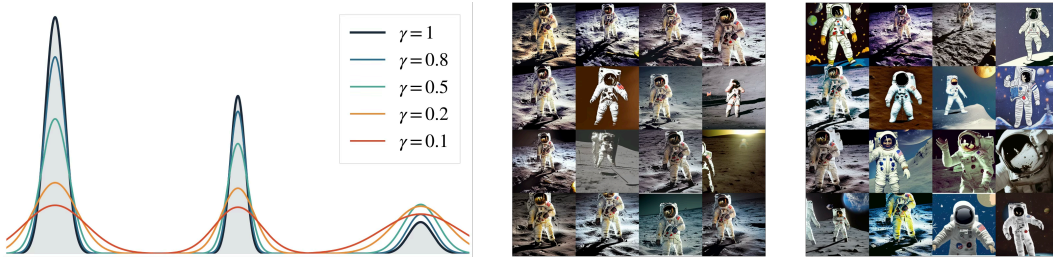


Figure 1: Left: The effect of increasing temperature (*i.e.* lowering γ) on a 1D Gaussian mixture. Middle: Sampling results with prompt “an astronaut on the moon” using Stable Diffusion. Right: The effect of our method on the same prompt and model.

Abstract

Diffusion models faithfully reproduce their training distribution, but also inherit its imbalances and leave rare or under-represented modes hard to reach. A natural inference-time remedy is to sample from the high-temperature target $p_0^{(\gamma)}(x) \propto p_0(x)^\gamma$ for $0 < \gamma < 1$, which flattens dominant modes and lifts rare ones. However, naive score scaling while correctly reweighting modes also inflates the per-mode variance, breaking the reverse diffusion process and degrading sample quality. We introduce *variance-corrective time shifting*, a training-free fix that queries the network at a shifted timestep and scales the resulting score by γ , canceling the variance inflation while preserving the mode reweighting. The correction turns simple temperature sampling into a practical diversity knob for pretrained diffusion and flow-matching backbones with no retraining, and we demonstrate consistent gains at minimal cost to sample quality and condition fidelity across DiT, Stable Diffusion and Motion Diffusion models. We further show that the timing of the temperature intervention enables coarse-to-fine control: high-noise stages drive compositional diversity across modes, while low-noise stages drive local appearance variation under a fixed composition.

1 Introduction

Diffusion models [11, 31, 33] have emerged as the dominant paradigm for generative modeling across images, video and 3D content. Beyond their high quality output, another key strength of diffusion models is that they offer a flexible inference-time control mechanism [10], which allows users to steer the sampling process without retraining the model. Yet, their samples often reflect the biases of

the data they are trained on: common modes are reproduced reliably, while rare or under-represented modes are unlikely to be generated. Although this behavior aligns the goal for matching the training distribution, the lack of diversity could be problematic in creative settings, where users care about exploring multiple plausible outcomes rather than a single output.

A natural way to surface rare modes is to draw from the high-temperature target

$$p_0^{(\gamma)}(x) := \frac{p_0(x)^\gamma}{Z_\gamma}, \quad Z_\gamma := \int p_0(x)^\gamma dx, \quad (1)$$

which flattens dominant modes and boosts the relative probability of rare events for $0 < \gamma < 1$, as shown in Fig. 1 (left). Here, $p_0(x)$ is the original data distribution and γ is the *inverse temperature* parameter that controls the strength of the effect, with $\gamma = 1$ recovering the original distribution and $\gamma \rightarrow 0$ approaching a uniform distribution, *i.e.* infinitely high temperature. In diffusion models, a neural network learns the score, the logarithmic likelihood gradient, and gradually denoises the random distribution $p_T(x)$ to the data distribution $p_0(x)$ guided by the score. The temperature sampling looks deceptively simple on score: the identity $\nabla_x \log p_0^{(\gamma)}(x) = \gamma \nabla_x \log p_0(x)$ holds strictly and conveniently cancels the unknown normalizer Z_γ . One might therefore hope to obtain $p^{(\gamma)}$ by multiplying the learned score by γ throughout the reverse diffusion process.

However, naive score scaling conflates two effects. It achieves the *macroscopic* goal of reweighting modes, but it also induces an unwanted *microscopic* variance expansion that introduces more noise than expected, as demonstrated in Fig. 1 (left). The excess variance breaks the strict correspondence between the score of the noised tempered marginal $\nabla_x \log p_t^{(\gamma)}(x)$ and the reverse diffusion process, as shown in Sec. 3.1. It lowers sample quality and produces off-manifold and blurry samples.

To mitigate this issue, we introduce *variance-corrective time shifting*, a training-free fix that decouples the two effects: rather than querying the network at the true timestep t , we query it at a shifted timestep $\tilde{t}$ whose smaller noise level cancels the variance inflation. Concretely, presenting the model with the noisy sample x_t at a smaller noise level and scaling the resulting score by γ lands the effective conditional variance back on the original schedule while preserving the mode reweighting. This cancelation is a very accurate approximation under an assumption of well-separated modes on the data (Sec. 3.1), which empirically performs well for natural images and motion generation considered in our experiments.

We show that with this method, sampling diversity is improved on a range of models, including DiT [22], Stable Diffusion [5, 24, 26] and Motion Diffusion Model [35], at minimal cost to sample quality and condition fidelity. We also explore the effects of temperature scaling on the different stages of the diffusion process. Similarly to the natural coarse-to-fine reverse diffusion process, applying $\gamma(t) < 1$ at early (high-noise) timesteps yields compositional diversity across modes, while applying it at later (low-noise) timesteps yields local appearance variation under a fixed composition.

In summary, our contributions are:

- We enable simple temperature sampling as a training-free mechanism for enhancing diversity in pretrained diffusion models, where naive application fails to produce usable samples.
- We identify the *variance expansion problem* of naive temperature sampling in diffusion models, which breaks the reverse diffusion process and degrades sample quality.
- We propose *variance-corrective time shifting*, a zero-cost correction that recovers the desired tempered marginal by querying the network at a shifted timestep.
- We demonstrate the effectiveness of our method across various models and modalities, showing consistent gains in diversity at minimal cost to sample quality and condition fidelity.

2 Related work

Diffusion models [11, 31–33] have quickly achieved superior performance compared with GANs [7] and became the gold standard for generative modeling. Beyond the high quality itself, another key strength of diffusion models is the flexible inference-time control mechanism [1, 4, 10, 30], which allows users to steer the sampling process without retraining the model.

Pioneered by Kirkpatrick et al. [15], temperature scaling or sampling has been used in various deep learning field [9, 12, 20, 29]. However, the majority of the work in generative modeling focuses on *lowering* the temperature to sharpen the distribution and produce higher-quality samples. One exception is TSR [37], which applies temperature scaling to control the variance of the output distribution. However, this method suffers from the mentioned variance expansion problem, and cannot be used to produce high-quality and diverse samples when raising the temperature. In contrast, our method is designed to mitigate this issue and enable effective high temperature sampling for diversity enhancement.

Several recent methods enhance diversity in pretrained diffusion models without retraining, each approaching the problem from a different angle. Sehwal et al. [28] bias the sampler towards low-density regions of the data manifold, but they require a pre-trained classifier, whereas our approach is agnostic to data-distribution and requires no auxiliary annotation. CADS [27] anneals the conditioning signal at high noise to broaden conditional generation, but it operates on the conditioning, rather than the data distribution itself, and therefore tends to produce samples that are less faithful to the condition. Closer to our goal of sampling from $p_0^{(\gamma)}$, Feynman-Kac correctors (FK) [30] use sequential Monte Carlo particle filtering to correct the target distribution exactly, which is mathematically clean and distribution-agnostic but requires running a large batch of particles in parallel. Although it performs well in the annealed setting ($\gamma > 1$), this method struggles with high temperatures ($\gamma < 1$) in our experiments. A parallel technique particle guidance [2] adds a pairwise repulsive potential across particles in a sampled batch, inducing non-i.i.d. batch-level diversity. It operates *across* particles, while our method operates *within* each particle.

