Colored Noise Diffusion Sampling

Hadar Davidson **Noam Issachar** **Sagie Benaim**

The Hebrew University of Jerusalem

Abstract

Diffusion models achieve state-of-the-art image synthesis, with their generative trajectories fundamentally exhibiting a spectral bias, resolving low-frequency global structures early and high-frequency fine details later. Conventional stochastic differential equation (SDE) solvers fail to account for this dynamic, naively injecting uniform white noise throughout the entire process and misusing the finite energy budget. In this work, we establish a mathematical framework that reconsiders SDE inference as a targeted, frequency-decoupled energy transfer. Leveraging this framework, we introduce Colored Noise Sampling (CNS), a novel, training-free stochastic solver. Rather than injecting uniform white noise, CNS utilizes a dynamic, timestep- and frequency-dependent schedule that more efficiently allocates injected energy toward structurally unresolved frequency bands. By actively exploiting the model's inherent spectral bias, CNS systematically steers the generated distribution toward the true data manifold. Extensive experiments demonstrate that CNS significantly outperforms standard ODE and SDE baselines as an inference-time sampler substitution across diverse architectures (SiT, JiT, FLUX). Compared to standard sampling on ImageNet-256, CNS achieves substantial unguided FID reductions, improving from 8.26 to 6.27 on SiT-XL/2, 32.39 to 26.69 on JiT-B/16, and 11.88 to 8.31 on JiT-H/16, while yielding consistent relative FID improvements with Classifier-Free Guidance. Project page is available at `https://hadardavidson.github.io/CNS/`.

1 Introduction

Diffusion models have established a new standard in photorealistic image synthesis, defining the state-of-the-art in high-fidelity generation [16, 62, 26]. Crucially, the sampling trajectory of these models exhibits a *spectral bias* [65, 20]. This inductive property dictates that diffusion models inherently resolve low-frequency global structures during early sampling steps, while filling in high-frequency fine details in later steps.

Current sampling algorithms [61, 62, 60] leave this structure unexploited. Standard stochastic methods based on Stochastic Differential Equations (SDEs) naively inject uniform white noise, completely disregarding how the frequency spectrum of the generated image dynamically evolves over time. To address this inefficiency at its root, we introduce a new class of stochastic solvers that actively leverages this spectral bias. By tailoring the injected noise to the specific denoising timestep, we improve generation fidelity without requiring any additional training.

Recognizing the importance of spectral bias, several recent works have attempted to leverage it. One line of research alters the training framework, interpolating from spectrally non-uniform or temporally evolving noise distributions [8, 58, 17]. Other approaches operate at inference time, introducing ad-hoc modifications such as frequency-decoupled operations [46, 71], internal activation reweighting [59], or step-size schedule adjustments [28]. While these methods yield measurable improvements, they remain fundamentally constrained by their underlying use of spectrally uniform solvers. This

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: **Colored Noise Sampling (CNS)**. Samples from SiT-XL/2 on ImageNet-256 (with CFG); all images within a triplet share the same seed. The insets highlight regions where the methods differ most: CNS resolves fine high-frequency detail more sharply than either baseline.

naturally leads to our guiding question: *How can we actively exploit the spectral bias of diffusion models to design a fundamentally new, general-purpose sampler that improves generation fidelity?*

To answer this, we first establish a mathematical framework to control the generated distribution via frequency-aware noise injection. Geometrically, sampling trajectories resemble non-orthogonal rotations toward the data manifold [65]. This implies that diffusion models do not arbitrarily discard initial noise; rather, a significant structural component of this signal is preserved and mapped into final image features [68, 37].

A key observation of our work is that this signal-preserving transfer also applies to the continuous noise injected by SDE solvers throughout the trajectory. Furthermore, this process is frequency-decoupled: injected noise in a specific frequency band maps directly to spatial features in that same band. By ensuring our frequency-aware adjustments remain strictly variance-preserving, requiring only that the total injected energy per step remains normalized, we demonstrate that the classic Langevin requirement for uniform white noise [42] can be safely relaxed without pushing intermediate states further from the training distribution than standard sampling already does.

Building on this framework, we construct a timestep- and frequency-dependent noise schedule. By analyzing the progression rates of different frequency bands during generation, we propose that the network's ability to convert injected noise into coherent image features significantly depends on how structurally "resolved" that specific band is at a given timestep. This insight allows us to reconsider SDE sampling as a targeted *energy injection* process. Rather than uniformly distributing the finite injected noise budget, our approach utilizes a dynamic schedule based on the expected evolution of the trajectory, allocating energy to the frequency bands where it is most needed. This principled allocation steers the output toward the true data manifold, yielding strictly higher-fidelity generation.

To validate our approach, we conduct extensive experiments across diverse architectures and modalities, including latent-space generation (SiT [36]), pixel-space generation (JiT [29]), and state-of-the-art text-to-image synthesis (FLUX [25, 26]). Evaluated primarily via the Fréchet Inception Distance (FID) [14], empirical results demonstrate that our method significantly outperforms standard ODE and SDE baselines. On ImageNet-256 [54], we achieve substantial FID reductions under both unguided and Classifier-Free Guidance (CFG) [15] settings, while maintaining robust stability across varying discretization steps. Fig. 1 compares samples across solvers at matched seeds, isolating where CNS improves against standard sampling. Furthermore, our sampler proves effective when integrated into text-to-image pipelines like FLUX, improving automatic human-preference scores.

To summarize, our main contributions are:

- We establish a mathematical framework that reframes SDE noise injection as a targeted energy transfer, and demonstrate that the standard Langevin requirement for spectrally uniform white noise can be safely relaxed to resolve spectral gaps.
- We introduce Colored Noise Sampling (CNS), a novel, training-free stochastic solver that actively leverages spectral bias by dynamically allocating injected noise energy toward structurally unresolved frequency bands.
- We validate CNS as a robust, general-purpose sampler across diverse architectures (SiT, JiT, FLUX). On ImageNet-256, CNS achieves substantial unguided FID reductions (e.g., 8.26 to 6.27 for SiT-XL/2, 32.39 to 26.69 for JiT-B/16) and relative CFG improvements ranging from $\sim$8% to $\sim$50% over standard ODE and SDE baselines.

2 Related Work

Samplers for Diffusion Models. Sampling in diffusion models is a highly researched domain primarily focused on numerically mitigating discretization errors. Prominent advancements include higher-order solvers [66, 69, 35] that maintain fidelity at low step counts, dynamic solver alternation [33, 76], and state reparameterizations that smooth integration pathways [74, 73, 5]. While these methods successfully reduce truncation errors and accelerate generation, they remain agnostic to the evolving spatial structure of the state. Our approach is fundamentally orthogonal: rather than strictly optimizing numerical precision, we optimize the allocation of stochastic energy by explicitly exploiting the model's spectral bias.

Leveraging Spectral Bias in Diffusion Models. A prominent line of research exploits the spectral bias of diffusion models during training by altering noise distributions. These methods rely on empirical heuristics to modify initial [58] or temporally-evolving noise distributions [17, 53], or introduce formally grounded frequency-dependent processes like EqualSNR [8]. However, fundamentally altering the learning objective demands costly model retraining. In sharp contrast, our approach overcomes this barrier by harnessing spectral bias exclusively at inference time, requiring no retraining but only a one-time calibration costing seconds to minutes (App. E). To circumvent retraining costs, a separate line of work leverages spectral bias via inference-only modifications. These methods introduce ad-hoc adjustments to the generation pipeline, such as applying frequency-decoupled operations to the predicted state [46, 71], dynamically reweighting internal network activations [59], adjusting step-size schedules [28], or coupling spectral bias with positional encodings [20]. While effective, these techniques treat the underlying stochastic solver as a static black box. Our work targets this unexplored component: rather than modifying the network or its outputs post-hoc, we directly embed the spectral bias into the core sampling mechanism itself. We compare against these methods directly in App. F.2.

3 Method

We now introduce our approach. Sec. 3.1 outlines standard diffusion background. Sec. 3.2 and 3.3 formalize the specific generative phenomena, inference-time spectral bias and noise energy preservation, that we leverage to build our framework. Building on these principles, Sec. 3.4 analyzes the spectral gap induced by standard SDEs. Finally, Sec. 3.5 details how CNS dynamically colors injected noise to actively steer the generated spectrum toward the true data manifold.

3.1 Background: Diffusion Models and Sampling Dynamics

Diffusion Models [16, 62] and Flow Matching [32, 34] can be unified under Stochastic Interpolants [3]. Given data $x_0 \sim p_{\text{data}}$ and a noise prior $\epsilon \sim \mathcal{N}(0, I)$, these models construct intermediate states x_t by:

$$x_t = \alpha_t x_0 + \sigma_t \epsilon \tag{1}$$

flowing from pure noise at $t = 1$ to clean data at $t = 0$. In order to reverse this process, a network is trained to approximate the velocity v_θ, the clean data x_θ, or the noise ϵ_θ. Because x_t is a simple affine combination, these parameterizations are algebraically equivalent and deterministically yield the marginal score $s_\theta \approx \nabla_{x_t} \log p_t(x_t)$ required for stochastic sampling. During inference, these learned predictions are substituted into reverse-time differential equations and integrated using deterministic or stochastic solvers (complete parameterization relations are detailed in App. A).

Sampling Dynamics. Deterministic sampling via Probability Flow ODEs integrates the predicted velocity ($dx_t = v_\theta dt$). While efficient, discrete numerical errors inevitably accumulate, causing the state to drift off the true data manifold [60]. Stochastic solvers address this drift by simulating the generative process as a reverse-time SDE. Introducing a diffusion coefficient $g(t) > 0$ and a reverse-time Wiener process $\bar{\mathrm{w}}$, the dynamics expand to:

$$dx_t = \left(v_\theta(x_t, t) - \frac{1}{2} g(t)^2 s_\theta(x_t, t) \right) dt + g(t) d\bar{\mathrm{w}} \tag{2}$$

This continuously counterbalances white Gaussian noise injection with a gradient step along the predicted score. By natively correcting accumulated discretization errors, SDE solvers anchor the trajectory firmly to the data distribution [61, 62].

Power Spectral Density and Noise Colors. The frequency composition of the injected noise ϵ is characterized by its Power Spectral Density (PSD). Letting $\hat{\epsilon}(f) = \mathcal{F}(\epsilon)$ denote its Fourier transform, the PSD $S(f)$ evaluates the expected energy at frequency f:

$$S(f) = \mathbb{E}[|\hat{\epsilon}(f)|^2] \tag{3}$$

The shape of $S(f)$ defines the noise "color" [43] as illustrated in Fig. 2. Standard Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ possesses a constant $S(f)$, injecting equal energy across all frequencies (*white noise*). Conversely, non-uniform spectra produce colored noise, such as high-frequency dominant *blue noise*. Due to Fourier's orthogonality, Parseval's theorem ensures that integrating the PSD yields the total spatial energy: $\mathbb{E}[\|\epsilon\|^2] = \int S(f) df$. Consequently, standard SDEs fundamentally operate by blindly injecting a fixed, frequency-agnostic white noise energy budget at every generative step.

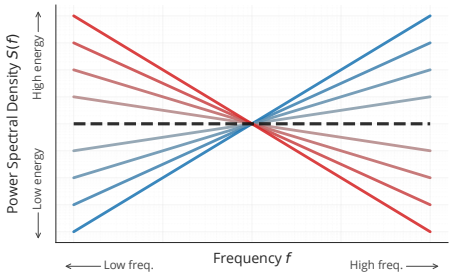


Figure 2: PSD of different colored noises. The spectra transition smoothly from high-frequency dominant blue noise, through uniform white noise (center black line), to low-frequency dominant red noise.

3.2 Spectral Bias of Diffusion Models

Spectral bias is a well-documented inductive property that extends beyond training optimization to fundamentally govern the inference dynamics of diffusion models [47, 49, 65]. Rather than resolving the image uniformly, generation follows a staggered frequency evolution.

Following [20], we quantify this band-wise progression using the bounded index $\gamma(f, t) \in [0, 1]$ (complete derivation detailed in App. B). Visualizing this γ-matrix (Fig. 3) directly exposes these generation dynamics: low-frequency structures resolve early in the generation process. In contrast, high-frequency details evolve at a gradual rate, only fully materializing at the very end of the sampling trajectory. This provides a precise temporal map dictating exactly when specific frequency bands are actively being "built" by the network.

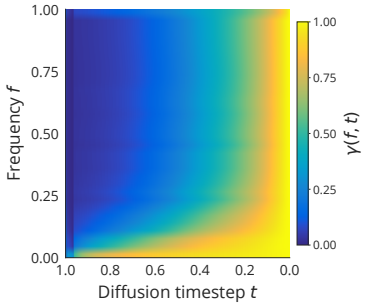


Figure 3: Temporal progression of frequency bands during sampling.

3.3 Structural Preservation and Energy Transfer in Diffusion Models

The mapping from the prior $\mathcal{N}(0, I)$ to the data distribution p_{data} is not an arbitrary coupling between the two spaces. Empirical evidence shows that the inference process preserves significant information from the initial noise realization [70, 68, 63], naturally following minimal-distance trajectories [19].

To explain this geometrically, Wang and Vastola [65] demonstrate that sampling trajectories are surprisingly low-dimensional. Rather than taking an unconstrained walk across the latent space, these trajectories effectively resemble 2D rotations of $\theta \approx 1$ radian from the initial noise state toward the target data manifold. In high-dimensional spaces, where independent random vectors are nearly orthogonal, this rotational angle yields a remarkably high expected cosine similarity ($\cos(\theta) \gg 0$).

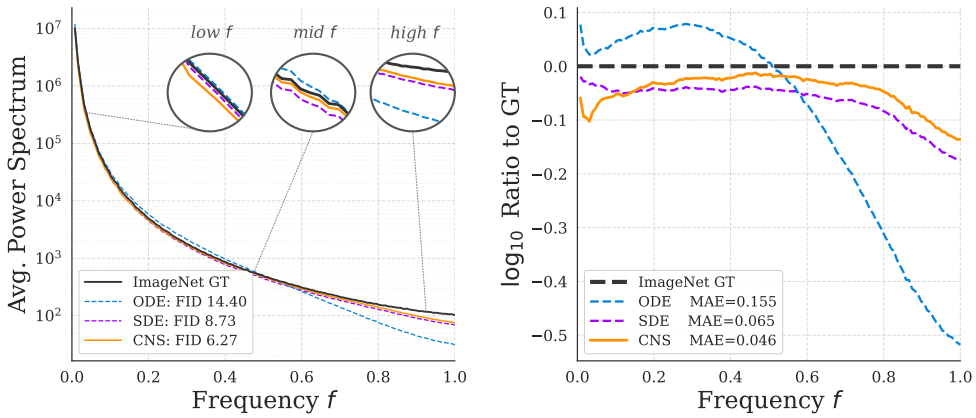


Figure 4: **The Spectral Gap Across Sampling Methods.** *(Left)* PSDs of the generated distributions versus the PSD of the ground truth ImageNet. Standard ODE sampling over-generates low-frequency structures and under-generates high-frequency details, while standard SDE sampling exhibits an energy deficit across the entire spectrum. *(Right)* The signed $\log_{10}$ error relative to the ground truth (black dashed line). By dynamically reallocating the injected noise budget, CNS better aligns the generated spectrum with the true data manifold, mitigating the spectral gap and achieving the lowest log-space Mean Absolute Error (MAE) across frequencies.

This mathematically confirms that the diffusion process does not generate novel structures from scratch, but rather preserves a substantial portion of the initial structural signal—a phenomenon we empirically visualize across spatial frequencies in Fig. 5a.

This rotational perspective has profound implications for the sampling dynamics. Since rotations preserve the L_2 norm (i.e., the vector's energy), inference acts as a signal transfer mechanism that retains a significant portion of the initial noise's energy. The model deterministically maps this preserved noise onto the structured spatial features of the final image, a property implicitly leveraged by recent works optimizing initial noise selection [75, 2, 37]. This non-destructive property forms the foundational premise of our approach: it implies that by strategically controlling the energy injected into the process via noise, we directly control the structural features of the final image.

3.4 The Generated Image Distribution Spectrum

As established in recent works [8, 1, 72], a distinct discrepancy exists between the PSD of generated images and the true data manifold—known as the *spectral gap*. Let $S_{\text{data}}(f)$ and $S_0(f)$ denote the PSDs of the real and generated distributions, respectively. As illustrated in Fig. 4, neither deterministic (S_0^{ODE}) nor stochastic (S_0^{SDE}) sampling perfectly recovers the ground truth $S_{\text{data}}(f)$. Crucially, resolving this gap requires more than simple post-hoc spectral matching, to achieve true distributional matching, the restored energy must align with the coherent spatial structures of the target manifold.

Fortunately, as established in Sec. 3.3, diffusion inference fundamentally operates as a partial energy-preserving signal transfer. Beyond preserving the initial noise realization, we found that the diffusion process also maps the stochastic increments injected by SDE solvers directly into corresponding spatial frequencies of the final generated structure. We formalize this non-destructive, frequency-coupled mapping by isolating spatial frequency bands via a Fourier-space band-pass projection operator, $P_b[\cdot]$. This allows us to quantify the structural alignment between the accumulated injected noise, $\epsilon_{\text{cumul}} = \sum_{i=0}^{T-2} g(t_i) d\bar{\text{w}}_{t_i}$, and the final generated image x_0. By calculating their expected cosine similarity in Fig. 5b, we observe a significant positive correlation:

$$\mathbb{E}\left[\cos\left(P_b[\epsilon_{\text{cumul}}], P_b[x_0]\right)\right] \gg 0 \tag{4}$$

This strong alignment reveals a powerful theoretical pathway: strategically shaping the spectral profile of the injected noise provides a direct mechanism to steer the PSD of the generated distribution toward the target data manifold.

