Diversity in Image Generation via SDE Sampling

Pia Donabauer[1,**] Xiao Han[1,**] Johannes Schusterbauer[2]
[1]LMU Munich
[2]CompVis @ LMU Munich
{pia.donabauer, xiao.han, j.schusterbauer}@campus.lmu.de

Abstract

Diversity in image generation is crucial for developing expressive diffusion models, yet it remains underexplored, particularly in how its various dimensions interact with sampling procedures. In this work, we investigate how sampling strategies (ODE vs. SDE solvers, stochastic noise norms, classifier-free guidance strength and scheduling) affect quality and diversity in conditional image generation. Contrary to the commonly assumed trade-off between quality and diversity, we find that diversity is multi-faceted, since semantic and perceptual diversity can improve alongside sample fidelity under specific configurations. Our findings advocate for a more holistic view of diversity and discuss considerations for tuning generation models toward richer and more controllable outputs.

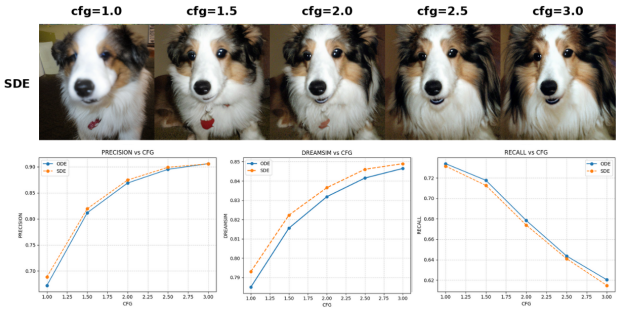

Figure 1. Effect of CFG scale on image quality and diversity. **Top**: Samples generated under SDE sampling with increasing CFG. **Bottom**: While quality (Precision) and perceptual diversity (DreamSim) improves with higher guidance, distributional diversity (Recall) declines.

Code — https://github.com/77Han329/CVPractical

1. Introduction

Diffusion models (DMs) have emerged as a powerful framework for high-resolution images generation, particularly in text-to-image tasks [25,44,47,50,55]. These models learn to reverse a stochastic process that progressively corrupts data with noise. By training a neural network to denoise this corruption step by step, DMs generate new samples that approximate the learned data distribution, starting from pure noise [29, 54, 55].

While most recent advances focus on improving sample efficiency or fidelity, an equally crucial aspect gaining attention is **sample diversity** [15, 21, 29, 37, 48, 52]. Despite generating high-quality and realistic images, DMs often fail to fully capture the variability of the training distribution [48, 57]. Even state-of-the-art models have shown to cover only around 77% of the training diversity [12]. This lack of diversity can lead to repetitive outputs; for instance, a model generating dogs may neglect less common species, poses, or compositions. In extreme cases, this limitation can cause mode collapse (where the model covers only a small subset of the data distribution) [56, 59] or memorization of training data [3, 60]. Such phenomena are especially problematic for data augmentation or synthetic dataset creation, where diversity is critical to ensure generalization [5].

A central factor influencing both diversity and quality is *classifier-free guidance* (CFG) [38]. CFG improves visual quality and prompt alignment by scaling the difference between conditioned and unconditioned model predictions [21, 39]. However, it has also been shown to reduce diversity [23,25,38], creating a quality-diversity trade-off [10, 27], reported via gains in quality (measured by Precision and Inception Score (IS)), and a reduction in diversity (measured by Recall and Fréchet Inception Distance (FID)) [15, 29, 48, 57]. However, the reliance on solely FID and Recall assumes a narrow view of what *diversity* means. FID combines sample quality with coverage evaluation, providing only limited interpretability [39]. Recall, while more targeted, only measures coverage relative to the training distribution. Both fail to capture what kind of variation is present in the samples, e.g., differences in viewpoint,

*Equal contribution.

shape, texture, or semantics. A model might generate fewer unique modes but with richer internal variation, which is not captured by traditional metrics. We argue that diversity in generative models is fundamentally multi-dimensional.

In this paper, we present a large-scale study analyzing how sampling design, including solver type, noise schedule, and CFG strength, affects quality and diversity under this richer perspective. Using the Scalable Interpolant Transformer (SiT) as a representative conditional diffusion model, we evaluate across multiple metrics to uncover trade-offs, sensitivities, and interdependencies. Our findings reveal that while CFG reduces Recall, it can in fact *increase* perceptual and semantic diversity, as shown in Figure 1, suggesting a more nuanced diversity–quality landscape than previously acknowledged. By reframing the analysis of diversity beyond Recall and FID, we aim to foster a more complete understanding of how diffusion models behave under different sampling strategies.

2. Background

We review the foundations of flow and diffusion models, their sampling strategies, and diversity metrics, focusing on their use in generative tasks.

2.1. Flow and Diffusion Models

Modern generative models transform simple noise into realistic images. Given data samples $\{x_i\}_{i=1}^N$ from an unknown distribution q over $\mathbb{R}^d$, the goal is to learn a model that generates new samples $x \sim \hat{q}$ approximating q.