A separate line of work modifies the training-time formulation to better represent long-tailed data, but does not directly increase the probability of sampling from the tails. Heavy-tailed diffusion [21] replaces the Gaussian prior with a t -distribution, so that the model can capture rare events present in the training data. This improves how the long tail is *represented* but does not change the relative sampling probabilities, so it does not on its own increase the chance of generating rare modes the way temperature sampling does.

3 Method

3.1 Variance-corrective time shifting for temperature sampling

For simplicity, we consider a variance-exploding (VE) diffusion with forward kernel $p(x_t | x_0) = \mathcal{N}(x_t; x_0, \sigma_t^2 I)$, where σ_t is a noise schedule that increases with time. We work with unnormalized densities below, since normalizing constants do not affect the score function or the sampling trajectory. Let $p_0(x)$ denote the data distribution and $p_t(x)$ the noisy marginal at time t . The reverse-time sampling process is defined by the SDE [33]:

$$dx = [f(x, t) - g(t)^2 \nabla_x \log p_t(x)] dt + g(t) dw, \quad (2)$$

where f and g are known functions determined by the forward process. Now consider the tempered distribution $p_0^{(\gamma)}$ obtained by raising p_0 to the power γ (and renormalizing), as defined in Eq. (1). Its corresponding unnormalized noisy marginal at time t is

$$p_t^{(\gamma)}(x) := \int p_0(x_0)^\gamma \mathcal{N}(x; x_0, \sigma_t^2 I) dx_0, \quad (3)$$

while the original noisy marginal is

$$p_t(x) = \int p_0(x_0) \mathcal{N}(x; x_0, \sigma_t^2 I) dx_0. \quad (4)$$

One might instead apply tempering *after* noising, by taking the (unnormalized) power

$$p_t(x)^\gamma = \left(\int p_0(x_0) \mathcal{N}(x; x_0, \sigma_t^2 I) dx_0 \right)^\gamma, \quad (5)$$

which we call *direct temperature scaling*. In this case the score is obtained by simple scaling:

$$\nabla_x \log p_t(x)^\gamma = \gamma \nabla_x \log p_t(x). \quad (6)$$

However, direct temperature scaling is generally *not* equivalent to noising the tempered data distribution, by comparing Eq. (3) and Eq. (5). Thus, the score of $p_t^{(\gamma)}(x)$ cannot be obtained by simply scaling the original score:

$$\nabla_x \log p_t^{(\gamma)}(x) \neq \gamma \nabla_x \log p_t(x). \quad (7)$$

Intuitively, $p_t^{(\gamma)}(x)$ reweights p_0 *before* Gaussian convolution, whereas $p_t(x)^\gamma$ applies the reweighting *after* noising. To further study this mismatch, we look at a simple example where p_0 is a multimodal distribution with well-separated modes, and we consider times t for which σ_t is small enough that the modes remain separated. In this regime,

$$p_t(x) \approx \sum_k \pi_k \mathcal{N}(x; \mu_k, \sigma_t^2 I), \quad (8)$$

where π_k is the probability mass of the k -th mode under p_0 and μ_k is the corresponding mean. Under this approximation, we can write down a closed-form expression for the unnormalized $p_t^{(\gamma)}$:

$$p_t^{(\gamma)}(x) = \sum_k \pi_k^\gamma \mathcal{N}(x; \mu_k, \sigma_t^2 I). \quad (9)$$

In contrast, direct tempering of p_t gives

$$p_t(x)^\gamma = \left(\sum_k \pi_k \mathcal{N}(x; \mu_k, \sigma_t^2 I) \right)^\gamma. \quad (10)$$

With the well-separated assumption, cross terms are negligible (see Sec. A), so

$$p_t(x)^\gamma \approx \sum_k \pi_k^\gamma \mathcal{N}(x; \mu_k, \sigma_t^2 I)^\gamma \propto \sum_k \pi_k^\gamma \mathcal{N}(x; \mu_k, \sigma_t^2 I / \gamma). \quad (11)$$

This motivates a simple time shift: choose $\tilde{t}$ such that $\sigma_{\tilde{t}}^2 = \gamma \sigma_t^2$. Then

$$p_{\tilde{t}}(x)^\gamma \approx \sum_k \pi_k^\gamma \mathcal{N}(x; \mu_k, \sigma_t^2 I). \quad (12)$$

Assuming we have access to a pre-trained neural network $\epsilon_\theta(x, t)$ that predicts the perturbation noise (ϵ -prediction), Tweedie’s formula [36] yields a score estimator $\nabla_x \log p_t(x) = -\epsilon_\theta(x, t) / \sigma_t^2$. We can then approximate the score of $p_t^{(\gamma)}$ by querying the network at the shifted time $\tilde{t}$ and applying temperature scaling:

$$\nabla_x \log p_t^{(\gamma)}(x) \approx \nabla_x \log (p_{\tilde{t}}(x)^\gamma) = \gamma \left(-\frac{\epsilon_\theta(x, \tilde{t})}{\sigma_{\tilde{t}}^2} \right), \quad (13)$$

which we refer to as *variance-corrective time shifting*. It effectively compensates the variance expansion while achieving the desired mode reweighting.

Although the derivation above is stated for variance-exploding diffusion, the same construction extends to other diffusion processes and samplers. We instantiate it for variance-preserving diffusion in Sec. B and for rectified-flow models in Sec. C.

3.2 Toy example

We validate our mode reweighting on a simple 2D Gaussian mixture,

$$p_0(x) = \sum_{k=1}^3 \pi_k \mathcal{N}(x; \mu_k, \sigma_{\text{data}}^2 I), \quad (14)$$

where the three centers $\{\mu_k\}_{k=1}^3$ are sampled once at random in $[-1.5, 1.5]^2$, the mixture weights are $\pi = (0.8, 0.1, 0.1)$ and σ_{data} is 0.01. We train a small ϵ -prediction model on samples from p_0 and generate samples using our temperature parameter γ .

To quantify mode reweighting, for each γ we draw $N = 10,000$ samples with DDPM [11], assign each sample to its nearest center, and report empirical reach probabilities $\hat{p}_k(\gamma)$ given by the resulting

Table 1: Toy example mode reweighting. For each γ we compare the target tempered mode weights $\pi^{(\gamma)}$ to empirical reach probabilities $\hat{p}(\gamma)$ and report $\text{KL}(\hat{p}(\gamma) \parallel \pi^{(\gamma)})$.

γ	mode 1		mode 2		mode 3		KL
	$\pi_1^{(\gamma)}$	$\hat{p}_1(\gamma)$	$\pi_2^{(\gamma)}$	$\hat{p}_2(\gamma)$	$\pi_3^{(\gamma)}$	$\hat{p}_3(\gamma)$	
0.1	0.381	0.464	0.309	0.227	0.309	0.309	0.021
0.5	0.586	0.640	0.207	0.167	0.207	0.192	0.007
1.0	0.800	0.794	0.100	0.091	0.100	0.115	0.001
5.0	1.000	0.999	0.000	0.001	0.000	0.001	0.003

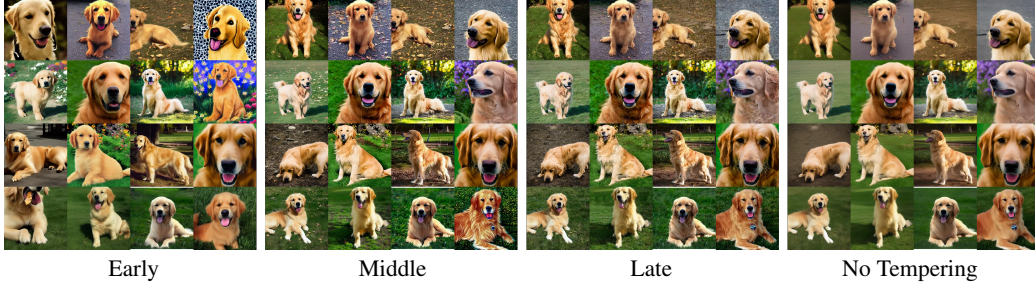


Figure 2: The effect of applying temperature scaling at different denoising stages.

frequencies. For reference, the ideal tempered mode weights are $\pi_k^{(\gamma)} := \pi_k^\gamma / \sum_j \pi_j^\gamma$. Table 1 summarizes the comparison between $\hat{p}(\gamma)$ and $\pi^{(\gamma)}$. It can be seen that with small γ , the empirical reach probabilities get closer to evenly distributed, while with large γ , the dominant modes are enhanced.