The Impact of Stochasticity. To restructure these dynamics, we quantify spectral deviation using the signed log error: $\varepsilon(f) = \log_{10}(S_0(f)/S_{\text{data}}(f))$. As shown in Fig. 4, comparing deterministic

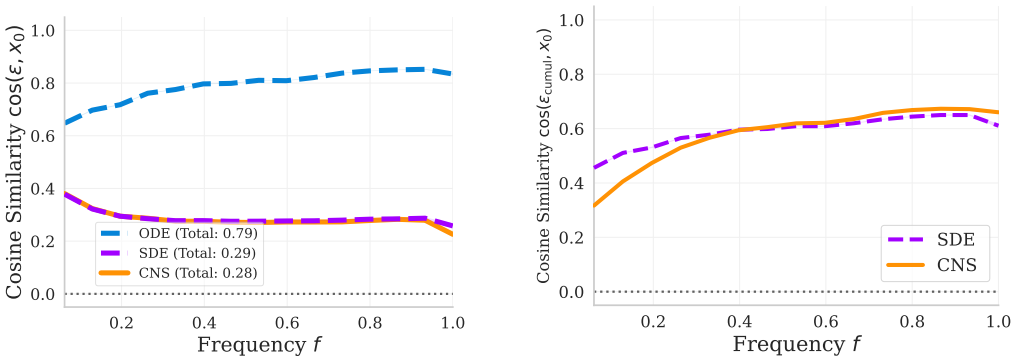


Figure 5: **Noise Signal Preservation and Transfer.** *(Left)* Initial Noise Persistence. Cosine similarity between initial noise and the final generated image. ODEs strongly preserve structural information across the spectrum; stochastic methods (SDE, CNS) still retain a significant, though reduced, amount of this initial signal. *(Right)* Cumulative Injection Transfer. Cosine similarity between total injected noise (ϵ_{cumul}) and the final generated image. This shows injected noise structure actively shapes the final features rather than serving as a temporary perturbation. Notably, CNS selectively routes this signal into higher frequency bands.

and stochastic solvers reveals that continuous noise injection fundamentally alters the final energy distribution. We suggest that this spectral divergence arises from inherent imperfections in the learned score function. During standard Langevin dynamics, the injected noise is not perfectly counterbalanced by the denoising drift. Consequently, score approximation errors cause unintended energy accumulations or deficits over the trajectory (App. D.1).

Crucially, the total stochastic energy injected over the generative trajectory is strictly bounded and mathematically independent of the time discretization (App. C.1). Because we cannot simply scale up the global noise injection to offset deficits without violating the underlying SDE (App. C.2), the stochastic noise acts as a strictly fixed *injected energy budget:*

$$\mathcal{E} = \int g^2(t)dt < \infty \tag{5}$$

Standard SDEs distribute this budget naively: uniform white noise allocates energy equally across the entire frequency spectrum ($S(f) = 1$). By transitioning to targeted *colored noise*, we treat this as a zero-sum game: we dynamically decrease energy allocation for structurally resolved frequency bands, freeing up the budget to inject energy into lagging frequencies. This principled reallocation steers the generated output toward the true data manifold without pushing the intermediate latents out-of-distribution (App. C.3).

3.5 Colored Noise Sampling (CNS)

CNS actively mitigates the spectral gap by repurposing the SDE's stochastic energy leak to steer the generated profile. As derived in App. D.1, the *effective* energy a generated sample absorbs from noise injection is highly state-dependent. Specifically, the band-wise energy absorption rate depends strictly on the correlation between the current spectral state and the local score error. Because this absorption efficiency varies, uniform white-noise injection is highly suboptimal—it allocates the finite energy budget on frequency modes that are already sufficiently resolved. An optimal strategy must therefore dynamically adapt the noise spectrum to the timestep t and frequency band f.

To formalize this active reallocation, we introduce a frequency-dependent scaling weight $\beta_f(t)$ to the

Algorithm 1 CNS Sampling

```
# gamma: [T, F] completion matrix
# freq_bins: (H,W) radial bin map
x = torch.randn(x_shape)
for i, t in enumerate(torch.
    linspace(0, 1, T)):
    w = torch.randn_like(x)
    scale = torch.sqrt(1-gamma[i])
        # colored PSD
    W = torch.fft2(w)*scale[
        freq_bins]
    w_c = torch.real(ifft2(W))
    w_c = w_c / torch.std(w_c)
    v = model(x, t)
    s = (t*v-x)/(1-t) # score
    D = diffusion(x, t)
    x += (v+D*s)*dt +torch.sqrt(2*D)
        *w_c*torch.sqrt(dt)
```

Table 1: **Evaluation of Unguided Image Generation.** ImageNet-256 evaluation metrics without Classifier-Free Guidance across different sampling methods.

Model	Sampler	FID ↓	sFID ↓	IS ↑	Precision ↑	Recall ↑
SiT-XL/2	ODE	14.39	10.54	99.32	0.59	**0.67**
	SDE	8.26	6.32	131.65	0.68	**0.67**
	CNS (Ours)	**6.27**	**4.73**	**147.33**	**0.71**	0.65
JiT-H/16	ODE	12.41	9.35	43.61	0.63	0.62
	SDE	11.88	8.64	44.30	0.64	0.63
	CNS (Ours)	**8.31**	**7.48**	**45.97**	**0.66**	**0.65**
JiT-B/16	ODE	32.39	11.81	26.60	0.47	0.61
	SDE	36.24	14.32	25.86	0.46	**0.63**
	CNS (Ours)	**26.69**	**9.67**	**27.95**	**0.51**	**0.63**

standard SDE noise increment. This colored-noise modification scales the stochastic Itô energy term for a given frequency f from $\frac{1}{2}g^2(t)$ to $\frac{1}{2}g^2(t)\beta_f^2(t)$ (App. D.2). To maintain the overall stability of the generative process, we enforce a strict global variance-conservation constraint, ensuring the average injected energy across all dimensions remains constant: $\frac{1}{D}\sum_{f=1}^{D}\beta_f^2(t) = 1$. In App. D.2, we demonstrate that a frequency band's capacity to absorb injected stochastic noise into a permanent structure is strictly governed by its progression ratio, tracked by the $\gamma(f,t)$-matrix (Sec. 3.2). As a band approaches a fully resolved state ($\gamma(f,t) \to 1$), the score error correlation decays. The network treats excess injected variance primarily as transient energy to be dissipated, severely diminishing the rate of permanent energy conversion.

To maximize the utility of the finite energy budget, CNS dynamically routes energy away from these resolved bands and actively channels it into lagging frequencies with a high structural deficit. We formulate the CNS allocation schedule such that the variance multiplier is strictly proportional to this structural deficit (further details in App. D.2.3). To satisfy the global energy constraint, we normalize this profile by its Root Mean Square (RMS) across frequencies:

$$\beta(f,t) = \frac{\sqrt{1-\gamma(f,t)}}{\sqrt{\frac{1}{D}\sum_{f'}\left(1-\gamma(f',t)\right)}} \tag{6}$$

This dynamic coloring systematically increases the efficiency of the injected energy, yielding a generated distribution spectrally closer to the true data manifold (Fig. 4). The exact integration of this schedule into standard SDE solvers is detailed in Alg. 1. Calibration of $\gamma(f,t)$ and computational cost are shown to be negligible in App. E.

4 Experiments

We evaluate the performance and robustness of CNS across three key areas: (1) Class-Conditional Generation (Sec. 4.1), demonstrating superiority over ODE/SDE baselines across pixel and latent spaces, varying solver orders, and CFG settings; (2) Text-to-Image Generation (Sec. 4.2), integrating CNS into state-of-the-art flow-matching architectures to enhance visual fidelity while preserving strict semantic alignment; and (3) Ablations and Orthogonality, validating our schedule design choices and showing that CNS provides benefits orthogonal both to models trained with alternative noise distributions and to inference-time spectral methods. Further experiments are reported in App. F.

4.1 Class-to-Image Generation

We evaluate CNS on class-conditional image generation (ImageNet-256 [54]), comparing it to standard ODE and SDE baselines [62, 60]. We assess both pixel-space x-prediction (JiT-H/16, JiT-B/16 [29]) and latent-space v-prediction (SiT-XL/2 [36]) models using their official pre-trained weights (Tab. 1). We measure generation quality, diversity, and manifold coverage using established metrics: Fréchet Inception Distance (FID) [14], spatial FID (sFID) [40] for high-frequency structural assessment, Inception Score (IS) [56], and Precision and Recall [24]. For fair comparison against

Table 2: **Evaluation of SiT-XL/2 by Solver Order.** ImageNet-256 evaluation metrics comparing different solvers and sampling methods. Solvers are categorized by their weak convergence order for SDEs (or deterministic order for ODEs).

Solver	Method	FID ↓	sFID ↓	IS ↑	Prec. ↑	Rec. ↑
Euler-Maruyama [38] (1st-Order Weak)	ODE	14.39	10.54	99.32	0.59	**0.67**
	SDE	8.26	6.32	131.65	0.68	**0.67**
	CNS (Ours)	**6.27**	**4.73**	**147.33**	**0.71**	0.65
Heun [13, 22] (2nd-Order Weak)	ODE	9.35	6.38	126.06	0.67	**0.68**
	SDE	8.00	5.49	132.72	0.69	0.67
	CNS (Ours)	**5.99**	**4.78**	**149.78**	**0.71**	0.65
Stochastic RK [51, 52] (2nd-Order Weak)	SDE (SRK2)	8.14	5.69	132.53	0.69	**0.67**
	SDE (SRK2S)	8.77	6.36	129.68	0.68	**0.67**
	CNS (Ours, SRK2)	**5.91**	4.77	**149.41**	**0.71**	0.65
	CNS (Ours, SRK2S)	5.97	**4.73**	148.55	0.70	0.66
Deterministic RK [10] (5th-Order)	ODE (dopri5)	9.04	6.04	126.49	0.67	0.67

Table 3: **Evaluation of Guided Sampling (CFG)** across SiT and JiT architectures on ImageNet-256.

Model	Sampler	CFG Scale	FID ↓	sFID ↓	IS ↑	Precision ↑	Recall ↑
SiT-XL/2	ODE	1.5	2.15	4.60	258.09	0.81	**0.60**
	SDE	1.5	2.06	4.49	**277.50**	**0.83**	0.59
	CNS (Ours)	1.5	**1.98**	**4.46**	257.68	0.81	**0.60**
JiT-H/16	ODE	2.2	3.92	7.72	62.86	0.76	0.59
	SDE	2.2	2.08	7.59	65.88	**0.79**	0.61
	CNS (Ours)	2.2	**2.03**	**6.98**	**65.89**	**0.79**	**0.62**
JiT-B/16	ODE	3.0	5.83	**8.71**	61.47	0.81	0.43
	SDE	3.0	4.54	11.30	63.02	0.82	0.45
	CNS (Ours)	3.0	4.39	9.48	**63.22**	**0.83**	0.45
	CNS (Ours)	2.6	**4.19**	9.01	60.27	0.81	**0.47**

reported baselines, we utilize originally published metric values wherever available. To maintain identical evaluation conditions across reproduced methods, we fix the initial seed across all methods. We match the sampling steps of the original works: 250 for SiT-XL/2 and 50 for JiT models. All quantitative metrics are computed using 50,000 generated samples against the standard 10,000 reference images. We additionally ran the full FID-50K evaluation across 8 widely and evenly spaced seeds for three settings, holding all hyperparameters and the γ-matrix fixed (App. F.3). Qualitative visual comparisons of CNS against SDE and ODE baselines are provided in App. F.9.

Furthermore, Fig. 6 evaluates our sampler across varying time discretizations. Once a sufficient number of steps for proper stochastic simulation is reached, CNS consistently outperforms standard sampling methods. Further details are provided in App. F.5.

Solver Order. While deterministic ODEs rely on standard Taylor expansions, SDE solvers require Itô-Taylor expansions and are categorized by *strong* (pathwise) and *weak* (distributional) convergence orders. Because generative metrics evaluate distributional alignment, weak convergence is our primary focus. As shown in Tab. 2, we evaluate CNS across solvers with varying weak orders (1st-order Euler-Maruyama; 2nd-order Heun, SRK2, SRK2S). CNS outperforms all tested solvers on SiT-XL/2. Mathematical overviews of these solvers are in App. F.4.

Classifier-Free Guidance (CFG). CFG [15] significantly improves sample fidelity by extrapolating the conditional prediction away from the unconditional baseline. In Tab. 3, we demonstrate that CNS consistently outperforms standard ODE and SDE samplers under CFG across SiT-XL/2, JiT-H/16, and JiT-B/16 (using Euler, 250 steps for SiT, 50 for JiT). When optimal CFG scales diverge between methods, we report the best-performing scale for both the baseline and CNS.

Table 4: **FID-30K (↓) on 64^2 Datasets.** Comparison of baseline models and BNDM variants.

	Prior Methods				BNDM Framework		
Dataset	**IHDM**	**DDPM**	**DDIM**	**IADB**	**ODE**	**SDE**	**CNS (Ours)**
AFHQ Cat	11.02	9.75	9.82	9.19	7.95	18.80	**7.49**
LSUN Church	17.76	13.07	16.46	13.12	10.16	66.71	**8.70**

The absolute margins under guidance are smaller than in the unguided setting, which is expected as guidance already recovers some of the fidelity that stochastic correction otherwise supplies. The remaining gains are nonetheless consistent across models and scales, and statistically unambiguous. Over 8 widely spaced seeds with a full FID-50K evaluation per run, the SDE and CNS confidence intervals do not overlap even at the tightest margin presented (App. F.3). Qualitative comparisons of CFG-enhanced generation are in Fig. 1.

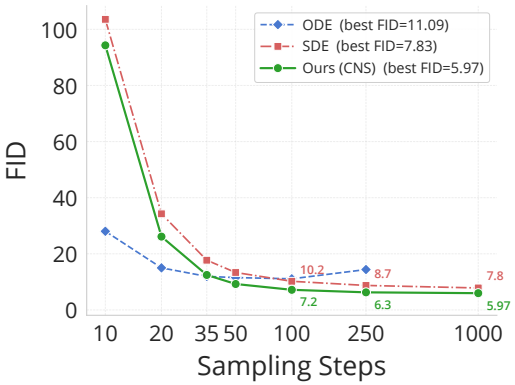


Figure 6: FID-50K vs sampling steps for different samplers.

Orthogonality to Alternative Noise Training. Recent methods [58, 8, 17] alter the training framework to explicitly exploit spectral bias. To confirm CNS is not rendered redundant by such modifications, we integrate our sampler into Blue Noise for Diffusion Models (BNDM) [17], which trains a U-Net [50] with a time-evolving white-to-blue noise distribution. As shown in Tab. 4, applying CNS to BNDM's pre-trained Iterative α-(de)Blending (IADB) [11] models still significantly improves generation quality. Evaluated across two 64^2 datasets ($30,000$ samples), this confirms CNS provides orthogonal inference-time benefits even to models inherently designed to leverage spectral bias. The SDE column in Tab. 4 injects standard white noise, which mismatches BNDM's trained noise distribution. In App. F.8 we implement a corrected stochastic baseline whose injected noise follows the trained white-to-blue schedule; it recovers most of the gap, yet CNS surpasses it on both datasets despite having no access to that schedule, having instead recovered the appropriate coloring empirically from its γ-matrix.

Comparison to Inference-Time Spectral Methods. Prior training-free methods exploit spectral bias by adjusting the predicted state [46, 71], the step-size schedule [28], or the network's internal activations [59], while leaving the solver untouched. We evaluate CNS against all four, each in its own published setting (App. F.2). CNS achieves lower FID than [46] at every NFE on ADM-G/ImageNet-128, more than doubles the gain of [71] on the EDM codebase, and substantially exceeds [28] on ADM-G/ImageNet-64. Across all comparisons, the solver-level correction is not redundant with post-hoc adjustment, and the two compose.

Table 5: **Quantitative evaluation on DrawBench.** Evaluated at 50 steps, CFG scales $w = 3.5,\ 4.0$ for FLUX1, FLUX2, respectively.

Sampler	**ImageReward ↑**	**CLIPScore ↑**	**Aesthetic ↑**
	FLUX.1-dev		
ODE	0.965	0.681	5.787
SDE	0.990	0.689	5.804
CNS (Ours)	**1.012**	**0.693**	**5.812**
	FLUX2.klein		
ODE	0.984	**0.735**	5.233
SDE	0.924	0.733	5.291
CNS (Ours)	**1.005**	**0.735**	**5.295**

4.2 Text-to-Image Generation

Moving beyond class-conditional synthesis, we demonstrate the broad applicability of our method by integrating it into complex downstream generation tasks. Specifically, we apply CNS as a sampler substitution within state-of-the-art Text-to-Image (T2I) flow-matching architectures: FLUX.1-dev

[25] and FLUX.2-klein [26]. We evaluate this integration by comparing CNS against both standard white-noise SDEs and the deterministic ODEs that serve as the default solvers for these models.