Diffusion Models (DMs) simulate a forward process that gradually adds noise to data, and then learn how to reverse this corruption step by step. The forward corruption process is represented by [35]:

$$x_t = \alpha_t x^* + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where x^* is a real image, ϵ is Gaussian noise, and α_t, σ_t control how much signal and noise are present at time t [55].

Flow Matching Models (FMs) [2, 32] avoid adding noise explicitly. Instead, they learn a time-dependent velocity field $v(x, t)$ that describes how a sample x should move over time to reach the data distribution. Samples are generated by solving an ordinary differential equation (ODE):

$$\frac{dx_t}{dt} = v(x_t, t), \quad x_0 \sim \mathcal{N}(0, I). \tag{2}$$

The result is a smooth, deterministic trajectory that maps noise into realistic samples. Because no randomness is added during sampling, flows are typically faster and more stable [33].

Stochastic Interpolants [1] unify diffusion and flow models. They define flexible interpolated paths between data and noise using[1]:

$$x_t = \alpha_t x^* + \sigma_t \epsilon, \quad t \in [0, 1], \tag{3}$$

crucially, with the freedom to design α_t and σ_t independently of the noise process. This allows the model to disentangle the sampling path from the training process, offering enhanced flexibility.

2.2. Sampling Strategies

Once trained, generative models can run the reverse process and therefore sample new data using either deterministic (ODE) or stochastic dynamics (SDE) [53].

ODE Sampling. [7, 24] Defined in 2, ODE-based models follow a deterministic path defined by the learned velocity field $v(x_t, t)$ with $x_t \sim \mathcal{N}(0, I)$. ODE sampling is often faster and results in cleaner images, but may lead to reduced variability since randomness is not reintroduced during sampling.

SDE Sampling. [35] In contrast, stochastic sampling reintroduces noise at each step using a stochastic differential equation:

$$dx_t = \left[v(x_t, t) - \tfrac{1}{2} w_t s(x_t, t)\right] dt + \sqrt{w_t} dW_t, \tag{4}$$

where $s(x_t, t) = \nabla \log p_t(x)$ is the score function, w_t is a time-dependent noise scale, and W_t is a Wiener process. This added noise may contribute to increased sample diversity.

SDE sampling allows for different diffusion coefficient schedules that govern noise injection over time [66]:

- **Sigma**: The diffusion coefficient $\sigma(t)$ typically follows a variance-preserving or variance-exploding schedule, controlling for the rate and intensity of injected noise as a function of time.
- **Constant**: A simplified formulation where the diffusion coefficient remains fixed throughout the sampling process.
- **Increasing-Decreasing**: A non-monotonic schedule where the diffusion coefficient increases during the early phase of sampling and decreases later.

Different solvers (e.g., Euler, Heun, DPM++) integrate the underlying (S)DEs using distinct numerical schemes, influencing the stability, accuracy, and effective stochasticity of the sampling process [66].

2.3. Classifier-Free Guidance (CFG)

CFG [21] improves the alignment between the generated image and a conditioning signal (e.g., text prompt or class

[1] Stochastic interpolants and diffusion models often share the same formula, but their meanings differ. While diffusion models a noise corruption process over long horizons ($t \to \infty$), stochastic interpolants define a flexible interpolation path over $t \in [0, 1]$, decoupling path design from diffusion-specific constraints.

label) without needing an external classifier [23, 29]. The model is trained to predict noise both with and without conditioning [10]. During sampling, these predictions are linearly combined:

$$\tilde{\epsilon}(x, t; c) = \epsilon(x, t; c) + w\left[\epsilon(x, t; c) - \epsilon(x, t)\right], \quad (5)$$

where $\epsilon(x, t; c)$ is the noise prediction given conditioning c (e.g., a prompt), $\epsilon(x, t)$ is the unconditional prediction, and $w > 1$ is the guidance scale. Larger w improves alignment but can reduce variety [15,21,25,29,38,48] by pushing samples toward the most likely modes [25, 38].

In flow-based models, a similar formula combines conditional and unconditional velocity fields:

$$v_\zeta(x, t; y) = \zeta v(x, t; y) + (1 - \zeta)v(x, t; \emptyset), \quad (6)$$

with ζ playing the role of w.

2.4. Scalable Interpolant Transformer (SiT)

The Scalable Interpolant Transformer (SiT) [35] extends the Diffusion Transformers (DiT) architecture [43] by replacing diffusion-based denoising with flow-based stochastic interpolants. Instead of predicting noise, it learns a velocity field $v(x_t, t)$ to guide a smooth trajectory from noise to image.

Though trained using Flow Matching (i.e., via an ODE), the same model can also be used for SDE-based sampling. This is possible because the score function $s(x, t)$ can be derived from the velocity:

$$s(x_t, t) = -\frac{1}{\sigma_t}\mathbb{E}[\epsilon \mid x_t = x]. \quad (7)$$

This design reduces sampling steps and computational cost while maintaining high sample quality and supporting various sampling configurations, making it suitable for analyzing quality-diversity trade-offs.