Although idealized, this toy setup captures a structural property shared by many real-world generative tasks. Natural data such as images or human motion can often be modeled as an unbalanced mixture of well-separated modes, where dominant categories or behaviors crowd out rarer ones. Under this approximation, the mode-reweighting effect demonstrated above predicts that lowering γ at sampling time will redistribute mass toward underrepresented modes, which is the mechanism we exploit to enhance diversity in the experiments that follow.

3.3 Different stages of the diffusion process

The reverse diffusion process is inherently coarse-to-fine: early denoising steps, operating at high noise levels, determine the global composition and mode identity of the generated sample, while later steps progressively refine structural details and texture. This suggests that the effect of tempering depends not just on its strength but on *when* along the trajectory it is applied. To explore this empirically, we relax the constant- γ setting used so far and replace γ with a sampling-time schedule $\gamma(t)$, applied step-by-step using the same time-shifting and score-scaling rule. We treat $\gamma(t)$ as a practical knob that lets us control which portion of the denoising trajectory is reweighted.

To study how each stage contributes, we partition the T sampling steps into three equal segments: Early ($T - \frac{2}{3}T$), Middle ($\frac{2}{3}T - \frac{1}{3}T$) and Late (steps $\frac{1}{3}T - 0$), where step T is the noisiest. We choose $\gamma(t) < 1$ inside one segment at a time while keeping $\gamma(t) = 1$ elsewhere.

We find the effect of temperature scaling to be qualitatively distinct across the three stages, as shown in Fig. 2. The *Early* segment is where noise is high and the score encodes global structure. Applying $\gamma(t) < 1$ here steers sampling particles towards different attractors, increasing overall compositional diversity and enabling the sampling of different modes. Applying it during the *Middle* segment preserves the global composition and mode identity but introduces variation in mid-level structural details. Applying it during the *Late* segment, when the score refines local appearance, yields samples with identical overall structure but differing in fine-grained details.

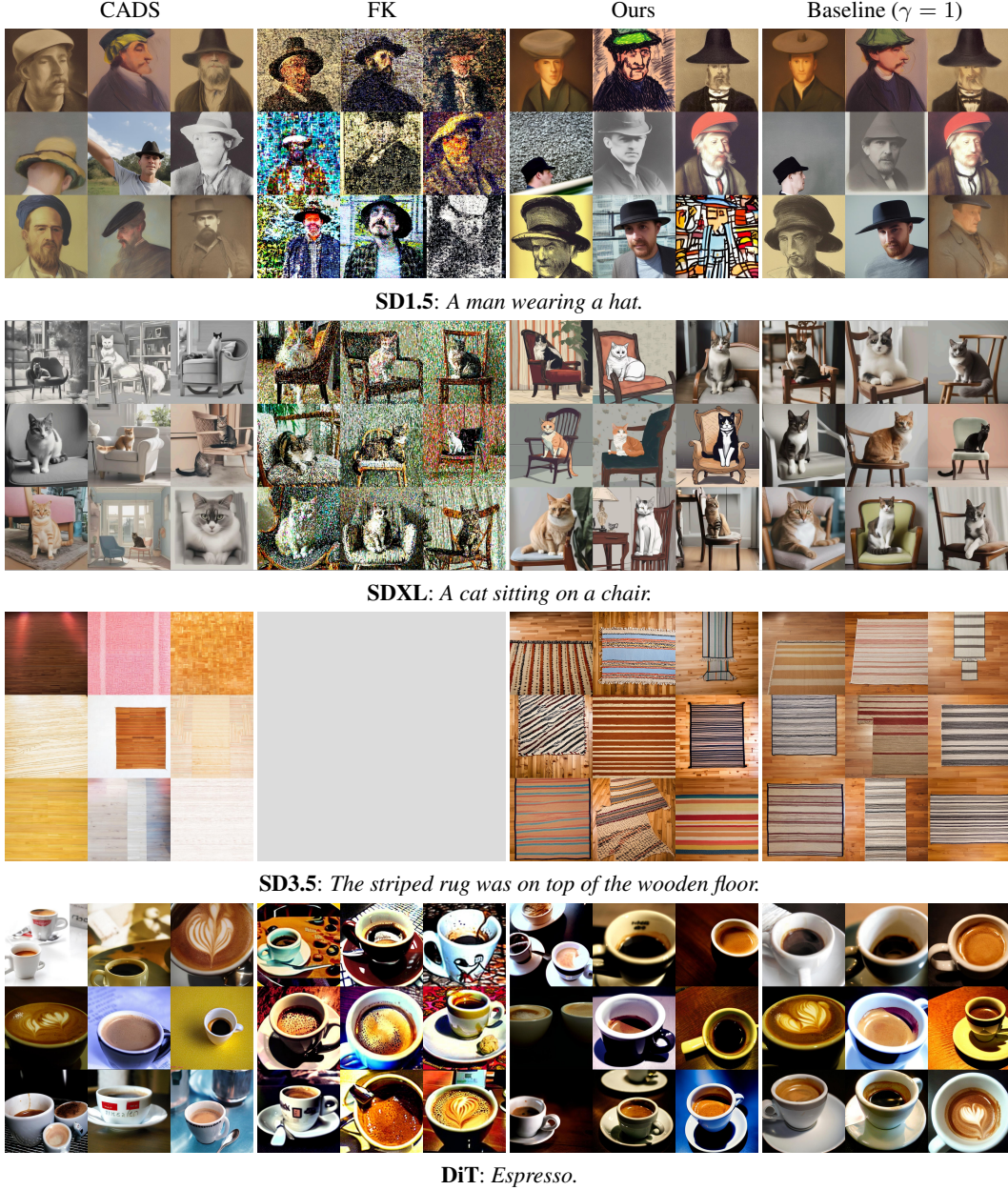


Figure 3: Qualitative comparison across model architectures. FK is not compatible with flow-matching models and is omitted for SD3.5.

4 Experiments

4.1 Fidelity and diversity metrics

The CLIP Maximum Mean Discrepancy (CMMD) [14] is the modern standard for perceptual quality of generative models, as it uses richer CLIP embeddings and an unbiased kernel MMD estimator, making it more effective at detecting semantic distribution differences and substantially more computationally efficient than FID [8]. CMMD rewards a generator for matching the reference data distribution, whereas our goal is to deliberately deviate from the training distribution by sampling from a tempered version $p^{(\gamma)}$ for diversity. As a result, CMMD penalizes our method for doing exactly what we aim for. We report CMMD for completeness and only as a partial reference for the quality of generated samples. For text-conditioned models, we report CMMD on the COCO [16]

Table 2: Quantitative comparison of diversity-enhancement methods. CMMD [14] is reported for reference (see Sec. 4.1). Error bars are $\pm 2\sigma$. Best and second-best entries per model group are **bold** and underlined.