We evaluate generation quality and text-alignment across two comprehensive prompt benchmarks. DrawBench [55] probes specific T2I failure modes like complex text rendering (Tab. 5), while GenEval [9] tests precise compositional attributes such as object counts and spatial positioning (Tab. 11). Performance is measured using ImageReward [67] for human preference, CLIPScore [12] for semantic consistency, and Aesthetic Score [57] for visual appeal. Crucially, these evaluations confirm that the dynamic stochastic modifications introduced by CNS enhance overall visual fidelity without degrading the underlying model's text comprehension or corrupting complex compositional instructions.

4.3 Ablation Study

We conduct an ablation study on SiT-XL/2 (Euler, 250 steps), summarized in Tab. 6. While several alternative configurations outperform the standard white-noise SDE, our derived CNS formulation consistently achieves the highest overall fidelity. Specifically, we demonstrate that scaling the injected variance without adhering to our normalization constraint significantly degrades quality, empirically validating the necessity of a strictly finite energy budget. Furthermore, perturbing the CNS schedule, via partial white-noise corruption, static colored spectra, or temporal shuffling, consistently yields inferior results, confirming the optimality of our dynamic, state-dependent allocation. Finally, we establish that replacing our schedule with Multifractional Brownian Motion (mBm) [44], a process generating time-varying colored noise via a shifting Hurst parameter $H(t)$, also falls short of CNS performance. Further mathematical details for all ablations are provided in App. F.7.

Table 6: **Key Ablation Studies (FID-10K).** Full ablation results are in Appendix F.7.

Scenario	FID ↓	sFID ↓	IS ↑
CNS (Ours)	**9.61**	**18.17**	**143.20**
White-noise SDE	11.82	19.15	107.75
Deterministic ODE	17.05	23.57	98.43
50% White Noise	10.64	19.08	136.36
Shuffled Schedule	10.46	19.03	138.02
Constant Spectrum	10.53	19.14	137.63
Scale 0.90 Energy	16.17	42.03	111.29
Scale 1.05 Energy	20.46	29.73	96.17
mBm	11.88	19.62	130.22

5 Conclusion

In this work, we address a fundamental inefficiency in standard diffusion SDE solvers: the uniform injection of white noise, which ignores the model's inherent spectral bias and squanders the finite generative energy budget. By reconceptualizing SDE inference as a targeted energy transfer, we introduce Colored Noise Sampling (CNS), a novel stochastic sampler. CNS actively exploits spectral bias by dynamically reallocating injected noise toward structurally unresolved frequency bands. As a sampler substitution at inference time, CNS systematically steers the generated distribution toward the true data manifold. Extensive experiments validate that CNS outperforms standard baselines across diverse architectures, yielding substantial FID reductions and enhanced visual fidelity without any model retraining.

Limitations and Future Work. The primary limitation of CNS is its reliance on an SDE framework, rendering it incompatible with deterministic ODE solvers. Because stochastic sampling intrinsically requires a high step budget to prevent discretization error accumulation, standard ODEs remain preferable for ultra-fast inference. A second limitation is that CNS applies a single dataset-averaged schedule to every sample; this is by design, as γ tracks the model's inductive dynamics, which are highly consistent across samples, but it leaves per-sample variation unexploited. Future work will explore extending frequency-dependent energy routing into deterministic paradigms for low-step sampling, deriving a per-instance, state-adaptive variant that measures band progression online, and applying CNS to video generation to leverage the temporal frequency dimension.

6 Acknowledgments

This research was supported by The Israel Science Foundation (grant No. 2416/25).

References

[1] K. Adamkiewicz, B. Moser, S. Frolov, T. C. Nauen, F. Raue, and A. Dengel. When pretty isn't useful: Investigating why modern text-to-image models fail as reliable training data generators. *arXiv preprint arXiv:2602.19946*, 2026.

[2] D. Ahn, J. Kang, S. Lee, J. Min, M. Kim, W. Jang, H. Cho, S. Paul, S. Kim, E. Cha, et al. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*, 2024.

[3] M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[4] B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[5] T. Chen, H. Zheng, D. Berthelot, J. Gu, J. Susskind, and S. Zhai. Tada: Improved diffusion sampling with training-free augmented dynamics. *arXiv preprint arXiv:2506.21757*, 2025.

[6] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[7] B. Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[8] F. Falck, T. Pandeva, K. Zahirnia, R. Lawrence, R. Turner, E. Meeds, J. Zazo, and S. Karmalkar. A fourier space perspective on diffusion models. *arXiv preprint arXiv:2505.11278*, 2025.

[9] D. Ghosh, H. Hajishirzi, and L. Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[10] E. Hairer, S. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2008. ISBN 9783540566700. URL `https://books.google.co.il/books?id=F93u7VcSRyYC`.

[11] E. Heitz, L. Belcour, and T. Chambon. Iterative α-(de) blending: A minimalist deterministic diffusion model. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–8, 2023.

[12] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021.

[13] K. Heun et al. Neue methoden zur approximativen integration der differentialgleichungen einer unabhängigen veränderlichen. *Z. Math. Phys*, 45(23-38):7, 1900.

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[15] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[16] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[17] X. Huang, C. Salaun, C. Vasconcelos, C. Theobalt, C. Oztireli, and G. Singh. Blue noise for diffusion models. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024.

[18] A. Hyvärinen, J. Hurri, and P. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* Computational Imaging and Vision. Springer London, 2009. ISBN 9781848824911. URL `https://books.google.co.il/books?id=pq_Fr1eYr7cC`.

[19] N. Issachar, M. Salama, R. Fattal, and S. Benaim. Designing a conditional prior distribution for flow-based generative models. *arXiv preprint arXiv:2502.09611*, 2025.

[20] N. Issachar, G. Yariv, S. Benaim, Y. Adi, D. Lischinski, and R. Fattal. Dype: Dynamic position extrapolation for ultra high resolution diffusion. *arXiv preprint arXiv:2510.20766*, 2025.

[21] M. Kamien and N. Schwartz. *Dynamic Optimization, Second Edition: The Calculus of Variations and Optimal Control in Economics and Management*. Dover Books on Mathematics. Dover Publications, 2013. ISBN 9780486310282. URL https://books.google.co.il/books?id=liLCAgAAQBAJ.

[22] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[23] P. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2011. ISBN 9783540540625. URL https://books.google.co.il/books?id=BCvtssom1CMC.

[24] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

[25] B. F. Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[26] B. F. Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

[27] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/larsen16.html.

[28] H. Lee, H. Lee, S. Gye, and J. Kim. Beta sampling is all you need: Efficient image generation strategy for diffusion models using stepwise spectral analysis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4215–4224. IEEE, 2025.

[29] T. Li and K. He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

[30] D. Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, 2011. ISBN 9781400842643. URL https://books.google.co.il/books?id=xQHEjXy8rlUC.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[32] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[33] E. Liu, X. Ning, H. Yang, and Y. Wang. A unified sampling framework for solver searching of diffusion probabilistic models. In *The Twelfth International Conference on Learning Representations*, 2024.

[34] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[35] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

[36] N. Ma, M. Goldstein, M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, and S. Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024.

[37] J. Mao, X. Wang, and K. Aizawa. Guided image synthesis via initial image editing in diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5321–5329, 2023.

[38] G. Maruyama. Continuous markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo*, 4(1):48–90, 1955.

[39] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pages 7176–7185. PMLR, 2020.

[40] C. Nash, J. Menick, S. Dieleman, and P. W. Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021.

[41] M. Ning, M. Li, L. Zhang, L. Liu, M. B. Blaschko, A. A. Salah, and I. O. Ertugrul. Spectrum matching: a unified perspective for superior diffusability in latent diffusion. *arXiv preprint arXiv:2603.14645*, 2026.

[42] B. Øksendal. Stochastic differential equations. In *Stochastic differential equations: an introduction with applications*, pages 38–50. Springer, 2003.

[43] A. Oppenheim, R. Schafer, and J. Buck. *Discrete-time Signal Processing*. Prentice Hall International Editions Series. Prentice Hall, 1999. ISBN 9780130834430. URL `https://books.google.co.il/books?id=cR3CQgAACAAJ`.

[44] R.-F. Peltier and J. L. Véhel. *Multifractional Brownian motion: definition and preliminary results*. PhD thesis, INRIA, 1995.

[45] M. Plancherel and M. Leffler. Contribution à l'étude de la représentation d'une fonction arbitraire par des intégrales définies. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 30(1): 289–335, 1910.

[46] Y. Qian, Q. Cai, Y. Pan, Y. Li, T. Yao, Q. Sun, and T. Mei. Boosting diffusion models with moving average sampling in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8911–8920, 2024.

[47] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[48] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10674–10685. ieee, 2022.

[49] B. Ronen, D. Jacobs, Y. Kasten, and S. Kritchman. The convergence rate of neural networks for learned functions of different frequencies. *Advances in Neural Information Processing Systems*, 32, 2019.

[50] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[51] A. Rößler. Second order runge–kutta methods for itô stochastic differential equations. *SIAM Journal on Numerical Analysis*, 47(3):1713–1738, 2009.

[52] A. Rößler. Runge–kutta methods for the strong approximation of solutions of stochastic differential equations. *SIAM Journal on Numerical Analysis*, 48(3):922–952, 2010.

[53] F. M. Ruscio and T. K. Rusch. Low-pass flow matching. *arXiv preprint arXiv:2606.02177*, 2026.

[54] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[55] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

[56] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[57] C. Schuhmann. Improved aesthetic predictor. *URL https://github.com/christophschuhmann/improved-aesthetic-predictor*, 2022.

[58] L. Scimeca, T. Jiralerspong, B. Earnshaw, J. Hartford, and Y. Bengio. Learning what matters: Steering diffusion via spectrally anisotropic forward noise. *arXiv preprint arXiv:2510.09660*, 2025.

[59] C. Si, Z. Huang, Y. Jiang, and Z. Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4733–4743, 2024.

[60] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[61] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[62] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[63] Ł. Staniszewski, Ł. Kuciński, and K. Deja. There and back again: On the relation between noise and image inversions in diffusion models. *arXiv preprint arXiv:2410.23530*, 2024.

[64] A. van der Schaaf and J. van Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770, 1996. ISSN 0042-6989. doi: https://doi.org/10.1016/0042-6989(96)00002-8. URL `https://www.sciencedirect.com/science/article/pii/0042698996000028`.

[65] B. Wang and J. J. Vastola. Diffusion models generate images like painters: an analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023.

[66] Y. Wu, Y. Chen, and Y. Wei. Stochastic runge-kutta methods: Provable acceleration of diffusion models. *arXiv preprint arXiv:2410.04760*, 2024.

[67] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[68] K. Xu, L. Zhang, and J. Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3024–3034. IEEE, 2025.

[69] S. Xue, M. Yi, W. Luo, S. Zhang, J. Sun, Z. Li, and Z.-M. Ma. Sa-solver: Stochastic adams solver for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:77632–77674, 2023.

[70] S. Yan, M. Li, B. Xinliang, J. Yang, Y. Zhang, G. Xiong, Y. Lan, T. Zhang, W. Zhai, and Z.-J. Zha. Beyond randomness: Understand the order of the noise in diffusion. *arXiv preprint arXiv:2511.07756*, 2025.

[71] M. Yu, L. Sun, J. Zeng, X. Chu, and K. Zhan. Elucidating the snr-t bias of diffusion probabilistic models. *arXiv preprint arXiv:2604.16044*, 2026.

[72] D. Zhang, T. Zhang, S. Ge, and S. Süsstrunk. Enhancing frequency forgery clues for diffusion-generated image detection. *arXiv preprint arXiv:2511.00429*, 2025.

[73] Q. Zhang and Y. Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

[74] K. Zheng, C. Lu, J. Chen, and J. Zhu. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. *Advances in Neural Information Processing Systems*, 36:55502–55542, 2023.

[75] Z. Zhou, S. Shao, L. Bai, S. Zhang, Z. Xu, B. Han, and Z. Xie. Golden noise for diffusion models: A learning framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17688–17697, 2025.

[76] D. Zou, E. Liu, X. Ning, H. Yang, and Y. Wang. Usf++: A unified sampling framework for solver searching of diffusion probabilistic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

A Continuous-Time Unified Framework Parameterizations

Diffusion Models [16, 62] and Flow Matching [32, 34] can be unified under the continuous-time framework of Stochastic Interpolants [3]. Given a target data distribution $x_0 \sim p_{\text{data}}$ and a tractable noise prior $\epsilon \sim \mathcal{N}(0, I)$, these models construct a probability path via the time-dependent state:

$$x_t = \alpha_t x_0 + \sigma_t \epsilon \tag{7}$$

where $t \in [0, 1]$. Boundary conditions are established such that x_0 strictly represents the clean data ($\alpha_0 = 1, \sigma_0 = 0$) and x_1 approximates the pure noise prior ($\alpha_1 \approx 0, \sigma_1 \approx 1$). Different generative frameworks utilize different schedules for this path. For instance, Variance-Preserving (VP) diffusion employs trigonometric schedules, while Flow Matching utilizes linear paths ($\alpha_t = 1 - t, \sigma_t = t$).

To learn to reverse this continuous-time probability flow, models approximate the conditional interpolant velocity $v_t = \dot{\alpha}_t x_0 + \dot{\sigma}_t \epsilon$. Because the intermediate state x_t is a simple affine combination of data and noise, predicting the velocity v_θ is algebraically equivalent to predicting the clean data $x_\theta \approx x_0$, the noise $\epsilon_\theta \approx \epsilon$, or the marginal score $s_\theta \approx \nabla_{x_t} \log p_t(x_t)$.

Omitting explicit (x_t, t) dependencies for brevity, these parameterizations are deterministically linked by the following relations:

$$v_\theta = \dot{\alpha}_t x_\theta + \dot{\sigma}_t \epsilon_\theta, \quad x_t = \alpha_t x_\theta + \sigma_t \epsilon_\theta, \quad s_\theta = -\frac{\epsilon_\theta}{\sigma_t} \tag{8}$$

A.1 Deterministic vs. Stochastic Sampling Dynamics

Deterministic sampling formulates the trajectory as a Probability Flow ODE (PF-ODE), directly integrating the predicted velocity:

$$dx_t = v_\theta(x_t, t)dt, \quad x_0 = x_1 + \int_1^0 dx_t \approx x_1 + \sum_i v_\theta(x_{t_i}, t_i)\Delta t_i \tag{9}$$

While computationally efficient and approximately invertible, this strict determinism lacks an inherent corrective mechanism. Consequently, discrete numerical approximations and network errors inevitably accumulate, causing intermediate states to gradually drift off the true data manifold and degrade final image fidelity [60].

Stochastic solvers address this drift by simulating the generative process as a reverse-time SDE. Introducing a time-dependent diffusion coefficient $g(t) > 0$ and a reverse-time Wiener process $\bar{\mathrm{w}}$, the dynamics expand to:

$$dx_t = \left(v_\theta(x_t, t) - \frac{1}{2} g(t)^2 s_\theta(x_t, t) \right) dt + g(t) d\bar{\mathrm{w}} \tag{10}$$

This process continuously counterbalances white Gaussian noise injection with a gradient step along the predicted score. The injected noise explores the local neighborhood, while score-based denoising actively pulls the state back toward high-density regions. By natively correcting accumulated discretization errors, SDE solvers anchor the trajectory firmly to the data distribution, yielding superior visual quality [61, 62].

B Spectral Evolution and the γ-Matrix Derivation

By analyzing the frequency spectrum of the clean variable prediction throughout the generation process, recent works have demonstrated a distinct, staggered evolution of frequency modes [65, 20]. The specific rate of this spectral evolution can be quantified using a progression map, or γ-matrix.

To formalize this band-wise progression, we evaluate the model's clean data prediction at each intermediate timestep t. Under a linear schedule, this is given by:

$$x_t^{pred} = x_\theta(x_t, t) = x_t - t v_\theta(x_t, t) \tag{11}$$

Let $X_0(f) = \mathcal{F}(x_0)(f)$ and $X^{pred}(f, t) = \mathcal{F}(x_t^{pred})(f)$ denote the spectral components at frequency band f for the final generated latent and the intermediate prediction, respectively. By

measuring the resolved energy of this intermediate prediction relative to the final outcome, we define the bounded progress index $\gamma(f,t) \in [0,1]$ for every frequency band f:

$$\gamma(f,t) = 1 - \frac{|X_0(f) - X^{pred}(f,t)|^2}{|X_0(f)|^2} \tag{12}$$

This index isolates exactly how much of a specific frequency band's final structure has been resolved by the network at any given timestep t (see Alg. 2). Exact definition of frequencies in our framework is detailed in App. E.1 and $\gamma(f,t)$ convergence study in App. E.3

Numerical Stability of the Progress Index. Eq. 12 divides by $|X_0(f)|^2$, which is small in the high-frequency tail, we therefore provide a stability argument. The numerator is structurally tied to the denominator, making the ratio bounded in expectation. For the MSE-optimal predictor $X_{\text{pred}}(f,t) = \mathbb{E}[X_0(f)|x_t]$, the tower property gives $\mathbb{E}[\langle X_0, X_{\text{pred}}\rangle] = \mathbb{E}[|X_{\text{pred}}|^2]$, hence

$$\mathbb{E}[|X_0(f) - X^{pred}(f,t)|^2] = \mathbb{E}[|X_0(f)|^2] - \mathbb{E}[|X_{\text{pred}}(f,t)|^2] \leq \mathbb{E}[|X_0(f)|^2] \tag{13}$$

The ratio is therefore at most 1 in expectation, and the raw index non-negative, at every band and timestep. Both terms also decay together along the $1/f^\omega$ spectral tail, since X_{pred} estimates X_0 itself, so the ratio remains bounded even where the absolute energy is small. Deviations from this identity arise only from realization noise and from the network's departure from the MSE-optimal predictor, and the implementation absorbs them in three ways. First, the per-cell denominator is regularized ($S_{\text{final}} + 10^{-8}$, Alg. 2). Second, radial binning averages many cells per band, across channels and across the batch. Third, any residual pathological cell inflates the ratio and drives the raw index negative, where the clamp to $[0,1]$ maps it to $\gamma = 0$ – "unresolved", the conservative default, so no instability propagates into $\beta(f,t)$.