2.5. Sample Quality and Diversity Metrics

Evaluating the realism of a single generated image is intuitive, as humans can judge visual plausibility. However, evaluating the overall performance of a generative model is much more challenging [57]. Thus, a comprehensive evaluation of generative models requires both quality and diversity assessments. For completeness, we reviewed a wide range of metrics proposed in the literature. However, we do not include all of these in our main evaluation. More details on selection in Section 4.

2.5.1 Quality

Precision [28] measures the fraction of generated images that fall within the support of the real data distribution. It reflects how many generated samples are realistic.

Density [28] refines Precision by measuring how densely the generated samples populate regions of the data distribution. It incorporates kernel-based neighborhoods to assess the local concentration of generated points near real data.

Inception Score (IS) [51] measures how well a model captures the full ImageNet class distribution while still producing individual samples [7].

Fréchet Inception Distance (FID) [20] compares distributions of real and generated images via Inception-V3 embeddings [58], based on ImageNet [9].

sFID [40] is a spatial variant of FID, computed on intermediate spatial features rather than global averages. It better captures high-level structural coherence [10].

FID with DINOv2 features replaces Inception embeddings with self-supervised DINOv2 representations [42], which have been shown to better align with human judgments [57].

Deep Image Structure and Texture Similarity (DISTS) [11] combines structural and texture features extracted from the Visual Geometry Group (VGG) network to assess perceptual quality.

Peak Signal-to-Noise Ratio (PSNR) [65] quantifies pixel-wise reconstruction error via mean squared error (MSE), insensitive to perceptual quality.

Feature Similarity Index (FSIM) [64] leverages low-level image features such as phase congruency and gradient magnitude. Though designed for perceptual fidelity, it is rarely used in generative modeling due to scale limitations.

2.5.2 Diversity

Learned Perceptual Image Patch Similarity (LPIPS) [65] measures the perceptual distance between image pairs using deep feature activations (e.g., from AlexNet or VGG).

Recall [28] quantifies the fraction of the real data distribution covered by the generated samples. High recall indicates broad coverage of the true distribution [10, 28].

DreamSim [17] computes perceptual similarity using a learned embedding space fine-tuned on human similarity ratings. It aggregates multiple encoder features (CLIP, DINO, OpenCLIP) to build a representation space that correlates more closely with human judgments than any single encoder.

Pretrained Embedding Distance (e.g., DINO, CLIP) compares generated samples by extracting features from models like DINO [42] or CLIP [46]. Similarity measures are then used to assess diversity.

Vendi Score [16] quantifies internal diversity by computing the entropy of a similarity matrix among generated samples. It is reference-free, but is computationally expensive and may fail to detect diversity gaps relative to the training distribution [57].

Validation Loss has been shown to correlate with human and automatic diversity/quality assessments [13, 14, 45]. Hence, it can serve as a practical indirect proxy for diversity.

Structural Similarity Index (SSIM) [61] measures perceived image similarity based on luminance, contrast, and structural consistency. Although less common in generative modeling, it has been applied in conditional generation settings like image translation or inpainting [49].

Chromatic Diversity (HSV) [63] evaluates the spread of generated samples in HSV color space. It requires pairwise color statistics and neighborhood sampling.

Image Retrieval Score (IRS) [12] measures how well generated images retrieve corresponding real samples in feature space. It offers interpretable insight into alignment with the data distribution but is computationally expensive.

Trace of Covariance [41] evaluates the spread of generated samples in a semantic embedding space by computing the trace of the covariance matrix. Higher trace indicates broader dispersion and thus higher diversity in that space.

Coverage [28] extends Recall by evaluating how much of the real data distributino is represented within the generated distribution. It quantifies the proportion of real samples that are sufficiently close to at least one generated image.

3. Related Work

Recent work on conditional diffusion models has focused on improving generation quality through CFG, tuning sampling strategies, or proposing alternative evaluation metrics. However, the interaction between sampling configurations and nuanced dimensions of diversity along with quality remains underexplored.

3.1. Guidance and Sampling Dynamics

The effects of CFG are commonly discussed [10, 21]. Dhariwal and Nichol [10] observe that increasing CFG improves IS and Precision, while worsening Recall and FID. This trade-off has been confirmed by several subsequent studies [15, 29, 48], which report that higher guidance produces sharper, more class-aligned samples, but narrows the distribution of generated outputs.

To mitigate this effect, Moufad et al. [38] introduced a repulsive Rényi-divergence term that discourages overconcentration near the conditioning signal, fostering diversity. Similarly, Sadat et al. [48] propose injecting temporally decaying noise into the conditioning input, allowing greater exploration early in generation, supporting variation. Kynkäänniemi et al. [29] presented Interval Guidance, which applies CFG only at intermediate timesteps, reducing over-saturation and improving FID. However, none of these studies examine whether these techniques preserve or enhance diversity in a semantic or perceptual sense; they primarily evaluate performance using Recall, FID, Precision, and IS. Notably, these approaches conducted the experiments under fixed sampling configurations.