Model	Method	CMMD↓	Vendi↑	Fidelity↑	Model	Method	CMMD↓	Vendi↑	Fidelity↑
DiT $\gamma = 0.20$	Baseline	0.550	13.45 ± 0.73	0.879 $\pm$ 0.062	SDXL $\gamma = 0.94$	Baseline	0.755	12.70 ± 0.68	0.903 ± 0.035
	CADS	0.753	15.03 $\pm$ 0.34	0.745 ± 0.094		CADS	0.899	13.89 $\pm$ 0.46	0.810 ± 0.063
	FK	3.138	13.15 ± 0.99	0.745 ± 0.116		FK	1.341	12.94 ± 0.71	0.894 ± 0.039
	Ours	<u>0.609</u>	13.86 ± 0.74	<u>0.859 $\pm$ 0.063</u>		Ours	0.719	<u>13.12 $\pm$ 0.65</u>	0.904 $\pm$ 0.032
SD 1.5 $\gamma = 0.88$	Baseline	0.546	14.13 ± 0.50	0.819 ± 0.058	SD 3.5 $\gamma = 0.78$	Baseline	<u>0.604</u>	<u>12.21 $\pm$ 0.86</u>	0.924 $\pm$ 0.036
	CADS	0.638	14.80 $\pm$ 0.29	<u>0.820 $\pm$ 0.059</u>		CADS	1.427	8.29 ± 0.23	0.398 ± 0.059
	FK	1.249	13.89 ± 0.58	0.817 ± 0.062		FK	N/A	N/A	N/A
	Ours	<u>0.566</u>	<u>14.27 $\pm$ 0.49</u>	0.824 $\pm$ 0.056		Ours	0.571	12.43 $\pm$ 0.86	<u>0.897 $\pm$ 0.056</u>

validation set. For class-conditional models, we report CMMD on the ImageNet [3] validation set. We use 2,000 samples for all CMMD calculations.

We report two complementary metrics that decouple condition fidelity from sample diversity, so the trade-off curve induced by varying γ can be read off directly.

Condition fidelity. For text-conditioned image generation we use the VQA score [17], which scores prompt alignment by querying a vision–language model with the prompt rephrased as a yes/no question and reading out the probability of the affirmative answer. The VQA score has been shown to correlate well with human judgment of prompt–image alignment, in particular for compositional prompts where CLIP-based scores [25] often saturate. For class-conditioned image generation we report the top-1 accuracy of a pretrained classifier on the generated samples.

Sample diversity. Following CADs [27], we quantify the diversity of a set of samples drawn from the same condition using the Vendi score [6] (full definition in Sec. D). Intuitively, the Vendi score reports the effective number of unique elements in a set: a score of 1 means all samples are identical, while a score of n (the size of the set) means the samples are maximally diverse under the chosen similarity kernel. Unlike recall- or coverage-based diversity measures, it does not require a reference distribution, which is appropriate here because the reference is exactly the distribution we are deviating from.

4.2 Results and comparisons

Image generation. In this section we compare our method with others, including CADs [27] and Feynman–Kac [30], on Stable Diffusion 1.5, Stable Diffusion XL, Stable Diffusion 3.5 and DiT. For the same model, all generations are conditioned on the same prompt, γ and seed where applicable. Since we are interested in mode diversity that is controlled in the early stage of reverse diffusion, we apply the tempering only in the early stage for all methods throughout Sec. 4.2.

As shown in Fig. 3, our method consistently improves sample diversity across all models while preserving prompt fidelity. CADs degrades noticeably on SDXL and fails catastrophically on SD3.5, whereas Feynman–Kac injects excessive noise throughout generation and yields unusable samples. The gain on DiT is more modest, since the model is less expressive than the Stable Diffusion family and therefore offers less headroom for diversity gains without compromising fidelity.

Quantitatively, our method attains the best CMMD and fidelity scores on all models and ranks first or second on Vendi, as reported in Tab. 2. CADs trades conditional fidelity for higher Vendi score and breaks down on SD3.5, whereas our method lifts diversity without sacrificing alignment. The qualitative comparison in Fig. 3 further reveals greater stylistic variety in our outputs, an axis that the Vendi score may not fully reflect.

Motion generation. Reliable metrics for diversity and quality in human motion generation [35] are still being actively explored, so we opt to provide qualitative *video* results in the supplementary material. We also include a static image in Fig. 8. The diversity gain is more apparent in video format, so we encourage readers to refer to the supplementary video for a full comparison. Our method produces more diverse motion patterns while maintaining the same overall pose and trajectory, whereas CADs produces motion that diverges from the text prompt, similar to the case of SD3.5 in image generation.

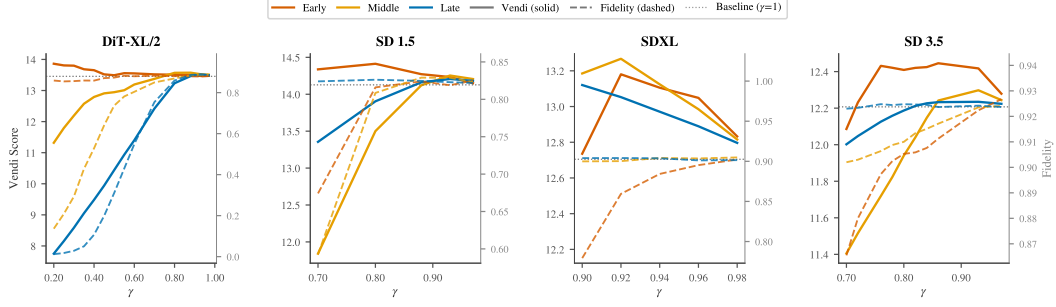


Figure 4: The effect of γ on diversity (Vendi, solid) and fidelity (dashed) for each model, averaged across all condition tiers.

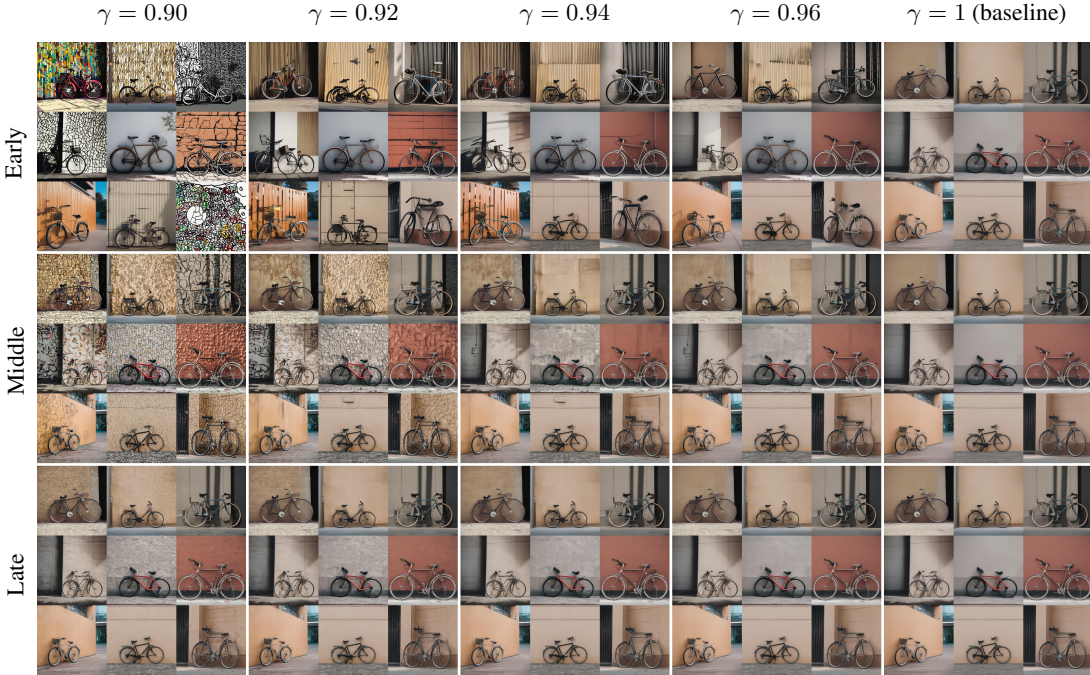


Figure 5: Qualitative effect of γ across denoising stages. Prompt: *A bicycle leaning against a wall* (SDXL).

4.3 Ablation study

Choice of γ and stage. We sweep γ over a range of values for each model and compare three denoising stages (early, mid and late), and different conditioning text and class tiers (loose, medium, tight), depending on the diversity headroom. Full setup and conditioning sets are given in Sec. E.