Algorithm 2 $\gamma(f,t)$ Matrix Computation

```
# ODE traj x: [T, B, C, H, W],
# x[0]=noise, x[T-1]=image
gamma_sum = torch.zeros(T, F)
for x in ODE_batches:
    t = torch.linspace(0, 1, T)
    v = (x[1:] -x[:-1]) / dt # [T-1,B,C,H,W]
    xp = x[:-1]+(1-t[:-1])*v # clean pred [T-
        1]
    xp = torch.cat([xp, x[-1:]]) # append
        final [T]
    X = torch.fft2(xp)
    Xf = torch.fft2(x[-1]) # final spectrum
    g = 1 -torch.abs(X-Xf)**2 / torch.abs(Xf)
        **2
    g = torch.clamp(g, 0, 1).mean(dim=C) # ch
        . avg -> [T,B,H,W]
    gamma_sum += bin_radially(g).mean(dim=B)
gamma = gamma_sum / N
```

C Theoretical Constraints on Stochastic Energy Injection

In the main text, we establish that the stochastic noise injected during the generative process acts as a strictly fixed *injected energy budget*. This appendix provides the formal derivations underpinning this core premise. First, in App. C.1, we demonstrate that the total injected stochastic energy is mathematically finite and strictly invariant to the time discretization of the solver. Building upon this, in App. C.2, we show that this finite energy budget cannot be globally scaled to offset spectral deficits without violating the underlying SDE and breaking the deterministic convergence to the target data manifold. Together, these theoretical constraints necessitate our zero-sum, frequency-band redistribution approach.

C.1 Invariance of Total Injected Energy to Timestep Discretization

In this section, we demonstrate that for a given continuous diffusion schedule $g(t)$ over a generative interval $[t_0, t_1]$, the total stochastic energy injected by the noise process is finite and strictly invariant to the number of discretization steps used to integrate the reverse SDE. This formalizes our conceptualization of the SDE's stochasticity as a strictly fixed, global injected energy budget.

Consider the continuous reverse-time SDE [4, 62]:

$$dx_t = \mu(x_t, t)dt + g(t)d\bar{\text{w}} \tag{14}$$

where the deterministic drift is defined as $\mu(x_t, t) = \left[v(x_t, t) - \frac{1}{2}g^2(t)\nabla_{x_t} \log p_t(x_t)\right]$, $d\bar{\mathrm{w}}$ is the standard reverse-time Wiener process, and $g : [t_0, t_1] \to \mathbb{R}_{\geq 0}$ is a continuous and bounded diffusion coefficient.

Let our discrete sampler be evaluated over N uniform integration steps using standard Euler-Maruyama numerical integration [23]:

$$x_{k+1} = x_k + \mu(x_k, t_k)\Delta t + g(t_k)\Delta\bar{\mathrm{w}}_k, \qquad \Delta\bar{\mathrm{w}}_k \sim \mathcal{N}(0, \Delta t\mathbf{I}) \tag{15}$$

Isolating a single spatial dimension for clarity, we define the injected stochastic increment at step k as:

$$\eta_k := g(t_k)\Delta\bar{\mathrm{w}}_k \tag{16}$$

and the total expected injected variance per dimension over the entire trajectory as:

$$\mathcal{E}_N := \sum_{k=0}^{N-1} \mathbb{E}[\eta_k^2] \tag{17}$$

Proof of Convergence. Because the diffusion coefficient $g(t_k)$ is deterministic at step k, the expected variance of the increment is:

$$\mathbb{E}[\eta_k^2] = \mathbb{E}\left[\left(g(t_k)\Delta\bar{\mathrm{w}}_k\right)^2\right] = g^2(t_k)\mathbb{E}[\Delta\bar{\mathrm{w}}_k^2] = g^2(t_k)\Delta t \tag{18}$$

Therefore, the total expected variance is:

$$\mathcal{E}_N = \sum_{k=0}^{N-1} g^2(t_k)\Delta t \tag{19}$$

This formulation is exactly the left Riemann sum of $g^2(t)$ over the interval $[t_0, t_1]$. By the continuity of $g(t)$, we can take the continuous-time limit:

$$\lim_{N\to\infty} \mathcal{E}_N = \int_{t_0}^{t_1} g^2(t)dt =: \mathcal{E} \tag{20}$$

Because standard diffusion schedules $g(t)$ are strictly bounded on a finite interval [62], the integral $\mathcal{E}$ evaluates to a finite constant.

Finite-N Discretization Error. Let $\mathrm{Err}(N) := |\mathcal{E}_N - \mathcal{E}|$ denote the variance discrepancy for a discrete solver. If $g^2(\cdot) \in C^1([t_0, t_1])$, the bounded error of the left Riemann sum guarantees:

$$\mathrm{Err}(N) \leq \frac{(t_1 - t_0)^2}{2N} \max_{t\in[t_0,t_1]} \left|\frac{d}{dt}g^2(t)\right| \tag{21}$$

This demonstrates that the convergence is $\mathcal{O}(1/N)$. Consequently, up to a minor, strictly bounded numerical integration error, any chosen number of steps N draws from the exact same finite pool of total injected energy.

C.2 Theoretical Demand for Injected Noise Variance Conservation

In this section, we demonstrate that a Langevin dynamics formulation in which the injected global noise variance is misaligned with the true score function—even under the assumption of an exact oracle score—yields a sampling process that fundamentally fails to converge to the target data distribution, p_{data}.

The Modified Reverse SDE and its Equivalent Flow. In continuous-time diffusion frameworks, the reverse generative process is governed by the stochastic differential equation (SDE) [62]:

$$dx_t = \left[v(x_t, t) - \frac{1}{2}g^2(t)\nabla_{x_t} \log p_t(x_t)\right] dt + g(t)d\bar{\mathrm{w}} \tag{22}$$

where $v(x_t, t)$ is the deterministic drift, $g(t)$ is the diffusion coefficient, and $d\bar{\mathrm{w}} \sim \mathcal{N}(0, dt\mathbf{I})$ is the reverse-time Wiener process.

Let us introduce a constant global variance multiplier β to the stochastic increment. The modified SDE becomes:

$$dx_t = \left[v(x_t, t) - \frac{1}{2} g^2(t) \nabla_{x_t} \log p_t(x_t) \right] dt + \beta g(t) d\bar{\mathrm{w}} \tag{23}$$

To analyze the exact marginal distributions generated by this modified SDE, denoted $\rho_t(x_t)$, we rely on the continuous-time equivalence between SDEs and ODEs. For any reverse-time SDE of the form $dx_t = \mu(x_t, t)dt + \sigma(t)d\bar{\mathrm{w}}$, there exists a deterministic Probability Flow ODE that shares the identical marginal density $\rho_t(x_t)$ at all times [4]:

$$dx_t = \left[\mu(x_t, t) + \frac{1}{2} \sigma^2(t) \nabla_{x_t} \log \rho_t(x_t) \right] dt \tag{24}$$

Substituting our modified drift and diffusion terms into this relation yields the equivalent PF-ODE for our perturbed system:

$$dx_t = \left[v(x_t, t) - \frac{1}{2} g^2(t) \nabla_{x_t} \log p_t(x_t) + \frac{1}{2} \beta^2 g^2(t) \nabla_{x_t} \log \rho_t(x_t) \right] dt \tag{25}$$

If we enforce the requirement that this modified process successfully reaches the target data manifold exactly along the true marginal trajectory, we must have $\rho_t(x_t) = p_t(x_t)$ for all t. Under this condition, the equivalent flow simplifies to:

$$dx_t = \left[v(x_t, t) - \frac{1}{2} g^2(t) \left(1 - \beta^2\right) \nabla_{x_t} \log p_t(x_t) \right] dt \tag{26}$$

However, the true target PF-ODE that corresponds to the unperturbed forward process is known to be:

$$dx_t = v(x_t, t)\, dt \tag{27}$$

This identity is recovered directly from the SDE-to-PF-ODE relation applied to the original reverse SDE ($\beta = 1$): the explicit score subtraction $-\frac{1}{2} g^2 \nabla_{x_t} \log p_t$ is exactly cancelled by the implicit Fokker–Planck correction $+\frac{1}{2} g^2 \nabla_{x_t} \log p_t$ contributed by the diffusion term, leaving only $v(x_t, t)$.

Equating the drift of our modified process with this target requires:

$$\tfrac{1}{2} g^2(t) \big(1 - \beta^2\big) \nabla_{x_t} \log p_t(x_t) = 0 \tag{28}$$

Since $g(t) > 0$ and the score $\nabla_{x_t} \log p_t$ is generically nonzero along the generative trajectory, this constraint is satisfied strictly if and only if:

$$\beta^2 = 1 \tag{29}$$

This mathematically confirms that any global rescaling of the injected stochastic noise disrupts the deterministic flow toward p_{data}.

The Rigid Coupling of Transport and Geometry. To rigorously justify why a mismatch ($\rho_t \neq p_t$) causes the process to fail, we examine the continuity equation governing the mass transport of the system. Let $v^*(x_t, t)$ denote the exact PF-ODE vector field. The temporal evolution of the mismatched density ρ_t driven by this static vector field is:

$$\begin{aligned} \frac{\partial \rho_t}{\partial t} &= -\nabla_{x_t} \cdot \Big(v^*(x_t, t) \rho_t(x_t) \Big) \\ &= -v^*(x_t, t) \cdot \nabla_{x_t} \rho_t(x_t) - \rho_t(x_t) \nabla_{x_t} \cdot v^*(x_t, t) \end{aligned} \tag{30}$$

This expansion exposes a fundamental structural incompatibility:

Advection ($v^* \cdot \nabla_{x_t} \rho_t$): The translation of mass acts upon the spatial gradient of the *mismatched* distribution ρ_t.

Compression/Expansion ($\rho_t \nabla_{x_t} \cdot v^*$): The local expansion and compression of probability mass is controlled by the score field, which is specifically tailored to the geometry of the true data distribution p_t.

Because ρ_t possesses a different spatial gradient than p_t, the static local flow routes mass incorrectly. An area in the latent space might receive an influx of probability mass dictated by v^* expecting the geometry of p_t, but instead encounters the density of ρ_t, inevitably creating artificial bottlenecks or overshoots. The score field does not act as a global attractor; rather, it serves as a rigidly coupled transport mechanism that remains valid only when the state distribution adheres strictly to its corresponding marginal trajectory.

Practical Implications of $\beta^2 \neq 1$. While $\beta^2 = 1$ is required for perfect distributional matching, the practical failure modes diverge depending on the direction of the scaling error:

- **Over-injection ($\beta^2 > 1$):** The excessive stochastic exploration outpaces the restorative capacity of the score-based drift. The state rapidly diffuses off the true manifold, resulting in severe degradation of sample fidelity.
- **Under-injection ($\beta^2 < 1$):** The restorative gradient steps overpower the diminished stochastic exploration. While this theoretically violates perfect marginal matching, the samples may not immediately drift off-manifold. Instead, the generative distribution aggressively contracts around local high-density modes. This preserves the visual coherence of individual samples but results in mode concentration and a measurable loss of generative diversity.

In App. C.3 and App. D.2 we demonstrate that in diffusion models, anchoring the density ρ_t to the true probability path requires relaxing the strict per-dimension variance constraint. Instead, the model enforces an average variance across all dimensions, satisfying the following condition:

$$\frac{1}{D}\sum_{f=1}^{D}\beta_f^2(t) = 1 \tag{31}$$

C.3 Model Robustness to Out-of-Distribution Spectral States

A fundamental assumption in our Colored Noise Sampling (CNS) framework is that actively modulating the per-frequency Signal-to-Noise Ratio (SNR) via the scaling vector $\beta_f(t)$ does not push the intermediate generative states so far Out-Of-Distribution (OOD) that the network fails. In this section, we demonstrate that this robustness is an inherent property of diffusion models. Specifically, because of the network's inductive spectral bias, standard inference trajectories intrinsically operate on highly OOD spectral states, yet the model reliably resolves these states into the true data distribution.

The Theoretical Training Paradigm. During standard continuous-time training (e.g., under a linear Flow Matching schedule, for simplicity), the model is optimized over exact marginals $p_t(x_t)$. Given a clean data sample $x_0 \sim p_{\text{data}}$ and a noise sample $\epsilon \sim \mathcal{N}(0, I)$, the intermediate state flows from pure noise at $t = 1$ to clean data at $t = 0$ via:

$$x_t = (1 - t)x_0 + t\epsilon \tag{32}$$

Applying the Fourier transform yields the spectral composition of the training marginals:

$$\widehat{x}_t(f) = (1 - t)\widehat{x}_0(f) + t\widehat{\epsilon}(f) \tag{33}$$

Later established in App. D.1.2, the expected clean data spectrum exhibits a power-law distribution $|\widehat{x}_0(f)| \approx C/f^{\omega_{\text{amp}}}$, while the noise prior is spectrally flat $|\widehat{\epsilon}(f)| \approx 1$.

If the generative process were to perfectly trace this theoretical training path, the proportion of the resolved image signal at any frequency band f would evolve strictly linearly across time. By isolating the data term and normalizing, the expected target progression ratio $\gamma_{\text{target}}(f, t)$ is defined as:

$$\gamma_{\text{target}}(f, t) = \frac{\mathbb{E}[|\widehat{x}_t(f) - t\widehat{\epsilon}(f)|]}{\mathbb{E}[|\widehat{x}_0(f)|]} = 1 - t \tag{34}$$

Under this idealized paradigm, every frequency band accumulates image information uniformly, maintaining the exact SNR balance dictated by the training marginals.

Inference-Time Spectral Divergence. However, during standard inference, the model does not generate the image uniformly. As quantified by the γ-matrix (Sec. 3.2), the neural network exhibits a profound spectral bias.

Both when initiated from perfect pure noise and integrated with high precision, and when given some intermediate state of the interpolated distribution, the empirical progression of the frequency bands, $\gamma_{\text{actual}}(f, t)$, drastically deviates from the linear target $\gamma_{\text{target}}(f, t) = 1 - t$. Low-frequency bands converge much faster than the training paradigm dictates, resolving their structural energy early in the generation process ($t \gg 0$), while high-frequency bands remain noise-dominated until the very end of the trajectory.

Implications for Colored Noise Sampling. This empirical divergence reveals a critical operational reality: *during standard inference, the intermediate latents x_t are already severely out-of-distribution with respect to the training marginal's PSD and SNR* [8].

At an intermediate timestep t, the theoretical training marginal $p_t(x)$ expects the low frequencies to be only partially built ($\sim 1-t$). In reality, the generated state x_t^{gen} possesses heavily saturated low frequencies ($\gamma_{\text{actual}} \approx 1$). Despite this massive spectral mismatch—where the local SNR of specific bands completely violates the theoretical training expectation—the network's learned score function $s_\theta(x_t, t)$ does not collapse. It continues to stably process this OOD spectrum, utilizing the combined velocity/score dynamics to reliably push the trajectory toward the true image distribution rather than reverting to noise.

This inherent spectral robustness directly validates the core mechanical assumption of Colored Noise Sampling. Because the model natively handles and corrects intermediate states with perturbed frequency-wise SNR distributions, intentionally shaping the injected stochastic energy via our variance scaling operator $\beta_f(t)$ operates safely within the model's established envelope of robustness. By ensuring that the global variance budget is strictly conserved ($\frac{1}{D}\sum \beta_f^2 = 1$), CNS leverages the network's inherent capacity to convert specifically routed frequency-band energy into coherent image structures without inducing catastrophic OOD failure.

D Theoretical Analysis of Spectral Dynamics and Colored Noise

In this appendix, we present a unified theoretical analysis of the spectral dynamics governing generative diffusion models. First, in App. D.1, we investigate the mechanistic origins of the observed spectral gap between ODE and SDE samplers, demonstrating how pathwise energy trajectories diverge due to imperfect score approximation. Building upon this foundation, in App. D.2, we formalize how Colored Noise SDEs could strategically reshape this continuous energy evolution on a strictly band-wise basis. We show that while ideal score conditions enforce a strict zero-sum energy redistribution, the time-dependent decay of state-error correlation allows CNS to fundamentally circumvent this constraint. This enables CNS to permanently and constructively alter the generated spectrum by dynamically routing variance to frequencies where structural conversion efficiency is maximized.

D.1 Theoretical Origins of the Generated Distributions Spectral Difference

To identify the mechanism behind the observed generated distribution PSDs differences (see Fig. 4), we theoretically analyze the temporal evolution of frequency-band energy during sampling for both ODE and SDE trajectories. The goal is to characterize *where* (which frequency bands and timesteps) and *how* (rate and direction of energy transfer) the two samplers diverge. This comparison provides a mechanistic explanation for why their final generated spectra differ and, in turn, clarifies the source of their quality gap.