3.2. Metric Behavior and Evaluation Limitations

While FID and related metrics are widely used, their interpretability is limited. Murphy [39] noted that a high FID may reflect poor sample quality, overconcentration, or incomplete coverage of the data distribution, without indicating which. Stein et al. [57] found that FID and IS correlate poorly with human judgments, and are biased against diffusion models due to their dependence on ImageNet-trained Inception features. They also show that while Precision and Density are tightly coupled, Recall behaves independently, underscoring that fidelity and diversity are not captured on a single axis.

Kynkäänniemi et al. [29] note that different diversity metrics favor different CFG intervals, and that no single metric is sufficient to diagnose the quality–diversity tradeoff. Yet even in this analysis, diversity is evaluated only via FID. While these studies raise important concerns about evaluation design, they do not investigate whether higher CFG might actually improve other types of diversity, even as Recall decreases.

3.3. Summary and Research Gap

Prior work has proposed methods such as Interval Guidance and feedback mechanisms to soften the quality-diversity trade-off, but these are usually tested under narrow configurations and evaluated using a small set of traditional metrics.

Most importantly, prior studies implicitly treat diversity as a one-dimensional property, often defined by coverage metrics such as Recall, without considering other forms of variation. We build on the assumption that diversity is fundamentally multi-faceted. To bridge this gap, we conduct a comprehensive analysis of how sampling components, including CFG strength, solver type, and noise schedule, shape different types of diversity. By combining coverage, perceptual, and semantic diversity, we aim to provide a more complete understanding of how generative behavior shifts under different sampling conditions.

4. Methodology

Our methodology is organized around three guiding research questions, with a particular focus on examining diversity and best sampling strategies from multiple perspectives.

Q1: How Many Samples Are Required for Robust Evaluation?

Before conducting configuration experiments, we assess the stability of evaluation metrics with respect to

sample size. Metrics such as FID are known to exhibit high variance at small scales, requiring large sample sizes (up to 50k) for stable estimates [10, 12], with most studies reporting FID@50k [12, 26, 28, 48], and others using FID@10k [34, 48]. In contrast, perceptual metrics such as DreamSim are computed on as few as 100 samples [28], and Precision and Recall are often estimated using 1k samples [10, 43].

Following [13, 14, 45], we additionally compute the model's validation loss on the ImageNet 2012 validation set[2].

We evaluate the following sample sizes:

- **Diversity metrics**: 10, 50, 100, 200 samples
- **Quality metrics**: 1k, 5k, 10k samples
- **Validation loss**: 10, 50, 100, 200, 1k, 5k, 10k samples

For each sample size across diversity metrics, we randomly draw the specific number from a pool of prior 10k generated images, using 5 different random seeds to account for variability. We report the averaged metric across all seeds.

The insights from this sensitivity analysis guide the design choices in subsequent experiments (Q2).

Q2: How Do Sampling Strategies and Parameters Influence Diversity and Quality?

We systematically evaluate how various sampling strategies and configuration parameters affect both diversity and quality. Each component is analyzed across multiple metrics.

- **Q2.1: ODE vs. SDE Sampling.** We compare deterministic (ODE) and stochastic (SDE) solvers to assess how noise injection influences sample variability and perceptual quality. We adopt the Euler solver,[3] a straightforward, fully deterministic method that adds no noise during inference.
- **Q2.2: Classifier-Free Guidance (CFG) Strength.** We vary the guidance scale in the range $[1.0, 1.5, 2.0, 2.5, 3.0]$ to analyze the trade-off between generation fidelity and diversity, as increased guidance typically enforces conditioning at the cost of variation.
- **Q2.3: SDE Diffusion Formulation.** We evaluate three types of diffusion coefficient schedules based on [66] *sigma* (default), *constant*, and *increasing-decreasing*.
- **Q2.4: SDE Noise Norm.** We explore different norm values for the stochastic noise term to assess their effect on sampling dynamics. Specifically, we test norm powers $\ell = \{1, 2, 3, 4\}$, where $\ell = 1$ is the default.
- **Q2.5: CFG Application Interval.** Recent work suggests that applying CFG throughout the entire sampling trajectory may degrade quality [29]. We therefore evaluate selective guidance intervals: $[0.3, 0.7]$, $[0.0, 0.5]$, $[0.2, 1.0]$, $[0.5, 1.0]$, and $[0.0, 0.7]$.

To reduce computational overhead, we fix certain configurations based on preliminary results. Details of this setup are provided in Section 5.

Q3: How Do Diversity and Quality Metrics Correlate?

Following [27, 57], we examine pairwise correlations between diversity and quality metrics to identify trade-offs, synergies, and redundancies. This analysis helps assess whether gains in one dimension (e.g., fidelity) systematically compromise another (e.g., diversity), and clarifies which metrics provide complementary versus overlapping insights.