We observe a consistent pattern as γ varies. At small γ , samples are dominated by noise artifacts and appear unnatural, as illustrated in the upper-left corner of Fig. 5, which translates into low Vendi and fidelity scores. As γ grows, samples become more natural and the Vendi score rises, peaking before falling back toward the baseline as γ approaches 1. Fidelity, in contrast, increases roughly monotonically with γ . A second observation is that tempering at the early stage typically produces the largest diversity gains, with SDXL being the sole exception. Guided by Fig. 4, we select for each model the smallest γ that yields a substantial gain in Vendi without a meaningful drop in fidelity, giving $\gamma = 0.20$ for DiT, $\gamma = 0.88$ for SD1.5, $\gamma = 0.94$ for SDXL and $\gamma = 0.78$ for SD3.5. These are the values used in Sec. 4.2. A detailed per-tier breakdown is given in Fig. 7.

Variance-corrective time shifting. Here, we make a comparison between our method without time shifting (i.e., direct tempering) and TSR [37]. TSR is effectively a direct tempering method



Figure 6: Effect of variance-corrective time shifting on DDIM across tempering methods and γ values. Prompt: *A horse in a field* (SDXL).

with a smooth $\gamma(t)$ where γ smoothly transits from 1 to the supplied value as the diffusion process progresses from noisy to clean. As shown in Fig. 6, TSR introduces increasingly more local details and eventually becomes noisy as γ decreases. Without time shifting, small γ values introduce excessive noise and little diversity due to the variance expansion problem. In contrast, our method with time shifting can maintain good sample quality while achieving the same level of diversity as direct tempering.

5 Discussion

We have presented variance-corrective time shifting, a training-free method that resolves the variance expansion of naive score scaling and turns temperature sampling into a practical diversity knob for pretrained diffusion and flow-matching models. Studying the effect of γ at different stages of the diffusion process across DiT, Stable Diffusion and Motion Diffusion models on a comprehensive set of conditions, we observe consistent gains in diversity over existing methods at minimal cost to sample quality and condition fidelity.

Our method is bounded by the capacity of the underlying model and cannot overcome its inherent limitations, see, for example, the failure case of a five-legged horse in the bottom-right of the mosaic in Fig. 6. Moreover, temperature scaling with γ is not a free lunch: it trades off fidelity against diversity, and the right balance has to be tuned per model. Our schedule for $\gamma(t)$ is a three-stage piecewise function obtained by grid search, and a more systematic understanding of how the effect of γ varies along the sampling trajectory together with principled ways to design better schedules are natural directions for future work. So is extending the theoretical analysis beyond the toy Gaussian and the well-separated modes assumption to time-varying $\gamma(t)$ and more complicated settings. We are also curious whether the same construction transfers to products of two guidance signals, whether the picture can be connected to entropy and information dynamics [34] and whether the way γ reshapes

generation may in turn reveal something about the structure of the dataset, the conditioning signal and the landscape of the learned distribution.

References

- [1] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 843–852, 2023.
- [2] Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-I.I.D. diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [5] 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.
- [6] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.
- [7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [12] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
- [13] 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.
- [14] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9307–9315, 2024.
- [15] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.
- [16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

- [17] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024.
- [18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [19] 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.
- [20] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.
- [21] Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. *arXiv preprint arXiv:2410.14171*, 2024.
- [22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- [23] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022.
- [24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [25] 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*, pages 8748–8763. PmLR, 2021.
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [27] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADs: unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.
- [28] Vikash Sehwal, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [29] Andy Shih, Dorsa Sadigh, and Stefano Ermon. Long horizon temperature scaling. In *International conference on machine learning*, pages 31422–31434. PMLR, 2023.
- [30] Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alán Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. *arXiv preprint arXiv:2503.02819*, 2025.
- [31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.
- [32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [33] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

- [34] Dejan Stancevic and Luca Ambrogioni. The information dynamics of generative diffusion. *arXiv preprint arXiv:2508.19897*, 2025.
- [35] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.
- [36] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.
- [37] Yanbo Xu, Yu Wu, Sungjae Park, Zhizhuo Zhou, and Shubham Tulsiani. Temporal score rescaling for temperature sampling in diffusion and flow models. *arXiv preprint arXiv:2510.01184*, 2025.

A Cross-term negligibility for Eq. (11)

The mixture approximation in Sec. 3.1 relies on the claim that, under the well-separated mode assumption, the cross terms in $p_t(x)^\gamma$ are negligible for all real $\gamma > 0$, not just integer values. We make this precise here.

Fix a point x close to center μ_k . Write the mixture as

$$p_t(x) = f_k(x)(1 + \delta_k(x)), \quad f_k(x) := \pi_k \mathcal{N}(x; \mu_k, \sigma_t^2 I), \quad (15)$$

where $\delta_k(x)$ collects the contributions from all other components:

$$\delta_k(x) := \sum_{j \neq k} \frac{\pi_j \mathcal{N}(x; \mu_j, \sigma_t^2 I)}{\pi_k \mathcal{N}(x; \mu_k, \sigma_t^2 I)}. \quad (16)$$

Expanding $\|x - \mu_j\|^2$ around μ_k , the per-component Gaussian ratio reads

$$\frac{\mathcal{N}(x; \mu_j, \sigma_t^2 I)}{\mathcal{N}(x; \mu_k, \sigma_t^2 I)} = \exp\left(-\frac{\|\mu_k - \mu_j\|^2 + 2(x - \mu_k)^\top(\mu_k - \mu_j)}{2\sigma_t^2}\right). \quad (17)$$

Restricting to the support of mode k , namely $\|x - \mu_k\| \leq r$ for some r on the order of σ_t , the Cauchy–Schwarz inequality bounds the linear term by $|2(x - \mu_k)^\top(\mu_k - \mu_j)| \leq 2r \|\mu_k - \mu_j\|$, so Eq. (17) yields the uniform estimate

$$\frac{\mathcal{N}(x; \mu_j, \sigma_t^2 I)}{\mathcal{N}(x; \mu_k, \sigma_t^2 I)} \leq \exp\left(-\frac{\|\mu_k - \mu_j\|(\|\mu_k - \mu_j\| - 2r)}{2\sigma_t^2}\right). \quad (18)$$

Under the well-separated assumption $\|\mu_k - \mu_j\| \gg \max(\sigma_t, 2r)$, every summand of $\delta_k(x)$ is exponentially small, so $\delta_k(x) \ll 1$ uniformly over the support of mode k .

For any real $\gamma > 0$ and $|\delta| < 1$, the generalized binomial series gives

$$(1 + \delta)^\gamma = \sum_{n=0}^{\infty} \binom{\gamma}{n} \delta^n = 1 + \gamma \delta + \frac{\gamma(\gamma-1)}{2} \delta^2 + \dots \quad (19)$$

Applying Eq. (19) to $p_t(x)^\gamma = f_k(x)^\gamma(1 + \delta_k)^\gamma$ gives

$$p_t(x)^\gamma = f_k(x)^\gamma [1 + \gamma \delta_k + O(\delta_k^2)]. \quad (20)$$

Because $\delta_k = O(\exp(-\|\mu_k - \mu_j\|^2/2\sigma_t^2))$ is exponentially small, all correction terms are negligible and

$$p_t(x)^\gamma \approx f_k(x)^\gamma = \pi_k^\gamma \mathcal{N}(x; \mu_k, \sigma_t^2 I)^\gamma. \quad (21)$$

Since the mode supports are disjoint under the well-separated assumption, summing the local approximations gives

$$p_t(x)^\gamma \approx \sum_k \pi_k^\gamma \mathcal{N}(x; \mu_k, \sigma_t^2 I)^\gamma. \quad (22)$$

This holds for all real $\gamma > 0$. The further proportionality

$$\mathcal{N}(x; \mu_k, \sigma_t^2 I)^\gamma \propto \exp\left(-\frac{\gamma \|x - \mu_k\|^2}{2\sigma_t^2}\right) \propto \mathcal{N}\left(x; \mu_k, \frac{\sigma_t^2}{\gamma} I\right) \quad (23)$$

follows by inspection of the Gaussian density.

B Temperature sampling for variance-preserving diffusion models

The main text motivates variance-corrective time shifting for temperature sampling. This appendix focuses on the VP case and, in particular, how to implement the time shift with an ϵ -predicting network.