Although the Probability Flow ODE and reverse-time SDE are constructed to share identical marginals $p_t(x)$ at each time t, their *pathwise* energy dynamics differ substantially due to the network's imperfect approximation of the score. Let the state energy at time t be defined as:

$$E_t := \frac{1}{2}||x_t||_2^2 \tag{35}$$

The factor of $\frac{1}{2}$ is a standard normalization that ensures $\nabla_x(\frac{1}{2}\|x\|_2^2) = x$, which cleanly aligns the energy drift expressions.

D.1.1 Pathwise Energy Dynamics in Continuous-Time Sampling

Let $v(x_t, t)$ denote the deterministic drift of the PF-ODE. The ODE trajectory is $dx_t = v(x_t, t)\, dt$. Applying standard differentiation, the expected energy progression is governed entirely by the alignment between the state and the velocity:

$$\frac{d}{dt}E_t^{\text{ODE}} = x_t^\top v(x_t, t) \tag{36}$$

We recall that the reverse-time SDE drift relates to the ODE drift via the score function. The SDE trajectory is:

$$dx_t = \left[v(x_t, t) - \frac{1}{2} g^2(t) s_\theta(x_t, t) \right] dt + g(t)\, d\bar{\mathrm{w}} \tag{37}$$

To evaluate the energy differential of this stochastic process, we apply Itô's lemma. Because the process is integrated backwards in time from $t = T$ to $t = 0$, we define the backward time variable $\tau = T - t$, such that $d\tau = -dt > 0$. The forward-in-τ SDE is $dx_\tau = -\left[v - \frac{1}{2} g^2 s_\theta \right] d\tau + g\, d\mathrm{w}_\tau$. Applying Itô's lemma for $f(x) = \frac{1}{2}\|x\|_2^2$ yields:

$$dE_\tau = x_\tau^\top dx_\tau + \frac{1}{2} \mathrm{Tr}\left[g^2(t) I \nabla_x^2 f(x_\tau) \right] d\tau = x_\tau^\top \left(-v + \frac{1}{2} g^2 s_\theta \right) d\tau + \frac{D}{2} g^2 d\tau + x_\tau^\top g\, d\bar{\mathrm{w}}_\tau \tag{38}$$

where D is the data dimension. Reverting to the forward-time derivative $\frac{d}{dt} = -\frac{d}{d\tau}$ and taking the expectation to zero out the Wiener noise, the expected energy drift of the SDE is:

$$\frac{d}{dt} \mathbb{E}[E_t^{\mathrm{SDE}}] = \mathbb{E}[x_t^\top v(x_t, t)] - \frac{1}{2} g^2(t) \mathbb{E}[x_t^\top s_\theta(x_t, t)] - \frac{D}{2} g^2(t) \tag{39}$$

The Ideal Score and Heat Cancellation. To understand this dynamic, assume an ideal oracle score $s_\theta = s^* := \nabla_{x_t} \log p_t(x_t)$. By integration by parts, the expected alignment of the data with its true score is strictly $\mathbb{E}[x_t^\top s^*] = -D$. Substituting this into the SDE drift:

$$\frac{d}{dt} \mathbb{E}[E_t^{\mathrm{SDE}}] = \mathbb{E}[x_t^\top v(x_t, t)] - \frac{1}{2} g^2(t)(-D) - \frac{D}{2} g^2(t) = \mathbb{E}[x_t^\top v(x_t, t)] = \frac{d}{dt} E_t^{\mathrm{ODE}} \tag{40}$$

This mathematically formalizes the ideal generative balance: the stochastic heat explicitly injected by the Itô noise term ($-\frac{D}{2} g^2$) is perfectly and exactly canceled by the restorative radial contraction of the true score ($\frac{D}{2} g^2$).

The Imperfect Score. In practice, the learned score contains approximation errors: $s_\theta(x_t, t) = s^*(x_t, t) + \epsilon(x_t, t)$. Substituting this imperfect score breaks the perfect heat cancellation:

$$\frac{d}{dt} \mathbb{E}[E_t^{\mathrm{SDE}}] = \frac{d}{dt} E_t^{\mathrm{ODE}} - \frac{1}{2} g^2(t) \mathbb{E}[x_t^\top \epsilon(x_t, t)] \tag{41}$$

Integrating this drift difference from $t = T$ (where $E_T^{\mathrm{SDE}} = E_T^{\mathrm{ODE}}$) down to $t = 0$, we isolate the exact energy gap between the generated distributions:

$$\mathbb{E}[E_0^{\mathrm{SDE}}] - E_0^{\mathrm{ODE}} = \int_T^0 -\frac{1}{2} g^2(t) \mathbb{E}[x_t^\top \epsilon(x_t, t)]\, dt = \int_0^T \frac{1}{2} g^2(t) \mathbb{E}[x_t^\top \epsilon(x_t, t)]\, dt \tag{42}$$

Thus, the macroscopic divergence between SDE and ODE generation is controlled strictly by the state-error correlation term $\mathbb{E}[x_t^\top \epsilon(x_t, t)]$. If nonzero, the "heat vs. contraction" balance is broken, and pathwise energy drifts away from the ideal trajectory.

D.1.2 State-Error Correlation and the Power Spectral Density Gap

By Parseval–Plancherel identity [45], this global correlation can be decomposed into independent frequency contributions:

$$x_t^\top \epsilon(x_t, t) = \sum_f \mathrm{Re}\left(\hat{x}_t(f)^* \, \hat{\epsilon}_t(f) \right) \tag{43}$$

We formulate this explicitly for a single frequency f. Defining the correlation functional as $\Gamma_f(t) := \mathbb{E}\left[\mathrm{Re}\left(\hat{x}_t(f)^* \, \hat{\epsilon}_t(f) \right) \right]$, the band-wise energy gap at $t = 0$ obeys:

$$\Delta_f := \mathbb{E}\left[||\hat{x}_0^{\mathrm{SDE}}(f)||^2 \right] - \mathbb{E}\left[||\hat{x}_0^{\mathrm{ODE}}(f)||^2 \right] = \int_0^T g^2(t) \Gamma_f(t)\, dt \tag{44}$$

Because $g^2(t) \geq 0$, the sign of Δ_f is determined exclusively by the accumulated sign of $\Gamma_f(t)$. To determine whether the network error ϵ induces positive or negative state-error correlation, we must account for neural network learning dynamics. Networks trained via Mean Squared Error (MSE) exhibit regression to the mean; when confronted with uncertainty, they output a conservative,

smoothed estimate [27]. Consequently, s_θ systematically underestimates the magnitude of the true score s^*. We model this approximation error as being anti-aligned with the true score: $\epsilon_t \approx -\alpha_f s_t^*$ where $0 < \alpha_f < 1$.

To determine the direction of the true score s^*, we invoke Tweedie's formula [7, 6], which states that the score is proportional to the difference between the expected clean data and the current noisy state: $\hat{s}^*(f,t) \propto (1-t)\mathbb{E}[\hat{x}_0(f) \mid x_t] - \hat{x}_t(f)$. Let N_f represent the energy of the initial noise prior, and R_f represent the target energy of the real data distribution. To ground this physically, natural images exhibit a well-documented $1/f^\omega$ power-law spectrum [64, 18], a structural property that fundamentally persists even within the compressed latent spaces of modern autoencoders [41]. Thus, we can approximate the real-data spectral target as a normalized $R_f \approx C/f^{\omega_{\text{pow}}}$ (with $\omega_{\text{pow}} \approx 2$), while the standard Gaussian prior maintains a flat white-noise spectrum $N_f = 1$. Comparing these frequency-dependent magnitudes defines three distinct generative regimes:

1. The Attenuation Regime ($R_f < N_f$): At frequencies where the target data energy is lower than the initial noise (typically high frequencies), the required evolution is attenuation. Because the expected clean signal magnitude is smaller than the current noisy state, the vector difference points inward. Thus, the true score acts to destroy noise: $\hat{s}^* \propto -c_1\hat{x}_t$. Because the network underestimates this inward pull, the resulting error points outward ($\hat{\epsilon}_t \propto +\alpha_f c_1 \hat{x}_t$). The error is positively aligned with the state, meaning their inner product is positive:

$$\Gamma_f(t) > 0 \implies \Delta_f > 0 \tag{45}$$

Conclusion: The SDE over-allocates energy relative to the ODE. The continuous noise injection is not fully dissipated because the learned score is too weak to adequately attenuate the state.

2. The Amplification Regime ($R_f > N_f$): At frequencies where the target structural magnitude is larger than the initial noise (typically low frequencies), the required evolution is amplification. The expected clean data vector has a significantly larger magnitude ($\|\mathbb{E}[\hat{x}_0(f)]\| > \|\hat{x}_t(f)\|$). Thus, the vector difference points outward: $\hat{s}^* \propto c_2\hat{x}_t$. The underestimation error therefore points inward ($\hat{\epsilon}_t \propto -\alpha_f c_2 \hat{x}_t$). The error is anti-aligned with the state, yielding a negative inner product:

$$\Gamma_f(t) < 0 \implies \Delta_f < 0 \tag{46}$$

Conclusion: The SDE under-allocates energy relative to the ODE. The SDE's constant stochastic disruption, coupled with a weakened score that cannot fully drive the necessary amplification, causes the trajectory to fall short of the ideal energy level.

3. The Crossover Point ($R_f = N_f$): The regime transition occurs exactly at the frequency where the inherent energy of the initial noise matches the target energy of the real data. Here, the expected magnitude of the clean data equals the magnitude of the current noisy state. The score provides a purely tangential (phase-rotational) pull, exerting zero radial pull. Consequently, the inner product of the state and the score (and therefore the error) is zero:

$$\Gamma_f(t) = 0 \implies \Delta_f = 0 \tag{47}$$

Conclusion: At the exact crossover frequency, the radial approximation error vanishes, and the ODE and SDE energy trajectories perfectly match.

D.2 Spectral Impact and Energy Dynamics of Colored Noise SDEs

As proven in App. C.1, the total energy injected by the stochastic process is finite and strictly bounded by the diffusion schedule $g(t)$. We now seek to mathematically formalize how a Colored Noise SDEs reshapes the continuous evolution of this energy on a strictly band-wise basis.

D.2.1 Energy Dynamics under an Ideal Score Function

Uniform Energy Allocation in Standard SDEs. In a standard white-noise SDE, the injected stochastic power is distributed uniformly across all frequency bands at every timestep. Equivalently, each integration step injects a flat Power Spectral Density (PSD), ensuring all spatial frequencies receive an equal share of the finite injected energy budget.

The Colored Noise Variance Operator. We introduce a colored-noise process $d\widetilde{\mathrm{w}}_t$ by scaling the standard noise differently across frequency bands. In the Fourier domain, let:

$$B(t) = \mathrm{diag}(\beta_1(t), \beta_2(t), \ldots, \beta_D(t)) \tag{48}$$

where each $\beta_f(t) > 0$ is a real, time-dependent scaling weight for frequency index f, and D is the data dimensionality. The modified noise increment in the frequency domain is defined as:

$$d\widehat{\widetilde{\mathrm{w}}}_t = B(t)\, d\widehat{\mathrm{w}}_t, \qquad d\widehat{\widetilde{\mathrm{w}}}_t(f) = \beta_f(t)\, d\widehat{\mathrm{w}}_t(f) \tag{49}$$

Mapping back to the spatial domain via the inverse Fourier transform $\mathcal{F}^{-1}$ yields the linear operator form:

$$d\widetilde{\mathrm{w}}_t = \mathcal{F}^{-1} B(t) \mathcal{F}\, d\bar{\mathrm{w}}_t \tag{50}$$

The Global Variance-Conservation Constraint. To preserve the global stability of the generative process (shown necessity in App. C.2), we require the total injected energy of CNS to precisely match the standard white-noise SDE. For our modified process, the expected variance is:

$$\mathbb{E}\left[||g(t)\, d\widetilde{\mathrm{w}}_t||^2\right] = g^2(t) \sum_{f=1}^{D} \beta_f^2(t)\, \mathbb{E}\left[||d\widehat{\widetilde{\mathrm{w}}}_t(f)||^2\right] = g^2(t) \sum_{f=1}^{D} \beta_f^2(t)\, dt \tag{51}$$

Because the standard SDE injects a total variance of $g^2(t) \cdot D \cdot dt$, matching this global energy budget imposes a strict variance-conservation constraint (i.e., the weights must have a Root Mean Square of 1):

$$\frac{1}{D} \sum_{f=1}^{D} \beta_f^2(t) = 1 \tag{52}$$

Band-Wise Energy Drift of Colored Noise SDE. Substituting this scaled noise process into the reverse-time SDE dynamics, the Fourier-domain state evolution for a specific frequency band f becomes:

$$d\widehat{x}_t(f) = \widehat{\mu}(x_t, t)_f\, dt + g(t)\beta_f(t)\, d\widehat{\mathrm{w}}_t(f) \tag{53}$$

Applying Itô's lemma to the band energy $E_t^{(f)} := \frac{1}{2}||\widehat{x}_t(f)||^2$, as derived in App. D.1, the expected band-wise energy drift under the CNS takes the form:

$$\frac{d}{dt}\mathbb{E}\left[E_t^{(f),\mathrm{CNS}}\right] = \mathbb{E}\left[\mathrm{Re}\left(\widehat{x}_t(f)^* \widehat{\mu}(x_t, t)_f\right)\right] - \frac{1}{2} g^2(t) \beta_f^2(t) \tag{54}$$

For comparison, under the standard white-noise SDE ($\beta_f(t) \equiv 1$), the corresponding band-wise drift is:

$$\frac{d}{dt}\mathbb{E}\left[E_t^{(f),\mathrm{SDE}}\right] = \mathbb{E}\left[\mathrm{Re}\left(\widehat{x}_t(f)^* \widehat{\mu}(x_t, t)_f\right)\right] - \frac{1}{2} g^2(t) \tag{55}$$

Hence, the colored-noise modification dynamically alters only the injected Itô heat term from $\frac{1}{2}g^2(t)$ to $\frac{1}{2}g^2(t)\beta_f^2(t)$, enabling targeted frequency-dependent amplification ($\beta_f^2 > 1$) or attenuation ($\beta_f^2 < 1$) while safely preserving the global energy budget.

Interpretation and Required Assumptions. For theoretical tractability (while noting this is an approximation due to the network's spectral bias), assume the ideal oracle score applies a per-band restoring drift that perfectly cancels the standard Itô energy injection. See App. C.3 for further details. Under these assumptions, the perfect score cancels each band at the standard unit scale. Consequently, scaling the injected noise in band f by $\beta_f(t)$ creates an intentional mismatch: $\beta_f^2 > 1$ yields a net energy addition in that band, while $\beta_f^2 < 1$ yields a net energy dissipation. Assuming the model reliably maps this altered stochastic energy into coherent spatial structures (detailed in Sec. 3.3), higher injected variance reliably yields higher generated structural energy in that frequency band.

D.2.2 Zero-Sum Spectral Constraints under an Ideal Score

We isolate the explicit effect of noise shaping by subtracting the standard SDE energy drift from the CNS drift, yielding the instantaneous excess energy injected into band f:

$$\frac{d}{dt}\left(\mathbb{E}\left[E_t^{(f),\mathrm{SDE}}\right] - \mathbb{E}\left[E_t^{(f),\mathrm{CNS}}\right]\right) = \frac{1}{2} g^2(t)\left(\beta_f^2(t) - 1\right) \tag{56}$$

Integrating this over the sampling trajectory $t \in [T, 0]$ formalizes the terminal excess-energy gap for frequency f:

$$\text{Excess}_f := \int_0^T \frac{1}{2} g^2(t) \sum_f \left(\beta_f^2(t) - 1\right) dt \tag{57}$$

Conclusion (Zero-Sum Redistribution). If we scale a specific frequency band such that $\beta_f > 1$, the generated energy in that band increases. However, due to the global conservation constraint ($\frac{1}{D}\sum \beta_f^2 = 1$), deviations from the unit baseline must sum strictly to zero:

$$\sum_{f=1}^{D} \left(\beta_f^2(t) - 1\right) = 0 \tag{58}$$

Therefore, summing the excess energies over all bands confirms global conservation:

$$\sum_{f=1}^{D} \text{Excess}_f = \int_0^T \frac{1}{2} g^2(t) \left(\sum_{f=1}^{D} (\beta_f^2(t) - 1) \right) dt = 0 \tag{59}$$

Under an ideal score, any spectral adjustment is fundamentally a zero-sum game: specific bands may achieve higher energy under CNS, but exclusively at the expense of others.

D.2.3 Breaking the Zero-Sum Constraint via Score Approximation Error

The zero-sum paradigm relies on a critical assumption: that the conversion rate of raw injected noise into permanent generated structural energy is constant across all frequency bands. However, as derived in App. D.1, the actual amount of injected raw noise that successfully converts into permanent structural divergence is strictly dictated by the state–error correlation term $\Gamma_f(t)$. We now demonstrate that this correlation is highly non-stationary and decays rapidly as the structural content of the image resolves. To quantify structural formation, we use the normalized bounded metric representing how "built" a specific frequency band f is at any time t derived in Sec. 3.2 and visually shown in Fig. 3.

Proposed Mechanisms for State-Error Correlation Decay To theoretically explain the decay in state-error correlation ($\lim_{\gamma_f(t)\to 1} \Gamma_f(t) = 0$), we postulate two complementary mechanisms: (1) the leading-order radial component of the true score vanishes as a band resolves, and (2) a possible transition of the network's error on unresolved high-frequency details toward phase-random scatter across consecutive late-phase steps.