This analysis is based on the samples generated in Q2.

4.1. Evaluation Metrics

No single metric reliably captures all facets of perceptual quality or diversity in generative models [5]. Many commonly used metrics rely on biased feature extractors [19], are sensitive to dataset size and hyperparameter choices, or suffer from limited interpretability [12]. To address these limitations, we adopt a multi-metric evaluation strategy for a more robust assessment of model behavior. Our metric selection is guided by prior work and includes measures that are widely used and validated in the literature.

4.1.1 Quality Evaluation

We evaluate generative quality using the following metrics: IS (↑, used in [10, 28, 48]), Precision (↑, used in [10, 27, 29, 48]), and FID (↓, used in [12, 28, 29, 48]). These metrics are computed between generated samples and the ImageNet 256×256 validation set [9]. To ensure consistency and comparability, we use the official implementations from [10].[4]

4.1.2 Diversity Evaluation

To measure diversity, we employ distributional, perceptual and semantic metrics that capture variability at different levels of abstraction.

We compute pairwise perceptual distances using LPIPS (↑, used as a diversity measure in [22, 31, 62, 67]) and

[2] https://www.image-net.org/index.php
[3] https://stable-diffusion-art.com/samplers/
[4] https://github.com/openai/guided-diffusion/tree/main/evaluations

DreamSim (↑, used for pairwise image diversity in [28]), averaged across all sample pairs.

Further, we compute Recall (↑, used in [27,29,38,48,57]) between generated samples and the ImageNet validation set, using the implementation from [10].[5]

We extract image embeddings using two pretrained models:

- **DINOv2** [42], using `facebook/dinov2-base`[6], since it has shown to best balance object-centric and holistic scene features [57].
- **CLIP** [46], using `openai/clip-vit-base-patch32`[7] as a widely adopted vision-language encoder.

For both models, we compute cosine similarity using two feature extraction strategies: (a) the `[CLS]` token and (b) spatially averaged patch embeddings (excluding `[CLS]`). We compare both strategies in our diversity evaluation (↑). The use of pretrained embeddings as diversity signals has been explored in [8, 28, 57].

Although FID is commonly treated as a quality metric, it also captures aspects of diversity by reflecting distributional coverage. We therefore include it in our diversity analysis as an auxiliary indicator, in line with recent work [29].

5. Experiments

We aim to identify which sampling strategies and configuration settings most efficiently enhance which parts of diversity and quality in conditional image generation,

We focus on class-conditional generation across 10 categories selected from the ImageNet-1K dataset,[8] a subset of ImageNet. This allows us to generalize across semantically diverse yet controlled categories.

5.1. Setup

All experiments are conducted using the SiT-XL/2 checkpoint,[9] trained on 256×256 ImageNet images [9]. SiT adopts the default configuration of DiT [43], but uses an optimized combination of diffusion coefficients. Specifically, SiT-XL employs a velocity-based diffusion model with a linear interpolant. Details on this setup are given in [35].

Since we focus on the effects of sampling, no model training is performed. We retain the original model configuration, including the exponential moving average (EMA) of model weights with a decay rate of 0.9999, as provided by the authors.

All samples are generated using 250 diffusion steps, following SiT's default setting. Prior work [35] has shown that for ODE solvers, FID tends to converge after roughly 60 steps; however, SDE sampling continues to improve sample quality and diversity significantly up to 250 steps.

[5] https://github.com/openai/guided-diffusion/tree/main/evaluations

[6] https://huggingface.co/facebook/dinov2-base

[7] https://huggingface.co/openai/clip-vit-base-patch32

[8] https://github.com/Jasonlee1995/ImageNet-1K

[9] https://github.com/willisma/SiT

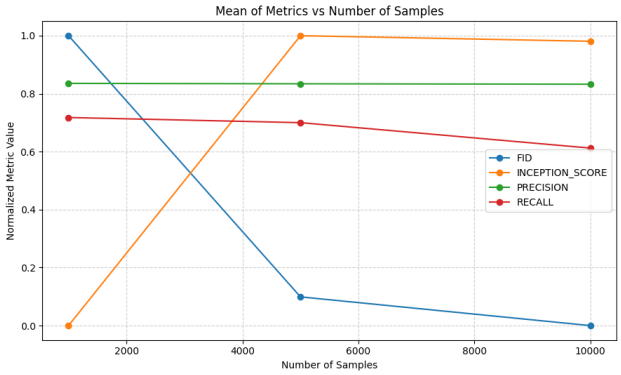

Figure 2. Mean values of FID, IS, Precision, and Recall across varying sample sizes.