B.1 Key identities

We use the standard VP parameterization

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (24)$$

so that an ϵ -predictor yields the score estimate

$$\nabla_x \log p_t(x) \approx -\frac{\epsilon_\theta(x, t)}{\sqrt{1 - \bar{\alpha}_t}}. \quad (25)$$

It is convenient to work with the VE-equivalent noise scale

$$\sigma_t^2 := \frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}, \quad (26)$$

which is monotone in t .

Time shifting defines a shifted time $\tilde{t}$ by matching the effective noise level in the normalized variable $u_t = x_t / \sqrt{\bar{\alpha}_t}$, i.e.

$$\sigma_{\tilde{t}}^2 = \gamma \sigma_t^2 \quad (\text{equivalently, } \sigma_{\tilde{t}} = \sqrt{\gamma} \sigma_t), \quad (27)$$

or, in terms of $\bar{\alpha}_t$,

$$\bar{\alpha}_{\tilde{t}} = \frac{\bar{\alpha}_t}{\bar{\alpha}_t + \gamma(1 - \bar{\alpha}_t)}. \quad (28)$$

B.2 Practical score implementation

Assume access to a pre-trained VP noise predictor $\epsilon_\theta(x_t, t)$ (standard ϵ -prediction, $\epsilon \sim \mathcal{N}(0, I)$). The temperature-scaled score follows directly from the VE result (13) by working with the normalized variable $u_t := x_t / \sqrt{\bar{\alpha}_t}$, in which the VP forward kernel reduces to a pure VE kernel $\mathcal{N}(x_0, \sigma_t^2 I)$ with $\sigma_t^2 = (1 - \bar{\alpha}_t) / \bar{\alpha}_t$. Writing p_t^u for the marginal of u_t and $p_t^{u,(\gamma)}$ for its tempered counterpart, the argument of Eq. (13) transports verbatim into u -space,

$$\nabla_u \log p_t^{u,(\gamma)}(u) \approx \gamma \cdot \nabla_u \log p_t^u(u), \quad (29)$$

with the shifted timestep $\tilde{t}$ chosen by nearest-neighbor lookup on the (monotone) schedule $\{\sigma_t\}_t$ to match the target noise level $\sigma_{\tilde{t}} = \sqrt{\gamma} \sigma_t$,

$$\tilde{t} \in \arg \min_{t'} |\sigma_{t'} - \sqrt{\gamma} \sigma_t|. \quad (30)$$

Since the VP network takes x (not u) as input, switching from t to $\tilde{t}$ while keeping u fixed requires rescaling the input,

$$\tilde{x} := \sqrt{\bar{\alpha}_{\tilde{t}}} u = \sqrt{\frac{\bar{\alpha}_{\tilde{t}}}{\bar{\alpha}_t}} x, \quad (31)$$

so that the network call resolves to $\epsilon_\theta(\tilde{x}, \tilde{t})$. Converting the u -space score (29) back to x -space at the original time t via $x = \sqrt{\bar{\alpha}_t} u$ contributes a Jacobian $1 / \sqrt{\bar{\alpha}_t}$, which yields the VE-parallel form

$$\nabla_x \log p_t^{(\gamma)}(x) \approx \gamma \cdot \left(-\frac{\epsilon_\theta(\tilde{x}, \tilde{t})}{\sqrt{\bar{\alpha}_t} \sigma_{\tilde{t}}} \right). \quad (32)$$

At $\gamma = 1$, $\sqrt{\bar{\alpha}_t} \sigma_t = \sqrt{1 - \bar{\alpha}_t}$ and the bracketed expression reduces to the standard VP score conversion $-\epsilon_\theta(x, t) / \sqrt{1 - \bar{\alpha}_t}$, as expected.

The two variance factors collapse algebraically. Using $\sigma_{\tilde{t}}^2 = \gamma(1 - \bar{\alpha}_t) / \bar{\alpha}_t$,

$$\sqrt{\bar{\alpha}_t} \sigma_{\tilde{t}} = \sqrt{\bar{\alpha}_t \cdot \gamma(1 - \bar{\alpha}_t) / \bar{\alpha}_t} = \sqrt{\gamma} \sqrt{1 - \bar{\alpha}_t}, \quad (33)$$

so one factor of $\sqrt{\gamma}$ in the numerator of Eq. (32) cancels with the variance factor in the denominator and the expression simplifies to

$$\nabla_x \log p_t^{(\gamma)}(x) \approx -\frac{\sqrt{\gamma} \epsilon_\theta(\tilde{x}, \tilde{t})}{\sqrt{1 - \bar{\alpha}_t}}. \quad (34)$$

The compact form (34) matches the standard VP score conversion at the original time t with the network query relocated to $(\tilde{x}, \tilde{t})$ and a single $\sqrt{\gamma}$ scalar applied to the output. Operationally, this can be implemented as a wrapper that returns

$$\epsilon_{\text{temp}}(x, t) := \sqrt{\gamma} \epsilon_\theta(\tilde{x}, \tilde{t}) \quad (35)$$

in place of $\epsilon_\theta(x, t)$, leaving the VP sampler and time schedule unchanged at the original timestep t .

C Temperature sampling for flow-matching models

Modern open-weight text-to-image backbones such as Stable Diffusion 3 [5] are trained as flow-matching models [18, 19] rather than VE/VP diffusion. Variance-corrective time shifting still applies in this setting with no retraining, because a rectified-flow forward kernel is just another affine Gaussian and the analysis of Eq. (13) carries through after a change of variables.

C.1 Key identities

We use the rectified-flow parameterization

$$x_t = (1 - t)x_0 + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad t \in [0, 1], \quad (36)$$

so $t = 0$ is data and $t = 1$ is noise. The conditional kernel $p(x_t | x_0) = \mathcal{N}(x_t; (1 - t)x_0, t^2 I)$ is affine with signal scale $a_t = 1 - t$ and noise scale $b_t = t$. The model predicts the velocity

$$v_\theta(x_t, t) \approx \mathbb{E}[\epsilon - x_0 | x_t], \quad (37)$$

from which the noise prediction and the score follow algebraically:

$$\epsilon_\theta(x, t) = (1 - t)v_\theta(x, t) + x, \quad (38)$$

$$\nabla_x \log p_t(x) \approx -\frac{\epsilon_\theta(x, t)}{t} = -\frac{(1 - t)v_\theta(x, t) + x}{t}. \quad (39)$$

Defining the normalized variable $u_t := x_t/a_t = x_t/(1 - t)$ reduces the conditional to a pure VE form, $u_t | x_0 \sim \mathcal{N}(x_0, \sigma_e(t)^2 I)$, with effective noise scale

$$\sigma_e(t) := \frac{b_t}{a_t} = \frac{t}{1 - t}. \quad (40)$$

C.2 Time-shift for rectified flow

By the same argument as Eq. (13), the tempered marginal in u -space is approximated by direct tempering of p_t^u provided $\sigma_e(\tilde{t})^2 = \gamma \sigma_e(t)^2$. For the rectified-flow schedule this gives

$$\frac{\tilde{t}}{1 - \tilde{t}} = \sqrt{\gamma} \frac{t}{1 - t} \iff \tilde{t} = \frac{\sqrt{\gamma} t}{1 - (1 - \sqrt{\gamma}) t}. \quad (41)$$

For $\gamma < 1$ this yields $\tilde{t} < t$, i.e. a query at a slightly less noisy time. The corresponding input rescaling that keeps the normalized variable u fixed across the time shift is

$$\tilde{x} := \frac{a_{\tilde{t}}}{a_t} x = \frac{1 - \tilde{t}}{1 - t} x = \frac{x}{1 - (1 - \sqrt{\gamma}) t}. \quad (42)$$