We note that the framework presented here is intended as a plausibility argument rather than a rigorous derivation. In particular, we do not require the absolute magnitude of the score error $\|\epsilon\|$ to vanish; we merely suggest that the changing geometric and temporal properties of the terminal generative state could plausibly drive the accumulated correlation toward zero even while $\|\epsilon\| \gg 0$.

1. Tangential Dominance of the True Score. By Tweedie's formula [7], the true score in the frequency domain is proportional to the displacement from the current state toward its conditional clean estimate:

$$\hat{s}^*(f,t) \;\propto\; \mathbb{E}[\hat{x}_0(f) \mid x_t] - \hat{x}_t(f) \tag{60}$$

We suppress the schedule-dependent prefactor here because only the *direction* of $\hat{s}^*$ relative to $\hat{x}_t$ enters the radial inner product we are about to compute. When a frequency band approaches a fully "built" state ($\gamma_f(t) \to 1$), the network has correctly resolved the band's macroscopic content, so the conditional clean estimate is closely aligned with the current state in both magnitude and approximate direction. We characterize the residual mismatch as a small phase rotation:

$$\mathbb{E}[\hat{x}_0(f) \mid x_t] \;\approx\; e^{i\,\delta\phi_f(t)}\,\hat{x}_t(f), \qquad \delta\phi_f(t) \to 0 \;\text{ as }\; \gamma_f(t) \to 1. \tag{61}$$

This is strictly stronger than merely $\|\mathbb{E}[\hat{x}_0(f) \mid x_t]\| \approx \|\hat{x}_t(f)\|$: equal magnitudes alone leave room for an arbitrary chord-like displacement with a substantial radial component, whereas a small-angle phase rotation is tangential to leading order. Expanding:

$$\hat{s}^*(f,t) \;\propto\; \left(e^{i\delta\phi_f} - 1\right)\hat{x}_t(f) \;=\; i\,\delta\phi_f\,\hat{x}_t(f) \;+\; \mathcal{O}(\delta\phi_f^2). \tag{62}$$

The leading-order displacement $i\,\delta\phi_f\,\hat{x}_t$ is geometrically orthogonal to $\hat{x}_t$ in the complex plane (multiplication by i rotates by $90°$), so the radial inner product of state and score is purely a quadratic remainder:

$$\mathrm{Re}(\hat{x}_t(f)^*\,\hat{s}^*(f,t)) \;=\; \mathcal{O}(\delta\phi_f^2). \tag{63}$$

Carrying the early-phase underestimation model $\hat{\epsilon}_t \approx -\alpha_f \hat{s}^*$ from App. D.1 forward, the same quadratic suppression propagates to the state-error correlation:

$$\mathrm{Re}(\hat{x}_t(f)^*\,\hat{\epsilon}_t(f)) \;\xrightarrow{\gamma_f(t)\to 1}\; 0. \tag{64}$$

2. Transition to Phase-Random Error on Unresolved Details. During the early phases of band formation ($\gamma_f(t) \ll 1$), MSE training induces a temporally coherent radial bias—specifically, a systematic underestimation of the score's amplitude along the state direction. However, this coherence breaks once the band's macroscopic magnitude is established ($\gamma_f(t) \to 1$). At this terminal stage, the network's residual task shifts to resolving fine sub-structural details, such as sharp edges and exact phase alignments. Because these high-frequency features are difficult to infer from a condensed latent representation, the network's errors abandon any consistent directional bias, transitioning instead into erratic, phase-random fluctuations across consecutive late-phase integration steps.

To see how such a transition could matter, it is useful to write the band-wise inner product in polar form:

$$\mathrm{Re}(\hat{x}_t(f)^*\,\hat{\epsilon}_t(f)) = |\hat{x}_t(f)|\,|\hat{\epsilon}_t(f)|\,\cos(\theta_\epsilon(t) - \theta_x(t))\,. \tag{65}$$

The contribution of the late steps to the accumulated energy gap is governed by the time average of $\cos(\theta_\epsilon - \theta_x)$ along their trajectory. In the early phase this relative phase is presumably concentrated near 0 or π for many consecutive steps (the radial directions associated with the underestimation model), so per-step contributions would add coherently with a fixed sign. If, in the terminal regime, the relative phase instead varied erratically across the final taken steps and behaved roughly as if uniform on $[0, 2\pi)$, the signed per-step contributions would tend to cancel under temporal accumulation, and the late-phase contribution to Γ_f could plausibly collapse toward zero — even while each individual step still satisfied $\|\hat{\epsilon}_t(f)\| \gg 0$. We offer this as a speculative mechanism that, together with the geometric argument above, may help account for the proposed decay of Γ_f.

State-Dependent Efficiency and the Breakdown of Zero-Sum Redistribution. Together, these proposed mechanisms provide a theoretical framework explaining why the state-error correlation $\Gamma_f(t) = \mathbb{E}\left[\mathrm{Re}\left(\hat{x}_t(f)^*\hat{\epsilon}_t(f)\right)\right]$ collapses once a frequency band is structurally resolved. Under this framework, when a band is fully built ($\Gamma_f(t) \approx 0$), the SDE manages to properly balance stochastic noise injection and score-driven contraction. Here, the same Γ_f that governs the band-wise SDE-vs-ODE gap (Eq. 44) acts as a *conversion gain*. It quantifies the rate at which raw injected stochastic variance is translated into coherent radial growth, rather than dissipating as incoherent fluctuation. As $\Gamma_f \to 0$, this conversion gain vanishes. Consequently, any surplus variance ($\beta_f^2 > 1$) enters the instantaneous variance budget but fails to materialize as coherent image content; instead, subsequent denoising steps treat it as ambient noise and remove it. As a result, the global conservation of injected variance ($\sum(\beta_f^2 - 1) = 0$) does not enforce a zero-sum conservation of *generated* energy, effectively dismantling the zero-sum redistribution constraint.

Ultimately, the true efficacy of colored noise—its *energy absorbance efficiency*—is dictated by the temporal intersection of the variance scaling $\beta_f(t)$ and the decaying structural correlation $\Gamma_f(t)$. Permanently altering the generated spectrum requires injecting variance while a band remains in structural deficit ($\gamma_f(t) \ll 1$). This highlights a fundamental property of stochastic generation: macroscopic spectral divergence is not determined by the total volume of injected energy, but by its timing.

Derivation of the Proposed CNS Allocation Schedule. Having established that energy conversion efficiency decays as $\gamma_f(t) \to 1$, we formulate the optimal time-dependent injection strategy. To maximize the utility of the finite variance budget, the system must avoid injecting excess energy into "built" bands where it will be wastefully dissipated. Instead, it must dynamically route variance into "unbuilt" lagging frequencies where the correlation $\Gamma_f(t)$ remains high, maximizing the structural retention of the injected heat. While a naive maximization strategy might suggest a "bang-bang" controller [30, 21]—routing the entire variance budget exclusively to the single least-built frequency at any given moment—we hypothesize that energy conversion efficiency is fundamentally bounded

by the local magnitude of the injected noise. Overwhelming a single frequency band with excessive stochastic variance surpasses the local restorative capacity of the score network, degrading the model's ability to cohesively integrate or dissipate that noise. Consequently, the optimal schedule must avoid greedy allocation. The variance scaling factor $\beta_f^2(t)$ should instead be smoothly and directly proportional to the structural deficit of the frequency band, $(1 - \gamma_f(t))$. To satisfy the global energy conservation constraint ($\frac{1}{D}\sum \beta_f^2(t) = 1$) at every timestep, we normalize this deficit across the frequency spectrum by its Root Mean Square (RMS):

$$\beta_f(t) = \frac{\sqrt{1 - \gamma_f(t)}}{\sqrt{\frac{1}{D}\sum_{f'=1}^{D}(1 - \gamma_{f'}(t))}} \tag{66}$$

This schedule exhibits several highly desirable theoretical properties:

- **Initialization as White Noise:** At $t = T$, all bands are entirely unbuilt ($\gamma_f(T) = 0$). The allocation evaluates to $\beta_f(T) = 1$ for all frequencies. The generative process naturally begins as a standard uniform white-noise SDE.
- **Dynamic Routing:** As generation progresses, bands accumulate structure at staggered rates. The schedule autonomously drains the variance budget from bands approaching $\gamma_f \approx 1$ and actively routes it into lagging frequencies.
- **Avoidance of Local Saturation:** By scaling proportionally rather than utilizing a binary "bang-bang" injection, the schedule prevents any single frequency band from being overwhelmed by stochastic variance, ensuring the noise remains within the operable restorative capacity of the score network.
- **Optimal Retention:** By shifting the mathematical weight exclusively to regions where the state–error correlation is strongest, the schedule mathematically circumvents the limitations of the zero-sum energy game, ensuring that every unit of injected variance is optimally converted into permanent macroscopic structure.

E Methodological and Experimental Details

This appendix provides the comprehensive methodological and experimental details required to reproduce the Colored Noise Sampling (CNS) framework. While the main text establishes the theoretical premise of spectral energy control, this section translates these concepts into a robust and readily deployable inference algorithm. First, in App. E.1, we formally define the multi-dimensional Fourier projection operators standardly used to isolate and evaluate isotropic radial frequency bands. Next, in App. E.2, we detail our hardware configuration alongside practical schedule relaxations—such as progression scaling and dynamic spectral tilting—that seamlessly integrate CNS into standard sampling pipelines, further enhance algorithmic robustness, and consistently improve synthesis results. The remaining sections quantify what CNS costs to deploy. App. E.3 measures the one-time calibration of the γ-matrix, showing that it converges within seconds to minutes and that downstream quality is essentially insensitive to how long it runs. App. E.4 measures the recurring per-step cost of noise coloring, which is negligible against a single model forward pass. Finally, App. E.5 consolidates both into a concrete protocol for deploying CNS on a new pre-trained model, including the hyperparameter selection procedure that produced every value in Tab. 7.

E.1 Isolation of Radial Spatial Frequency Bands

Let $x \in \mathbb{R}^{C\times H\times W}$ denote a continuous, real-valued multi-channel spatial tensor, such as a generated noise sample or an image. To accurately evaluate the frequency-coupled dynamics of the diffusion process, we must systematically isolate the signal components corresponding to specific spatial scales. This is achieved using the multi-dimensional discrete Fourier transform.

The 2D Discrete Fourier Transform and Shifted Grid. We begin by applying the 2D Discrete Fourier Transform (DFT), denoted as $\mathcal{F}$, independently across each channel c:

$$X_c(u, v) = \mathcal{F}[x_c](u, v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x_c(h, w)e^{-i2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)} \tag{67}$$

To analyze frequencies based on their spatial scale rather than their directional orientation, we shift the 2D frequency indices to center the DC (zero-frequency) component at the origin. Let the shifted frequency coordinates be denoted as (f_y, f_x), defined over the discrete domain:

$$f_y \in [-\lfloor H/2 \rfloor, \dots, \lceil H/2 \rceil - 1], \quad f_x \in [-\lfloor W/2 \rfloor, \dots, \lceil W/2 \rceil - 1] \tag{68}$$

Isotropic Radial Frequencies. Because natural images and standard diffusion noise priors are generally isotropic (rotationally invariant in expectation), we collapse the 2D frequency grid into a 1D measure of spatial scale. We define the radial frequency $\rho(f_y, f_x)$ as the Euclidean distance from the DC component:

$$\rho(f_y, f_x) = \sqrt{f_y^2 + f_x^2} \tag{69}$$

The maximum theoretical radial frequency, corresponding to the corners of the 2D Nyquist limit, is bounded by $\rho_{\max} = \sqrt{(H/2)^2 + (W/2)^2}$.

Discrete Band Partitioning. To perform statistical analysis across the spectrum, we partition the continuous range $[0, \rho_{\max}]$ into N_b discrete radial frequency bands. Each discrete coordinate (f_y, f_x) is mapped to an integer band index $b \in \{0, \dots, N_b - 1\}$ via nearest-integer scaling:

$$b(f_y, f_x) = \left\lfloor \frac{\rho(f_y, f_x)}{\rho_{\max}} (N_b - 1) \right\rceil \tag{70}$$

This mapping defines a set of mutually exclusive 2D frequency masks, B_b, where each set contains all coordinate pairs belonging to band b:

$$B_b = \left\{ (f_y, f_x) \in \mathbb{Z}^2 : b(f_y, f_x) = b \right\} \tag{71}$$

The Projection Operator and Hermitian Symmetry. Using these discrete sets, we define the band-pass projection operator $P_b[\cdot]$, which isolates the spatial signal residing exclusively in band b. This is achieved by taking the inverse 2D DFT of the masked spectrum:

$$P_b[x] = \mathcal{F}^{-1} \left[\mathbf{1}_{(f_y, f_x) \in B_b} \odot \mathcal{F}[x] \right] \tag{72}$$

where $\mathbf{1}$ is the indicator function and $\odot$ denotes element-wise multiplication.

A critical geometric property of the radial distance function ρ is its symmetry across the origin; if $(f_y, f_x) \in B_b$, then $(-f_y, -f_x) \in B_b$. Because the input spatial tensor x is real-valued, its Fourier transform inherently exhibits Hermitian conjugate symmetry ($X(-f_y, -f_x) = X^*(f_y, f_x)$). The isotropic symmetry of the mask B_b perfectly preserves this Hermitian property during the masking operation. Consequently, the inverse DFT guarantees that the projected tensor remains strictly real-valued:

$$P_b[x] \in \mathbb{R}^{C \times H \times W} \tag{73}$$

This allows us to treat $P_b[x]$ not as a complex Fourier series, but as a standard real vector in a lower-dimensional subspace $\mathcal{V}_b \subset \mathbb{R}^{CHW}$. Because the projection operator returns real values, we can seamlessly compute standard spatial metrics—such as the L_2 norm and cosine similarity—directly on the projected tensors.

E.2 Hardware Configuration and Empirical Relaxations

Hardware and Environment. All generative inference and evaluation experiments were conducted using PyTorch with Distributed Data Parallel (DDP) across a compute node equipped with four NVIDIA L40S (48GB VRAM) GPUs.

Empirical Schedule Relaxations. While App. D.2.3 derives the theoretically optimal $\beta(f, t)$ variance allocation schedule, we empirically found that introducing minor algorithmic relaxations helps maintain optimal generative stability. In practice, we implement three parameterized adjustments to our inference framework:

- **Progression Scaling:** To prevent premature variance routing, we soften the strictly computed structural progression matrix by introducing a constant scaling factor $c > 1$, such that the effective progression becomes $\tilde{\gamma}(f, t) = \gamma(f, t)/c$.

- **Dynamic Spectral Tilting:** For specific experiments, we apply a temporally evolving exponential tilt to the injected colored noise spectrum. This smooths the strict band-wise allocation, providing a more stable transition across adjacent frequency bands as generation progresses.
- **Energy Equilibrium Tuning:** Although App. C.2 derives a strict $\beta^2 = 1$ constraint, empirical score networks systematically underestimate the restorative gradient (App. D.2.3), skewing the reverse-time SDE's "heat vs. contraction" balance. We compensate with a micro-scaling of the total injected energy budget; the values used (0.98–0.999, Tab. 7) lie well inside the unit-energy stability basin characterized in Tab. 25.

Exact configuration values and hyperparameter selections across all evaluated architectures (SiT, JiT, FLUX) are comprehensively reported in Tab. 7.

Table 7: Sampling hyperparameters for the baseline architectures evaluated in our experimental setup. Standard experiments.

	SiT-XL/2		**JiT-B/16**			**JiT-H/16**	
Parameter	**w/o CFG**	**w/ CFG**	**w/o CFG**	**w/ CFG**	**w/ CFG**	**w/o CFG**	**w/ CFG**
Architecture & Data							
Space	Latent (VAE)		Pixel			Pixel	
Dataset	ImageNet-256		ImageNet-256			ImageNet-256	
Prediction	v-pred		x-pred			x-pred	
Frequency Bands	32		32			32	
Sampling Steps	250		50			50	
CNS Sampling Settings							
Guidance Scale	–	1.5	–	2.6	3.0	–	2.2
Solver	Euler/Heun	Euler	Euler	Euler	Euler	Euler	Euler
$\gamma(t, f)$ Power	0.75	0.5	0.5	0.5	0.5	0.5	0.5
$\gamma(f, t)$ Divider	1.73	25.0	5.0	10.0	14.0	7.5	25.0
Alpha Tilting Interpolation	$(0.15, -0.5)$	$(-0.1, 0.03)$	-	-	-	-	-
Alpha Tilting Interpolation Type	Exponential (0.75)	Linear	-	-	-	-	-
Noise Energy Scale	0.98	0.998	0.995	-	-	0.98	0.999

E.3 Calibration Cost and γ-Matrix Convergence

Alg. 2 estimates $\gamma(f, t)$ by averaging the band-wise progression of N ODE trajectories (generated with seeds disjoint from all evaluation seeds). Because this estimate must be recomputed for each model, dataset, or checkpoint, its cost and convergence rate determine the practical overhead of CNS.

Convergence criterion. For each (timestep, frequency-bin) cell we track the running mean and variance of the per-sample progress values, and report the 99.9% confidence half-width of the highest-variance cell. Calibration stops when this half-width falls below a threshold.

Empirical study. We snapshot the γ-matrix after N calibration trajectories and evaluate each snapshot with a full FID-30K run on SiT-XL/2 (unguided, 250 Euler steps, single L40S). Baselines are ODE (dopri5) 9.63 and SDE (Euler) 9.28.