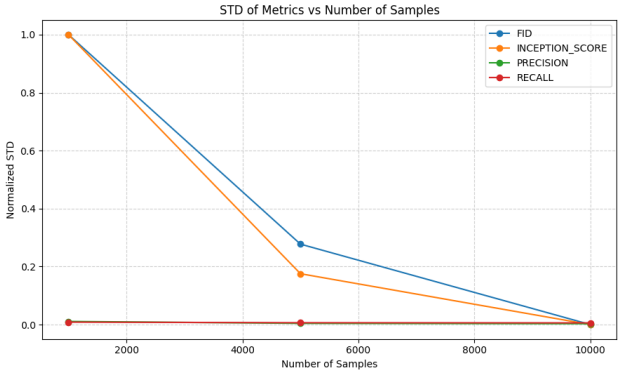

Figure 3. Standard deviation of FID, IS, Precision, and Recall across sample sizes.

5.2. Preliminary Analysis: Sample Size Sensitivity

We begin by analyzing the impact of sample size on metric stability and reliability, using SiT's default parameters.

Figure 2 shows that both FID and IS improve substantially from 1k to 5k samples, with only marginal gains beyond that. IS slightly declines after 5k, likely due to increased sampling noise. Figure 3 further reveals that the standard deviation of FID and IS is high at 1k samples (STD ≈ 1.0), but drops to 0.2–0.3 at 5k, and becomes negligible at 10k. These trends align with prior work recommending larger sample sizes (10k or 50k) for reliable FID computation [12, 26, 28, 34]. In contrast, Precision and Recall remain remarkably stable across all sample sizes. Both show low variance, and while Recall exhibits a slight decline at higher sample counts, these metrics are suitable even for small-scale evaluations. These findings suggest that a sample size of 5k offers a reasonable trade-off be-

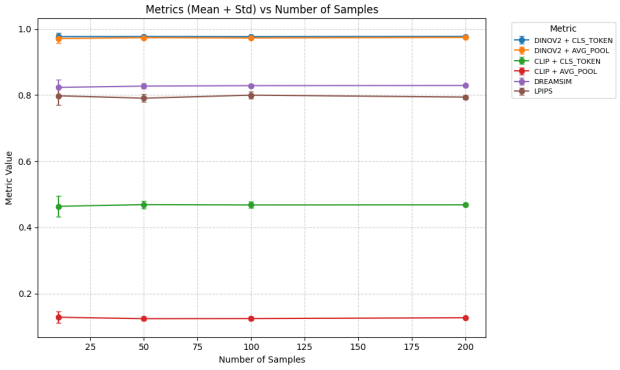

Figure 4. Mean and standard deviation of DINOv2, CLIP, DreamSim, and LPIPS across sample sizes. For DINOv2 and CLIP, pooling is done via either [CLS] token or spatial average.

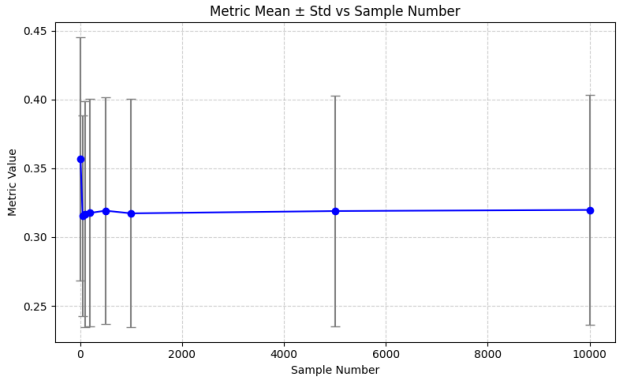

Figure 5. Mean and standard deviation of validation loss across sample sizes.

tween metric reliability and computational cost. Following [29], we adopt 5k as the default sample size for all subsequent experiments involving FID, IS, Precision, and Recall.

Figure 4 examines the effect of sample size on diversity metrics. DINOv2, CLIP, DreamSim, and LPIPS all exhibit low variance, even at small scales (10–200 samples). Minor instabilities for CLIP and LPIPS below 50 samples likely arise from batch-dependent feature variability. For DINOv2, representations based on the `[CLS]` token and spatial average pooling yield nearly identical results, with a slight advantage for `[CLS]`. Also CLIP shows that average pooling results in consistently lower diversity scores than `[CLS]` features. Therefore, we use the `[CLS]` approach in following experiments. Given their sensitivity, all four metrics are suited for evaluations under limited sampling budgets. We proceed with 200 samples per class in subsequent diversity analyses, as it strikes a good balance between efficiency and stability.