C.3 Practical velocity implementation

As in the VP case, the temperature-scaled score is most transparently derived in u -space, where the rectified-flow kernel reduces to VE with noise scale $\sigma_e(t) = t/(1 - t)$. Applying Eq. (13) at the shifted noise level $\sigma_e(\tilde{t}) = \sqrt{\gamma} \sigma_e(t)$ gives

$$\nabla_u \log p_t^{u, (\gamma)}(u) \approx \gamma \cdot \nabla_u \log p_t^u(u), \quad (43)$$

where p_t^u is the marginal of the normalized variable $u_t = x_t/(1-t)$ introduced in the previous subsection, and $p_t^{u,(\gamma)}$ is its tempered counterpart. The input rescaling (42) ensures the network call resolves to $\epsilon_\theta(\tilde{x}, \tilde{t})$, and converting the score back to x -space at the original time t via $x = (1-t)u$ contributes a Jacobian $1/(1-t)$, yielding the VE-parallel form

$$\nabla_x \log p_t^{(\gamma)}(x) \approx \gamma \cdot \left(-\frac{\epsilon_\theta(\tilde{x}, \tilde{t})}{(1-t)\sigma_e(\tilde{t})} \right). \quad (44)$$

Using $\sigma_e(\tilde{t}) = \sqrt{\gamma}t/(1-t)$, the variance factor collapses,

$$(1-t)\sigma_e(\tilde{t}) = \sqrt{\gamma}t, \quad (45)$$

so one $\sqrt{\gamma}$ cancels and Eq. (44) simplifies to $-\sqrt{\gamma}\epsilon_\theta(\tilde{x}, \tilde{t})/t$. The corresponding tempered noise prediction at the original time is therefore

$$\epsilon_{\text{temp}}(x, t) := \sqrt{\gamma}\epsilon_\theta(\tilde{x}, \tilde{t}), \quad (46)$$

identical in structure to the VP case. Composing with Eq. (38),

$$\epsilon_{\text{temp}}(x, t) = \sqrt{\gamma}[(1-\tilde{t})v_\theta(\tilde{x}, \tilde{t}) + \tilde{x}], \quad (47)$$

and converting back to a velocity at the original time t via $v = (\epsilon - x)/(1-t)$ yields the tempered velocity to plug into the unmodified flow-matching sampler:

$$v_{\text{temp}}(x, t) = \frac{\sqrt{\gamma}[(1-\tilde{t})v_\theta(\tilde{x}, \tilde{t}) + \tilde{x}] - x}{1-t} \quad (48)$$

with $\tilde{t}$ and $\tilde{x}$ given by Eqs. (41) and (42). At $\gamma = 1$ we recover $\tilde{t} = t$, $\tilde{x} = x$ and $v_{\text{temp}} = v_\theta$, as expected.

D Vendi score

The Vendi score [6] measures the effective number of distinct modes in a finite sample. Given n samples $x_1, \dots, x_n \in \mathcal{X}$ and a symmetric positive-semidefinite similarity kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ satisfying $k(x, x) = 1$, form the $n \times n$ Gram matrix K with $K_{ij} = k(x_i, x_j)$ and normalize it to $\hat{K} = K/n$. Let $\lambda_1, \dots, \lambda_n \geq 0$ be the eigenvalues of $\hat{K}$, which satisfy $\sum_i \lambda_i = 1$ by construction. The Vendi score is

$$\text{VS}(x_1, \dots, x_n) := \exp\left(-\sum_{i=1}^n \lambda_i \log \lambda_i\right) = \exp\left(H(\hat{K})\right), \quad (49)$$

where $H(\hat{K}) = -\text{tr}(\hat{K} \log \hat{K})$ is the von Neumann entropy of $\hat{K}$. The score equals 1 when all samples are identical (rank-one $\hat{K}$) and equals n when all samples are mutually orthogonal under k (identity $\hat{K}$), making it an interpretable effective count of distinct elements.

For image samples we follow CADs [27] and use cosine similarity on SSCD [23] features as the kernel: $k(x_i, x_j) = \phi(x_i)^\top \phi(x_j) / (\|\phi(x_i)\| \|\phi(x_j)\|)$, where ϕ is the SSCD encoder. SSCD features are designed to be invariant to photometric variation while remaining sensitive to semantic and compositional differences, making them well-suited for measuring diversity beyond low-level pixel statistics.

E Ablation setup and conditioning sets

We organize the ablations around a per-condition sweep. For each configuration we draw $K = 16$ samples per condition and report the Vendi score averaged over conditions within a prompt tier. Smaller γ values were excluded after preliminary runs showed unrecoverable loss of prompt fidelity. The diffusion trajectory is partitioned into three contiguous stages of equal timestep length (early, mid and late) and γ is applied within a single stage per run, with γ applied across the full trajectory included as an additional configuration. The untempered model at $\gamma = 1$ provides the baseline.

The following subsections enumerate the full conditioning sets. For text-conditioned models we use 40 prompts split across three tiers of decreasing diversity headroom: 15 single-concept prompts (L), 15 mildly descriptive prompts (M) and 10 compositional prompts with explicit attributes, counts or spatial relations drawn from T2I-CompBench [13]. For ImageNet class-conditional models we use 20 classes evenly split between a high-variance subset (scenes and broad concepts) and a canonical subset (objects with stereotyped appearance).

E.1 Text-to-image prompts

Tier L: single-concept prompts (15).

- a dog
- a landscape
- a portrait of a person
- a building
- a bowl of food
- a vehicle on a road
- a mountain
- an oil painting
- a still life
- a child playing
- a city street
- a flower
- a bird in flight
- a piece of furniture
- a dessert

Tier M: mildly descriptive prompts (15).

- a cat sitting on a chair
- a man wearing a hat
- a forest in autumn
- a car parked at night
- a bowl of soup on a table
- a woman reading a book
- a dog running on a beach
- a kitchen with morning light
- a bicycle leaning against a wall
- a horse in a field
- a coffee shop interior
- a sunset over water
- a chef preparing a meal
- a vase of flowers on a windowsill
- a bookshop with old books

Tier T: compositional prompts (10). Drawn from T2I-CompBench [13].

- the red hat was on top of the brown coat rack.
- the fluffy cat is on the left of the soft pillow.
- the rectangular picture frame was hung above the beige couch.
- the leather wallet was inside the brown purse.
- the black chair was on the left of the white table.
- the red apple was on top of the black plate.

- the striped rug was on top of the wooden floor.
- the silver laptop was on top of the black desk.
- the square box was next to the circular canister.
- the rough brick was on top of the smooth tile.

E.2 ImageNet classes

Subset H: high intra-class variance (10). Scenes and broad categories.

- seashore
- valley
- lakeside
- alp
- library
- restaurant
- jigsaw puzzle
- comic book
- bookshop
- toyshop

Subset C: canonical low-variance (10). Objects with stereotyped pose or appearance.

- great white shark
- hammerhead shark
- golf ball
- balance beam
- espresso
- pretzel
- soccer ball
- volleyball
- screw
- spotlight

Fig. 7 reports the per-tier breakdown of the γ sweep summarized in Fig. 4 of the main text. The trends from the averaged view carry over to each tier. Two additional observations emerge from the breakdown. First, at $\gamma = 1$ the loose tier attains higher diversity than the tight tier, confirming that the tier partition reflects a real gap in diversity headroom. Second, the diversity gain from our method is more pronounced on the tight tier than on the loose tier, consistent with the expectation that prompts with tighter constraints have more room to improve.

F Qualitative evaluation on Motion Diffusion models

As shown in Fig. 8, CADs noticeably degrades condition fidelity across all four prompts. Our method produces more diverse motions while remaining faithful to the text prompt. We encourage readers to view the accompanying video for a fuller comparison.

G Computational resources

We do not train any models for this paper. All experiments are conducted with off-the-shelf checkpoints that are publicly available and widely used in the community. We generate around 100k samples across all experiments used in the paper, which we estimate to require around 650 GPU-hours on an nVidia RTX 3090. The Vendi score is computed on CPU and takes around 10 hours for the full set of samples.