Table 8: Performance and fidelity metrics across trajectory counts

N trajectories	**Wall-clock**	**99.9% CI (worst cell)**	**FID**
1	16 s	N/A	6.81
10	22 s	$\pm$0.33	6.81
100	80 s	$\pm$0.10	6.80
1,000	11 min	$\pm$0.03	6.80
10,000	108 min	$\pm$0.01	6.80

Table 9: **Cross-CFG transfer of the γ-matrix.** SiT-XL/2, 50 steps, FID-50K. The mismatched arm samples at CFG 1.5 using a matrix calibrated in the unguided setting.

Arm	CFG	γ-matrix	FID-50K $\downarrow$	sFID $\downarrow$
matched	1.0 (unguided)	unguided	9.373	7.065
mismatched	1.5 (guided)	unguided	2.549	5.677
matched	1.5 (guided)	guided	2.548	5.679

Downstream quality is essentially independent of calibration size as FID varies by at most 0.01 across four orders of magnitude in N, and a single trajectory already delivers the full improvement. Recomputation per model or checkpoint is therefore a one-time cost of seconds.

E.3.1 Transfer of γ-Matrix Across Guidance Scales

We seek to show how specific the γ-matrix is to the setting that produced it by sampling SiT-XL/2 with CFG 1.5 while deliberately using the matrix calibrated in the unguided setting (both at 50 steps). Result are shown in Tab. 9. The mismatched arm is indistinguishable from the matched one (2.549 vs. 2.548 FID; 5.677 vs. 5.679 sFID), indicating that the matrix captures the model's inductive spectral dynamics rather than properties of the guidance setting.

E.4 Per-Step Computational Overhead

Each Colored Noise Sampling step adds a forward FFT, a frequency-domain scaling, an inverse FFT, and an energy renormalization to the standard SDE update (Alg. 1). We benchmark this on SiT-XL/2 (bfloat16, single L40S):

- **FLOPs.** The full coloring operation costs ≈ 52.4 MFLOPs, against 30.37 TFLOPs for a single model forward. A ratio of 1.7×10^{-6}.
- **Isolated latency.** At batch 128, the coloring takes 0.085 ± 0.003 ms versus 293.9 ± 1.1 ms for the model forward (0.029%).
- **End-to-end.** Under the evaluation configuration (250 steps, batch 32), per-step wall-clock rises from 62.04 to 62.18 ± 0.14 ms, an increase of 0.22%.

The overhead is negligible: four orders of magnitude below the model forward in FLOPs, under 0.03% in isolated latency, and 0.22% end-to-end. Also, since the injected noise is stationary within a timestep, the spectral shaping admits an approximate realization as a small spatial convolution, avoiding the transform pair entirely should it ever become a bottleneck.

E.5 Deployment Recipe for New Architectures

Given a new pre-trained model, deploying CNS requires one calibration pass and a short hyperparameter search. All values in Tab. 7 were produced by the procedure below: a coordinate-wise sweep-and-refine over each hyperparameter independently, scored on a cheap proxy metric rather than the full target evaluation, with a single full evaluation run at the end.

1. **Calibrate** γ. Generate 10–100 ODE trajectories at the target resolution and step count, using a representative conditioning set and seeds disjoint from evaluation. Apply the radial binning of Alg. 2. Stop when the 99.9% CI of the worst cell falls below a threshold.
2. **Divider tuning**. Using a cheap proxy metric (FID-5K for class-conditional models; ImageReward or CLIPScore for text-to-image), sweeping 8–10 values over 1.0–10 unguided or 5.0–35.0 with CFG. The optimum is broad and flat (App. F.6).
3. **Optional refinements.** One axis (see Tab. 7) at a time on the same proxy: static alpha tilting, then a dynamic tilting range around it, then 5–10 energy-equilibrium values in 0.975–1.025. All are skippable as a divider-only configuration already outperforms baselines.
4. **Full run.** Run the full target evaluation once, after all sweeps are complete.

F Additional Results and Extended Evaluations

This section provides extended empirical evaluations that supplement the findings presented in the main text. We first present additional quantitative benchmarks across varying architectures and sampling settings (App. F.1). Next, we detail the numerical integration schemes utilized for our stochastic differential equations high-order solvers (App. F.4) and analyze the robustness of CNS across varying numbers of discretization steps (App. F.5). To rigorously validate our theoretical constraints and design choices, we then provide a comprehensive ablation study evaluating the effects of global energy scaling, spectral perturbation, and temporal schedule misalignment (App. F.7). Finally, we elaborate on experiments demonstrating the orthogonality and generalization of CNS to models trained with alternative noise distributions (App. F.8). Finally, we conclude with extended qualitative visual comparisons (without guidance) that highlight the superiority of CNS over standard baselines (App. F.9).

F.1 Extended Generative Benchmarks

To demonstrate consistency across model architectures and evaluation frameworks, we provide supplementary performance tables. Tab. 10 reports the unguided JiT-H/16 results utilizing 100 sampling steps. Under these conditions, CNS strictly dominates both the ODE and standard SDE baselines across all tracked metrics. Furthermore, Tab. 11 (referenced in the main text) details the extensive GenEval [9] benchmarking for the FLUX model, demonstrating CNS's capacity to enhance complex text-to-image synthesis.

Table 10: **Evaluation of Unguided Image Generation Performing 100 Sampling Steps**. ImageNet-256 JiT-H/16 model evaluation metrics without Classifier-Free Guidance across different sampling methods.

Sampler	FID ↓	sFID ↓	IS ↑	Prec. ↑	Rec. ↑
ODE	11.46	10.49	44.71	0.63	0.64
SDE	8.95	7.98	46.38	0.65	**0.65**
CNS (Ours)	**8.57**	**7.16**	**46.72**	**0.66**	**0.65**

Table 11: **Quantitative compositional evaluation on GenEval.** All results are reported as accuracy scores $\in [0, 1]$ (higher is better). **Bold** indicates best performance per column within the model block.

Model & Sampler	**Overall**	**GenEval Task Breakdown**					
		Single Obj.	**Two Obj.**	**Counting**	**Colors**	**Color Attr.**	**Position**
FLUX.1-dev							
ODE	0.643	**0.988**	0.784	0.699	**0.826**	0.413	0.152
SDE	0.635	0.984	0.789	0.685	0.768	**0.433**	0.154
CNS (Ours)	**0.647**	**0.988**	**0.794**	**0.714**	0.803	0.420	**0.164**

F.2 Comparison to Inference-Time Spectral Methods

Sec. 2 positions CNS against training-free methods that exploit spectral bias without entering the solver. We test that positioning directly. Because each method is defined for a particular architecture and sampler, we evaluate in each method's own published setting.

Across four independent settings, CNS matches or exceeds each post-hoc method on its own ground, and composes with [71, 59, 28]. This is the behavior expected if the correction belongs inside the solver rather than around it.

Moving Average Sampling in Frequency Domain (MASF) [46]. Official code is unpublished, so we replicated the evaluation environment (ADM-G, ImageNet-128, FID-50K) and compare against the numbers reported in their Tab. 6, with all relative gains computed on the shared DDPM baseline column. Results are shown in Tab. 12. CNS achieves strictly lower FID at every NFE tested.

Table 12: **Comparison to MASF [46].** ADM-G, ImageNet-128, FID-50K.

NFE	DDPM	+ MASF	+ CNS (Ours)
10	25.24	19.03 (-24.6%)	**18.32** (-27.4%)
15	14.74	10.60 (-28.1%)	**9.70** (-34.2%)
20	10.38	7.33 (-29.4%)	**6.83** (-34.2%)
25	8.37	5.87 (-29.9%)	**5.40** (-35.5%)

SNR-t correction [71]. We integrated CNS into their official EDM codebase (CIFAR-10 32×32, 13-step Euler sampling). CNS requires an active stochastic term and is therefore run with EDM's stochastic sampler ($S_{\text{churn}} > 0$), whereas the baseline and [71] use EDM's default deterministic configuration. Relative gains are reported against the shared deterministic reference. Results are shown in Tab. 13. CNS alone more than doubles the improvement of the post-hoc correction, indicating that the solver-level fix subsumes most of what the correction targets. The two remain compatible, and their combination is best overall.

Table 13: **Comparison to the SNR-t correction [71].** CIFAR-10 32×32, 13-step Euler sampling.

Configuration	FID ↓	vs. EDM
Standard EDM (deterministic ODE)	10.65	—
+ [71]	9.65	-9.4%
+ CNS (Ours)	8.40	-21.1%
+ CNS + [71]	**8.19**	$\mathbf{-23.1\%}$

Beta Sampling [28]. Code is unavailable, so we reimplemented their ADM-G ImageNet-64 protocol. CNS requires an active stochastic term, so we evaluate FID-50K against ancestral DDPM rather than their deterministic DDIM. Both methods are therefore evaluated under identical conditions, and we report relative gains over the shared baseline rather than absolute values, since the baseline sampler differs from theirs. Results are shown in Tab. 14. CNS improves substantially over the baseline and over [28] at every step count. The two compose at 10 steps and at 50 and above.

Table 14: **Comparison to Beta Sampling [28].** ADM-G, ImageNet-64, FID-50K.

Steps	Baseline	+ Beta	+ CNS (Ours)	+ CNS + Beta
10	23.26	23.06	18.57	**17.09**
15	13.64	13.52	**9.17**	9.43
20	9.64	9.77	**6.07**	6.50
50	5.34	5.08	3.49	**3.28**
100	4.43	4.27	3.25	**3.12**

FreeU [59]. FreeU reweights U-Net internal activations, we evaluate in its native setting: Stable Diffusion 2.1 [48] with published code and parameters. Their paper reports only human evaluation, so we adopt DrawBench (200 prompts × 4 images, paired seeds), stochastic sampling, 50 steps, CFG 7.5. CNS uses a γ-matrix calibrated on a COCO-captions [31] proxy prompt corpus. Results are shown in Tab. 15. On this architecture FreeU is the stronger standalone method on ImageReward, and CNS the stronger on CLIPScore. The two are complementary rather than competing, they act on network internals and on the solver respectively, and their combination reaches 0.461 ImageReward, well beyond either alone.

F.3 Statistical Significance Across Seeds

All results in the main text fix a single shared initial seed across samplers, which isolates the effect of the sampler but leaves open whether the reported differences exceed seed-to-seed variance. Several of the CFG-regime margins are in the 2–7% range, where this question is material. We therefore ran the full FID-50K evaluation over 8 widely and evenly spaced seeds, holding the γ-matrix and all hyperparameters fixed at the values used for the published results. Results are shown in Tables 16, 17, 18. Three observations are to be made; first, confidence intervals are non-overlapping between

Table 15: **Comparison to FreeU [59].** Stable Diffusion 2.1, evaluated on DrawBench (200 prompts $\times$ 4 images), stochastic sampling, 50 steps, CFG 7.5.

Condition	ImageReward ↑	CLIPScore ↑	Aesthetic ↑
Baseline	0.141	26.05	5.44
FreeU	0.356	26.47	**5.48**
CNS (Ours)	0.287	**26.56**	5.44
FreeU + CNS	**0.461**	26.13	5.43

samplers in every setting tested, including the tightest margin in the paper (JiT-H/16 under CFG, 2.08 vs. 2.03). Second, every multi-seed mean closely matches its corresponding single-seed value, and the CNS means are in each case marginally better, indicating that the published numbers do not reflect a favorable seed. Finally, the seed standard deviation is small relative to the sampler gaps: for SiT-XL/2 unguided the CNS–SDE gap exceeds 20 standard deviations.

Table 16: **Seed variability, SiT-XL/2 unguided.** FID-50K over 8 widely and evenly spaced seeds.

Mean FID	Std	95% CI	Published (single seed)
6.207	0.091	[6.14, 6.27]	6.270

Table 17: **Seed variability, JiT-B/16 with CFG.** FID-50K over 8 seeds, all samplers.

Sampler (CFG scale)	Mean FID	95% CI
ODE (3.0)	5.83	[5.81, 5.85]
SDE (3.0)	4.52	[4.51, 4.54]
CNS (3.0)	4.38	[4.36, 4.39]
CNS (2.6)	**4.19**	[4.18, 4.20]

Density and Coverage. The precision and recall values in Tab. 1 and Tab. 3 differ only marginally between samplers. We additionally report density and coverage [39], whose manifold-based estimators address known failure modes of precision and recall. Results are shown in Tab. 19. Both metrics cleanly separate the stochastic samplers from the ODE, but do not separate CNS from the standard SDE: the density difference is within seed-level error, and coverage is identical. CNS therefore incurs no cost in fidelity or diversity as these metrics measure them, with its improvements concentrated in the distributional statistics captured by FID and sFID — which, per the preceding section, are themselves statistically significant.

F.4 Numerical Integration Schemes for SDEs

When discretizing the reverse-time Stochastic Differential Equation (SDE) of a generative model, $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + \mathbf{g}(\mathbf{x}, t)d\mathbf{w}$ [62], the choice of numerical solver dictates the approximation error. Unlike Ordinary Differential Equations (ODEs) which rely on standard Taylor series approximations, SDEs require Itô-Taylor expansions [23]. Consequently, SDE solvers are characterized by two distinct types of convergence orders:

- **Strong Order (Pathwise Accuracy):** A solver has strong order p if the expected error of a single trajectory scales as $\mathcal{O}(h^p)$ as the step size $h \to 0$, defined as $\mathbb{E}[\|\mathbf{x}_N - \mathbf{x}(T)\|] \leq Ch^p$. This measures how accurately the solver tracks a specific noise realization.
- **Weak Order (Distributional Accuracy):** A solver has weak order q if the error in the expectation of smooth test functions ϕ scales as $\mathcal{O}(h^q)$, defined as $|\mathbb{E}[\phi(\mathbf{x}_N)] - \mathbb{E}[\phi(\mathbf{x}(T))]| \leq Ch^q$.

In the context of generative modeling, our primary objective is to sample from the correct target distribution rather than accurately track a specific microscopic noise path. Because standard evaluation

Table 18: **Seed variability, JiT-H/16 with CFG (2.2).** FID-50K over 8 seeds, all samplers.

Sampler	Mean FID	95% CI
ODE	3.905	[3.888, 3.923]
SDE	2.061	[2.047, 2.075]
CNS (Ours)	**2.021**	[2.007, 2.035]

Table 19: **Density and coverage [39]** for the Tab. 3 settings, computed with default parameters over 8 widely spaced seeds ($\pm$ 95% CI).

Sampler	Density ↑	Coverage ↑
ODE	1.009 ± 0.011	0.952 ± 0.004
SDE	1.187 ± 0.004	0.978 ± 0.003
CNS (Ours)	1.191 ± 0.010	0.978 ± 0.002

metrics like Fréchet Inception Distance (FID) [14] measure distributional distances, the *weak* order of the solver is the dominant quantity of interest. In our experiments (Sec. 4.1), we evaluate solvers of varying weak orders:

1. **Euler-Maruyama** [38]: The foundational 1st-order weak (and 1/2-order strong) SDE solver, requiring 1 function evaluation per step.
2. **Stochastic Heun**: A 2nd-order weak predictor-corrector method requiring 2 function evaluations per step [13, 22].
3. **Stochastic Runge-Kutta (SRK)**: We utilize two high-order schemes derived by Rößler. **SRK2** [51] achieves weak order 2 (and strong order 1 for additive noise) using 2 evaluations per step. **SRK2S** [52] achieves strong order 1 even for general diagonal multiplicative noise while maintaining weak order 2, requiring 3 evaluations per step.

F.5 Robustness to Discretization Steps

To evaluate the numerical robustness of our approach, we conducted an experiment analyzing performance across varying discretization steps. As the number of sampling steps increases, CNS exhibits a monotonic decrease in FID, consistently outperforming the standard SDE baseline. Notably, CNS matches the peak FID of the ODE sampler using less than half the number of steps required by the standard SDE.

Table 20: **FID-50K across sampling budgets.** SiT-XL/2 unguided, first-order Euler.

Sampling steps (NFE)	10	20	35	50	100	250	1000
ODE	**28.03**	**15.00**	**11.95**	11.53	11.09	14.40	-
SDE	103.56	34.25	17.72	13.34	10.21	8.73	7.83
CNS (Ours)	94.30	26.13	12.46	**9.24**	**7.17**	**6.27**	**5.97**

Conversely, due to inherent limitations in the pre-trained model's underlying framework, the deterministic ODE sampler fails to maintain a monotonic improvement at high step counts; thus, we omit its results beyond the standard 250 steps. While CNS significantly accelerates stochastic convergence, it still shares the fundamental limitation of SDE solvers, requiring a larger minimum number of discretizations than ODEs to properly integrate the underlying differential equations.

F.5.1 Why Stochastic Samplers Underperform at Low Step Counts

In Tab. 1 and Tab. 5, there are results where the standard SDE is worse than the deterministic ODE. This reflects a known property of stochastic sampling rather than an untuned baseline. All baselines use the official pre-trained weights and the step counts of the original works, with the best guidance scale reported separately per method. At low step budgets, injected noise combined with

Table 21: **ODE/SDE crossover with step budget.** JiT-B/16 unguided, ImageNet-256, FID-50K.

Sampler	50 steps	75 steps	100 steps
ODE	**32.39**	31.97	32.56
SDE	36.24	**30.78**	**30.37**

Table 22: **FLUX.2-klein on DrawBench at 70 steps** (vs. 50 in Tab. 5).

Sampler	ImageReward ↑	CLIPScore ↑	Aesthetic ↑
ODE	0.998	**0.737**	5.233
SDE	1.007	0.736	5.281
CNS (Ours)	**1.072**	**0.737**	**5.329**

large discretization steps outgrows the drift's corrective capacity, while deterministic solvers avoid stochastic error accumulation entirely (Sec. 5). Both results sit at 50 steps, marginally within this regime. We show the ordering inverts as the budget grows in Tables 21, 22.