We also examine the sensitivity of the validation loss to

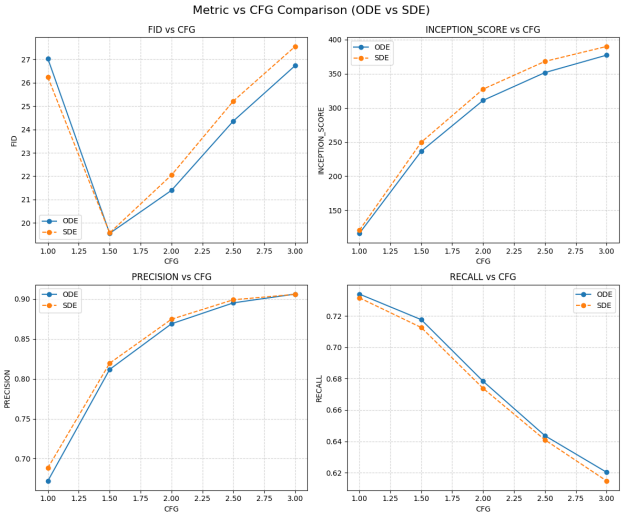

Figure 6. FID, IS, Precision, and Recall across different CFG scales for both ODE and SDE sampling.

sample size. Following [35], we segment the sampling trajectory into 8 equidistant timesteps $t \in (0, 1)$ and compute the loss at each, averaging across all but the final segment ($t = 1$) to reduce noise and instability in the signal. As shown in Figure 5, both the mean and standard deviation of validation loss fluctuate slightly at small sample sizes, indicating unstable estimates. From 100 samples onward, the mean stabilizes around 0.320 and the standard deviation converges to approximately 0.083. Increasing sample size beyond this point yields no improvements.

5.3. Sampling Configurations

We begin by evaluating the influence of sampling design on quality and diversity, starting with a joint comparison of ODE/SDE sampling. Instead of isolating these components, we analyze their interaction directly in early configuration experiments.

5.3.1 CFG Values

Figure 6 shows that the optimal CFG value varies across metrics. FID follows a U-shaped curve, reaching its minimum at CFG = 1.5, consistent with previous findings [10, 35, 48]. By contrast, IS and Precision increase with CFG, reflecting improved prompt alignment and visual fidelity [10, 38, 48]. Meanwhile, Recall decreases with increasing guidance, signaling reduced coverage of the training distribution and confirming the well-known quality–diversity trade-off [10,38,48,57]. Other trade-offs, such as FID vs. IS, have also been reported in [4, 57], and Precision vs. Recall in [29]. Interestingly, CFG = 1.5 emerges as an inflection point only for FID, but differs from the monotonic trends of the other metrics.

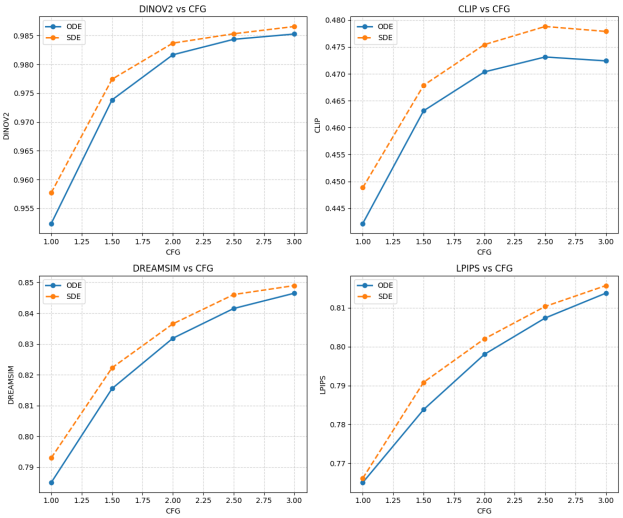

Figure 7. DreamSim, LPIPS, CLIP, and DINOv2 (CLS token) across CFG scales for both ODE and SDE sampling.

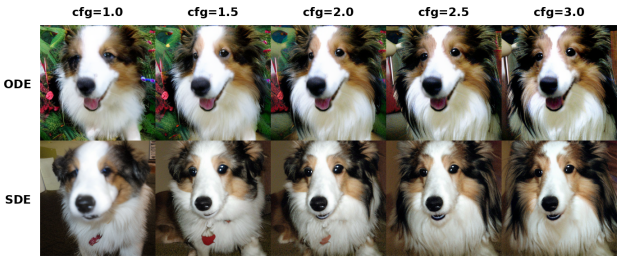

Figure 8. Qualitative comparison of ODE and SDE samples across increasing CFG values.

Figure 7 presents a different picture: DINOv2, CLIP, DreamSim, and LPIPS all increase with higher CFG. At first glance, this contrasts with Recall and literature, which report reduced diversity at stronger guidance. However, all four metrics (except CLIP, which saturates near CFG 3.0) continue to rise, suggesting a more nuanced view of diversity as assumed in prior work.

Rather than indicating disagreement among metrics, these trends reveal distinct *diversity dimensions*. Recall captures distributional coverage (i.e., how well generated data spans the real distribution), while LPIPS measures perceptual diversity, and CLIP, DreamSim, and DINOv2 capture higher-level semantic diversity. These latter metrics are sensitive to variation in attributes such as pose, background, color, or composition-dimensions that may increase even as overall distributional spread decreases.

This divergence highlights a key gap in prior work, which has largely relied on Recall or FID alone. Our findings emphasize the need to evaluate multiple facets of diversity, particularly when exploring the effects of CFG. Similar to findings in [29], we observe that different metrics peak under different configurations, suggesting that there is not just a straightforward negative trade-off between quality and diversity with increasing CFG, but rather more of a multidimensional view.