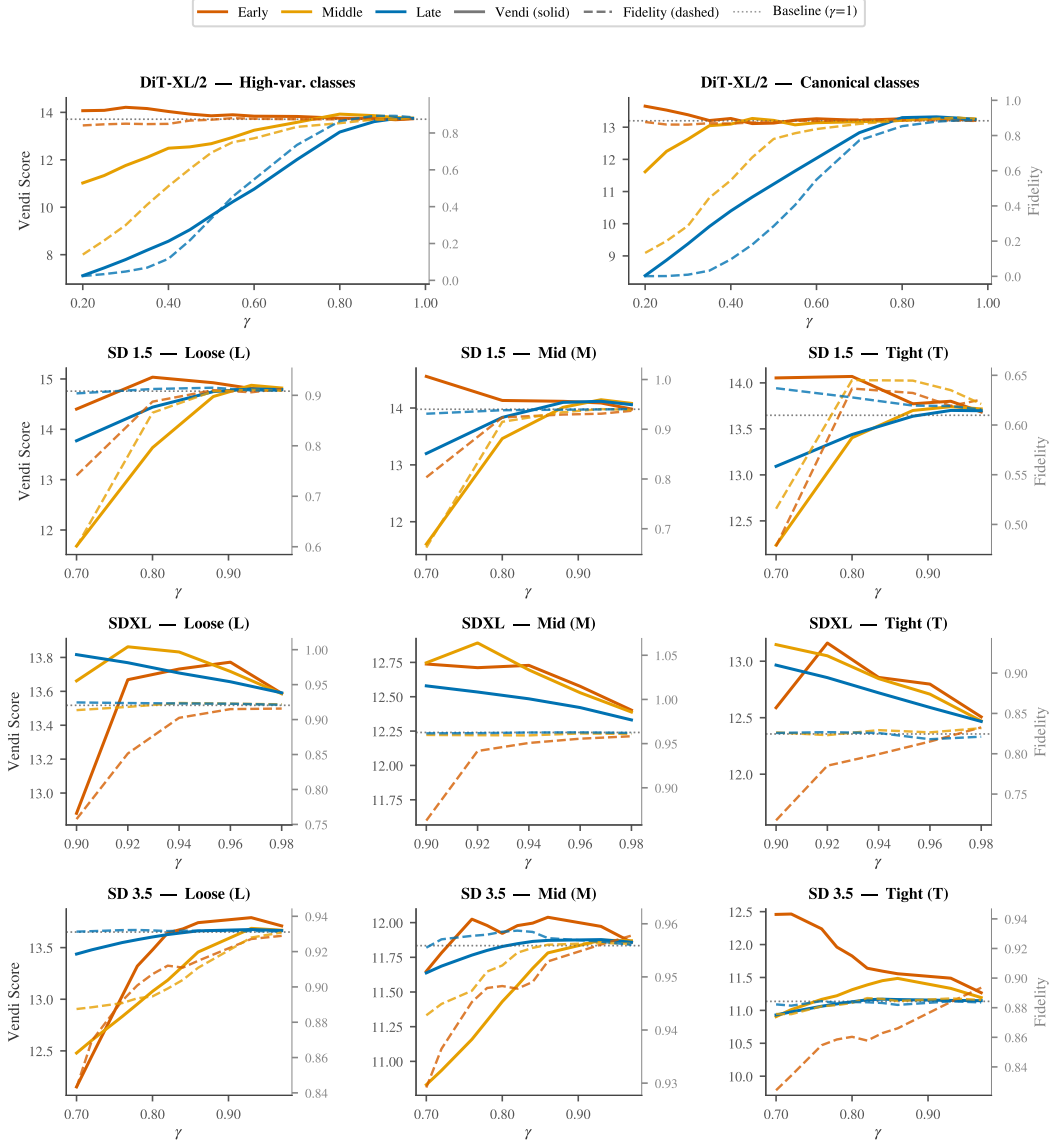


Figure 7: Per-tier breakdown of the γ sweep (full version of Fig. 4). Each row corresponds to one model and each column corresponds to one condition tier (prompt tier for text-conditioned models, class subset for ImageNet). Within each panel, γ is on the horizontal axis, and we overlay six curves: three denoising stages (early, middle, late) crossed with two metrics, Vendi (diversity, left axis) and VQA/Accuracy (fidelity, right axis). The dotted line marks the untempered baseline ($\gamma = 1$).

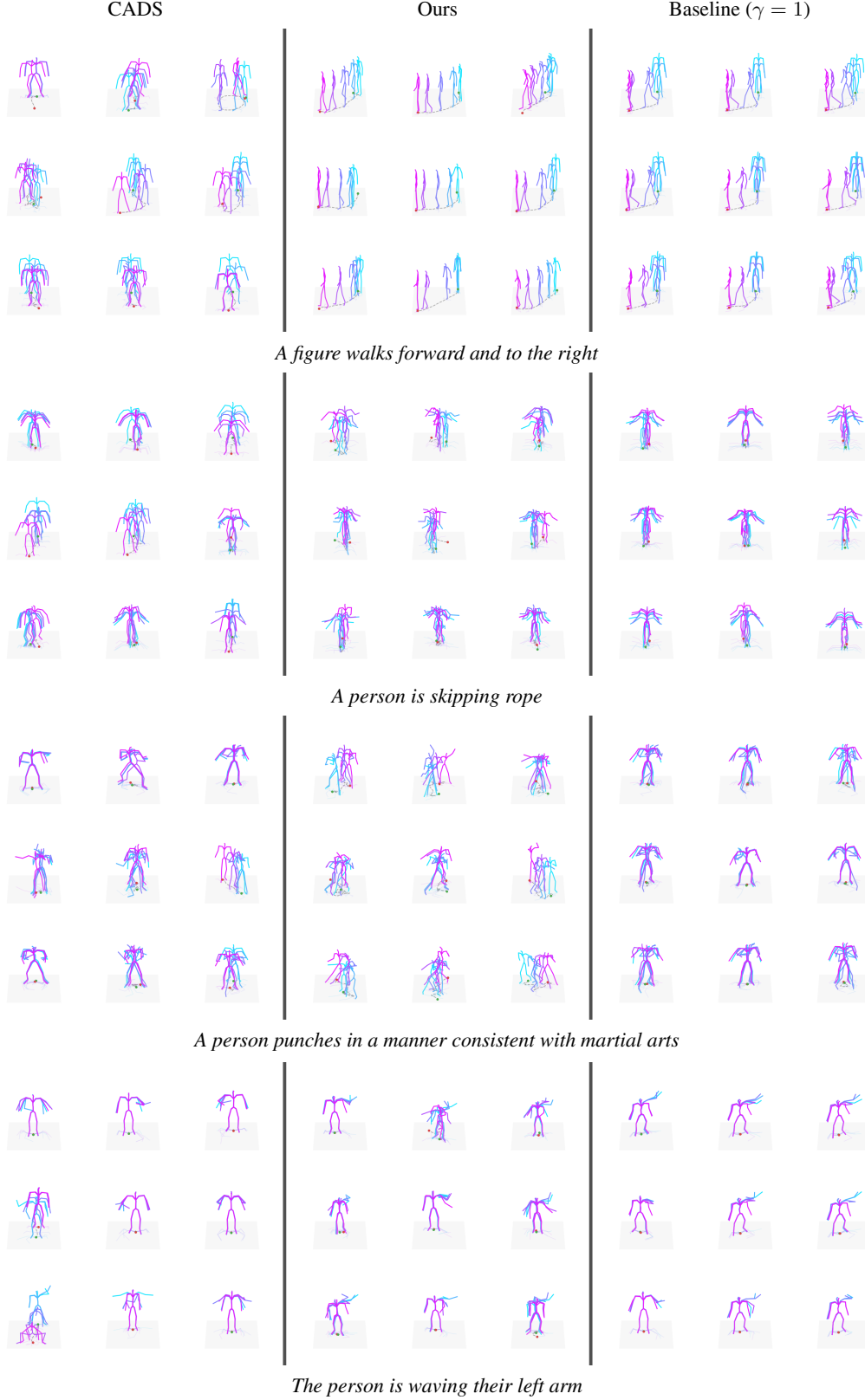


Figure 8: Qualitative comparison on the motion diffusion model [35]. Each row shows one text prompt, and each column shows samples from CADS, our method and the baseline ($\gamma = 1$). Five evenly spaced frames are rendered as skeletons, with color encoding time from cyan (earliest) to magenta (latest). A dashed floor trajectory traces the root joint (pelvis) path over the full sequence.