F.6 Robustness to Hyperparameter Selection

CNS introduces several hyperparameters (Tab. 7), We show that improvements do not depend on tuning all of them. We separate the mechanism from its refinements by progressively removing hyperparameters.

Removing Energy-Equilibrium Tuning. The energy-equilibrium parameter (App. E.2) compensates for the score network's systematic underestimation of the restorative gradient. Results in Tab. 23 demonstrate that it is not a requirement for CNS. By removing it CNS still improves substantially over both baselines. It is moreover unused entirely for JiT-B/16 with CFG, and in our BNDM and FLUX configurations (Tab. 7).

Table 23: **Contribution of individual hyperparameters.** SiT-XL/2 unguided, ImageNet-256, FID-50K. Progressively removing CNS hyperparameters degrades results but leaves both baselines outperformed.

Method	FID ↓	sFID ↓	IS ↑
CNS (full)	**6.27**	**4.73**	**147.3**
CNS w/o energy-equilibrium tuning	7.31	5.47	139.9
CNS (divider only)	7.52	6.06	138.3
Standard SDE	8.26	6.32	131.7
Standard ODE	14.39	10.54	99.3

A Single-Hyperparameter Configuration. In the limit, CNS can be reduced to a single scalar by fixing the γ-power to the theory-consistent square root and setting $\beta(f,t) \propto \sqrt{1-\gamma(f,t)}/\mathrm{div}$, leaving only the divider. We sweep it over an intuitive range in Tab. 24. Every value in the sweep outperforms both baselines, and the optimum is broad and flat — the range 1.25–1.75 spans 0.09 FID. A single noise-scale knob with a flat optimum is conventional for stochastic samplers, e.g. EDM's $S_{\text{churn}}, S_{\text{noise}}$ [22]; CNS's additional parameters refine the optimum rather than enable the mechanism.

Table 24: **Divider-only sweep.** SiT-XL/2 unguided, FID-50K, with all other CNS hyperparameters disabled.

Divider	1.10	1.20	1.30	1.40	1.75	2.00	2.50
FID-50K	7.84	7.61	7.54	**7.52**	7.61	7.69	7.87

F.7 Comprehensive Ablation Study

To empirically validate the theoretical constraints and design choices of the Colored Noise Sampling (CNS) framework, we conduct an extensive ablation study. All ablations are performed on the unguided SiT-XL/2 architecture using 250 Euler sampling steps and evaluated on ImageNet-256 (FID-10K, sFID, and Inception Score). The full quantitative results are detailed in Tab. 25. Below, we formalize the methodology and theoretical implications for each ablation category.

F.7.1 Global Energy Scaling (Validating the Variance Constraint)

In App. C.2, we mathematically established the necessity of the global variance-conservation constraint ($\frac{1}{D}\sum \beta_f^2 = 1$). To empirically prove this, we uniformly scaled the total injected energy budget by factors ranging from 0.50 to 2.00. The results are stark: any deviation from a tight neighborhood of the unit-energy budget drastically degrades generation fidelity. Scaling the energy down (e.g., 0.90) starves the SDE of the necessary stochastic exploration required to correct numerical drift, pulling performance toward the deterministic ODE baseline. Conversely, scaling the energy up by even 5% (1.05) destabilizes the process. The excessive injected Itô heat easily overpowers the restorative gradient of the score network, rapidly pushing the latent states out-of-distribution and destroying semantic coherence.

Table 25: **Comprehensive Ablation Studies.** Evaluated on FID-10K (↓), sFID (↓), and IS (↑). We isolate the specific effects of the CNS formulation against alternative baselines, partial noise corruption, temporal schedule permutations, global energy scaling, and multifractional Brownian Motion (mBm) schedules.

Sampling Scenario	FID-10K ↓	sFID ↓	IS ↑
CNS (Ours)	**9.61**	**18.17**	**143.20**
Baselines			
Deterministic ODE	17.05	23.57	98.43
Standard white-noise SDE	11.82	19.15	107.75
Temporal Partial Corruption			
25% White Noise	10.47	18.95	137.91
50% White Noise	10.64	19.08	136.36
50% Random Unit-Energy Spectrum	11.28	19.21	134.81
100% Random Unit-Energy Spectrum	12.26	19.80	127.09
Temporal Schedule Permutations			
Constant Spectrum (Temporal Mean)	10.53	19.14	137.63
Shuffled Schedule	10.46	19.03	138.02
Inverted Schedule (Reverse Time)	10.50	18.91	137.46
Global Energy Scaling (Variance Constraint Violation)			
Scale 0.50 (Half Energy)	106.82	278.69	9.83
Scale 0.75	53.29	161.09	34.11
Scale 0.90	16.17	42.03	111.29
Scale 1.01	11.37	20.21	133.84
Scale 1.05	20.46	29.73	96.17
Scale 1.10	50.63	49.70	46.96
Scale 1.25	171.12	91.28	5.70
Scale 1.50	256.56	134.67	2.12
Scale 2.00 (Double Energy)	327.45	198.37	1.55
Multifractional Brownian Motion (mBm) Variants			
White → Blue ($H: 0.5 \to 0.1$)	13.46	25.16	122.49
White → Blue ($H: 0.5 \to 0.25$)	11.88	19.62	130.22
Red → Blue ($H: 0.9 \to 0.1$)	266.61	224.77	2.97

F.7.2 Spectral Perturbation and Partial Corruption

To validate the precision of our band-wise allocation, we randomly sampled a predefined ratio (25%, 50%, 100%) of timesteps to inject alternative noise distributions rather than the optimal CNS schedule. These injected distributions were either uniform white noise or random unit-energy spectra (where per-band weights are randomized but strictly maintain the global energy budget). As demonstrated in Tab. 25, any corruption of the optimal $\beta(f,t)$ mapping monotonically degrades performance. This confirms that merely ensuring a unit-energy budget is insufficient; the energy must be explicitly routed to the specific frequency bands experiencing the highest structural deficit.

F.7.3 Temporal Schedule Permutations

As theorized in App. D.2.3, the state-error correlation decays as structures resolve, meaning the efficiency of energy injection is highly non-stationary. To prove that *when* energy is injected is just as critical as *where*, we dismantled the temporal alignment of the CNS schedule:

- **Constant Spectrum:** We averaged the dynamic CNS matrix across all timesteps to create a single, static colored noise profile.
- **Shuffled Schedule:** We randomly permuted the timesteps of the allocation matrix.
- **Inverted Schedule:** We applied the schedule backwards relative to the generative time.

While all three variants maintain the exact same total frequency-wise energy injection as the optimal CNS over the full trajectory, they completely fail to match its FID. By injecting energy at incorrect

Table 26: **Matching the injected noise spectrum to BNDM's training distribution.** FID-30K (↓) on 64^2 datasets.

Sampler	AFHQ Cat	LSUN Church
ODE	7.95	10.16
SDE, white noise (Tab. 4)	18.80	66.71
SDE, white→blue	7.83	8.86
CNS (Ours)	**7.49**	**8.70**

timesteps—often into bands that are already structurally resolved—they squander the finite variance budget on transient rotational noise, validating the necessity of a dynamic, state-aware schedule.

F.7.4 Alternative Noise Formulations (mBm)

Finally, we explored an alternative mathematical formulation for generating non-stationary colored noise: multifractional Brownian Motion (mBm) [44]. Unlike standard Wiener processes, mBm introduces a time-varying Hurst parameter, $H(t)$, allowing the stochastic increments to shift continuously between distinct noise colors (e.g., from $H = 0.5$ White noise to $H < 0.5$ Blue noise). Formally, this frequency-dependent shift can be understood through the harmonizable representation of mBm:

$$B_{H(t)}(t) = \frac{1}{C(H(t))} \int_{\mathbb{R}} \frac{e^{it\omega} - 1}{|\omega|^{H(t)+1/2}} d\tilde{W}(\omega) \tag{74}$$

where ω denotes frequency, $d\tilde{W}(\omega)$ is the complex Wiener measure, and $C(H(t))$ is a normalization constant. In this form, the exponent $H(t) + 1/2$ directly governs the spectral density, mathematically enforcing how the noise color evolves over time. While theoretically elegant, mBm dictates a rigid, parameterized shift across the entire spectrum. It lacks the highly granular, per-band structural awareness provided by the γ-matrix. Consequently, even meticulously tuned mBm schedules (e.g., White → Blue, $H : 0.5 \to 0.25$) fall short of performance of the proposed CNS framework, although optimized mBm configurations can still offer marginal improvements over the standard SDE.

F.8 Generalization to Alternative Noise Training

To demonstrate that our framework is orthogonal to alternative noise training paradigms, we evaluated CNS on the official pre-trained IADB models provided by the authors of BNDM [17]. Because BNDM's IADB models are trained as deterministic α-(de)blending with a unique time-evolving white-to-blue noise distribution, they have no native stochastic reverse process. As such, the SDE column of Tab. 4 uses a stochastic solver we derived for their forward process, with standard white-noise increments. On these models, white-noise injection is not merely suboptimal but a distribution mismatch: late in the trajectory the model expects blue-dominant noise, so the low-frequency energy that a spectrally flat process continues to inject is never properly counterbalanced. To confirm this mechanism, we implemented an additional stochastic sampler whose injected noise follows the trained white-to-blue schedule. Matching the injected spectrum to the training noise distribution restores stochastic sampling past the ODE on both datasets. Despite the alternative training objective, we observed that these models still exhibit a pronounced spectral bias. We empirically tracked this bias (Alg. 2) to compute the corresponding structural progression $\gamma(f, t)$-matrix and integrated our CNS method on top of the custom SDE solver. By dynamically shaping the injected noise spectrum according to the matching γ-matrix, CNS achieved significant generation improvements across two evaluated 64×64 datasets using the standard provided test batches (despite having no access to the training noise schedule).

F.9 Additional Visual Comparisons

In this section, we provide extended qualitative results demonstrating the efficacy of our proposed Colored Noise Sampling (CNS) framework compared to standard deterministic (ODE) and stochastic (SDE) baselines. Fig. 7 and Fig. 8 present generation samples across a diverse set of ImageNet classes. To ensure a fair and isolated evaluation of the sampling dynamics, all images within a given row are generated using the exact same noise realizations and class prompt.

As theoretically established earlier, standard uniform white-noise SDEs frequently struggle to resolve high-frequency spatial structures, often yielding blurry or structurally degraded textures. Conversely, while deterministic ODEs preserve structure, they suffer from accumulation errors that lead to over-smoothed, artificial appearances. By dynamically routing stochastic energy to structurally unresolved frequency bands, CNS consistently bridges this gap, yielding sharper fine details (e.g., fur, feathers, and foliage) and a more globally coherent output that better aligns with the true data manifold.

G Broader Impacts

Positive Societal Impacts. Colored Noise Sampling (CNS) enhances generation fidelity as a training-free sampler, circumventing the financial and environmental costs of model retraining. By accelerating convergence, CNS reduces the inference-time carbon footprint of deploying large-scale generative models, democratizing access to computationally efficient, state-of-the-art visual synthesis.

Potential Negative Societal Impacts and Mitigations. Conversely, improving photorealism inherently escalates dual-use risks, including the generation of deceptive content and misinformation. To mitigate these threats, CNS should be deployed alongside robust provenance guardrails, such as cryptographic watermarking and metadata embedding (e.g., C2PA). Furthermore, the unique frequency spectrum alterations introduced by our colored noise framework present a promising avenue for developing more resilient, spectral-based detection systems for AI-generated imagery.

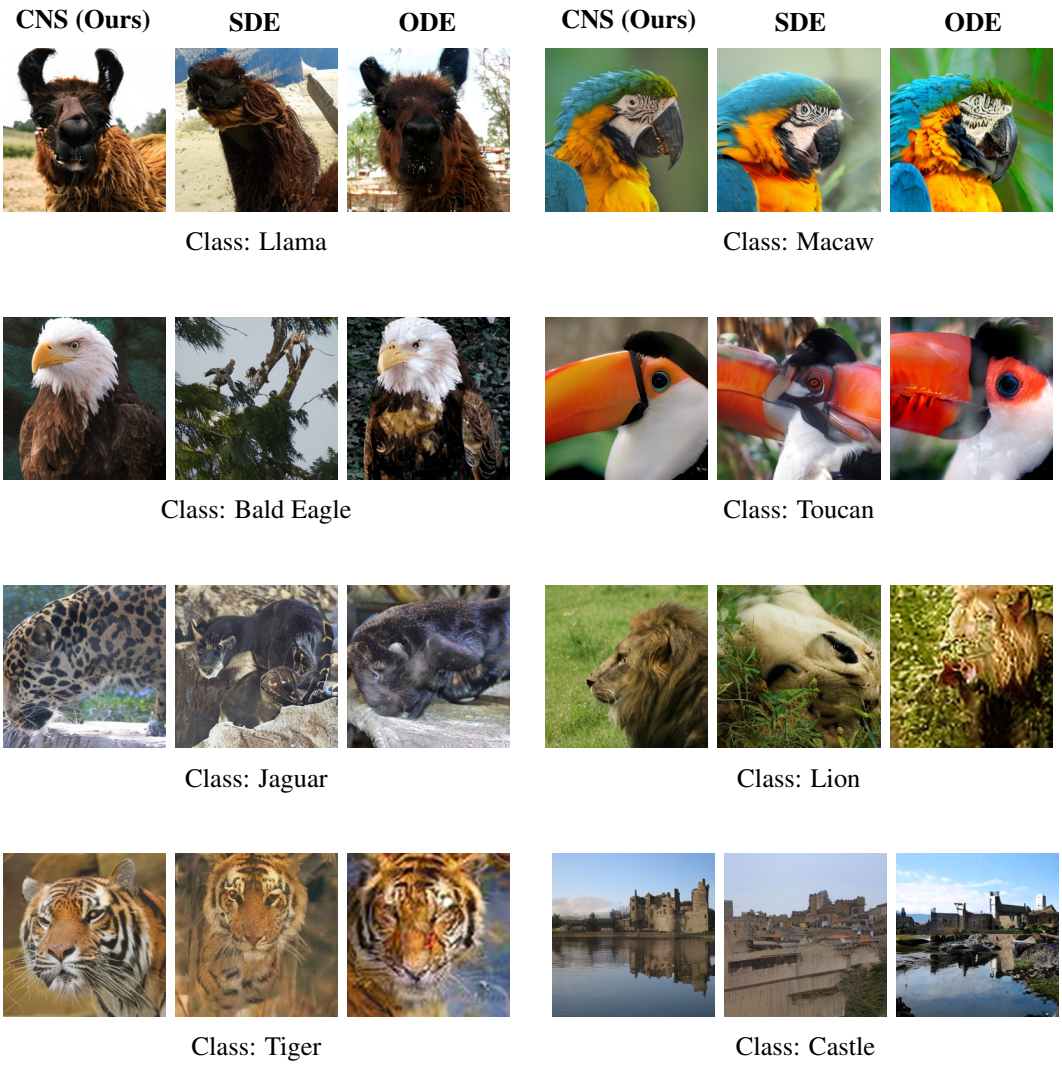


Figure 7: Visual comparison of samples generated using ODE, standard SDE, and our proposed CNS framework (without CFG). All images in a given triplet were generated using the same seed.



Figure 8: Visual comparison of samples generated using ODE, standard SDE, and our proposed CNS framework (without CFG). All images in a given triplet were generated using the same seed.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the mathematical framework and the introduction of the training-free stochastic solver, Colored Noise Sampling (CNS). The claims of improved generation fidelity are directly supported by the empirical results presented across SiT, JiT, and FLUX architectures in Sect. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated "Limitations and Future Work" subsection is included in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Theoretical derivations, core assumptions (such as the ideal oracle score), and complete mathematical proofs regarding variance conservation, energy transfer, and the origins of the spectral gap are detailed extensively in Appendices A D.2.3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: App. E outlines the experimental reproducibility settings. Our code will also be made publicly available ensuring reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code will be made publicly available upon publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec 4 and App E detail the sampling steps, CFG scales, solver types, datasets used, and evaluation metrics required to understand the results

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Main-text metrics (FID, sFID, IS, Precision, Recall) are computed over 50,000 generated samples per configuration, as is standard in the class-conditional generation literature. To establish that the reported differences exceed seed-to-seed variance, App. F.3 reports means and 95% confidence intervals over 8 widely and evenly spaced seeds, with a full FID-50K evaluation per run, for three model/guidance settings including the smallest margin reported in the paper. As well as density and coverage with 95% confidence intervals over the same seeds. Text-to-image results are reported over fixed prompt suites (DrawBench, GenEval) with paired seeds across samplers, without confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: App. E specifies the hardware environment used for inference and evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research develops an inference-time sampling algorithm for existing diffusion models and does not violate the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are discussed in App. G.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper introduces a training-free sampler substitution and does not release a new high-risk pre-trained model, image generator, or scraped dataset that requires usage restrictions or safety filters.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper utilizes standard open-source datasets (e.g., ImageNet) and heavily cites the creators of the pre-trained models utilized (SiT, JiT, FLUX).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or novel pre-trained models are introduced; the core contribution is a training-free stochastic solver for existing models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research evaluates automated generative metrics and does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study does not involve human subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large Language Models are not an important, original, or non-standard component of the core Colored Noise Sampling methodology developed in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.