Figure 8 provides qualitative support for the above trends. As CFG increases, generated images exhibit sharper details and improved realism, particularly in facial features and fur texture. Saturation and contrast also improve with stronger guidance, especially in fur and eye regions.

5.3.2 ODE vs. SDE

Across all CFG values, SDE consistently outperforms ODE in both metrics. This is likely due to the beneficial role of stochasticity during sampling, which helps maintain variability in outputs. This finding aligns with previous work showing improved performance of SDE-based sampling for the SiT model [35]. Based on these results, we use primarily SDE sampling in the following experiments, if not indicated differently.

Figure 8 shows that ODE outputs appear more consistent and prompt-aligned, although at the cost of reduced background and color variation. In contrast, SDE samples preserve more diversity in pose and style, even at higher CFG values.

5.3.3 SDE Formulation and Norm Strength

Figure 9 compares three SDE diffusion coefficient schedules *constant*, *increasing–decreasing*, and *sigma* across four norm strengths: 1.0, 2.0, 3.0, and 4.0. It shows that the *constant* schedule consistently underperforms the others, and its performance deteriorates further as the norm strength increases. This suggests that unmodulated noise injection leads to excessive perturbations, degrading both quality and diversity, especially at higher magnitudes.

In contrast, the *increasing–decreasing* and *sigma-based* schedules remain robust across all norm values. These adaptive formulations support high sample quality while sustaining or even enhancing diversity metrics. Specifically, the *increasing–decreasing* schedule achieves the highest LPIPS scores, indicating strong perceptual diversity. Meanwhile, the *sigma* schedule yields the best performance on CLIP and DreamSim, suggesting improved semantic variation. Metric fluctuations are minor across norm values, further highlighting the stability of these schedules.

These findings are supported by qualitative results in Figure 10. The *constant* schedule produces visibly degraded outputs, with noticeable blurring, distortion, and structural artifacts that worsen with norm strength, particularly beyond norm 2.0. In contrast, samples generated using *increasing–decreasing* and *sigma* schedules remain visually coherent and detailed across all norm levels.

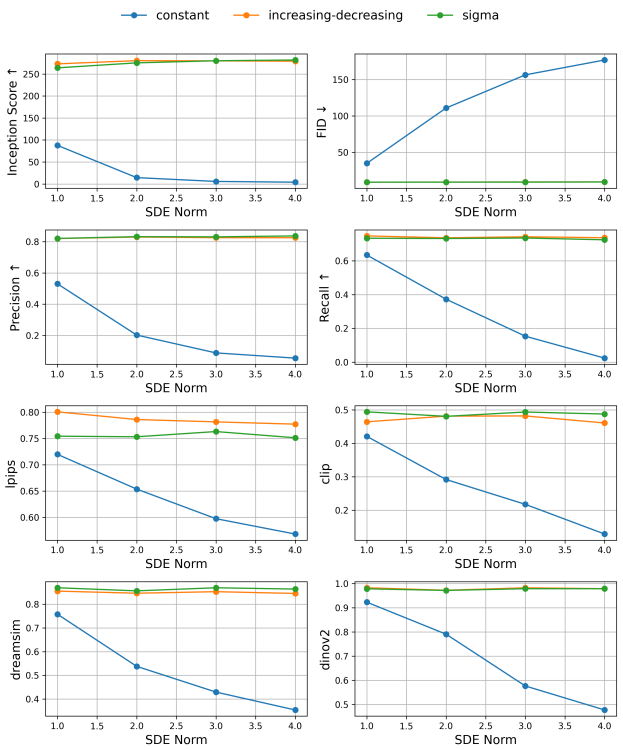

Figure 9. Effect of different SDE noise schedules and norm strengths on quality and diversity metrics.

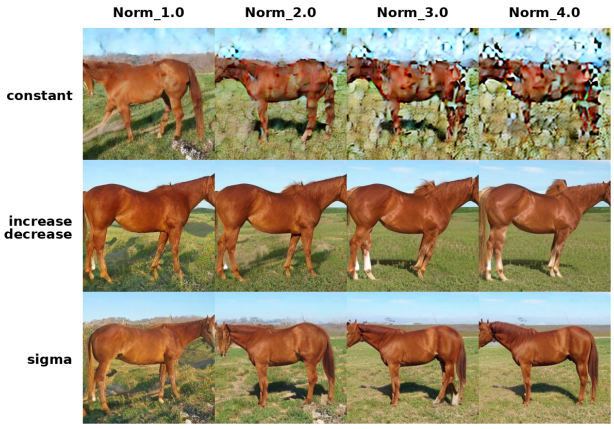

Figure 10. Qualitative comparison of SDE samples under different noise schedules and norm strengths.

These results confirm that adaptive normalization is essential for stable and diverse SDE sampling. Unlike the *constant* schedule, *increasing–decreasing* and *sigma* formulations are resilient to norm scaling and allow for improved image quality without compromising diversity.

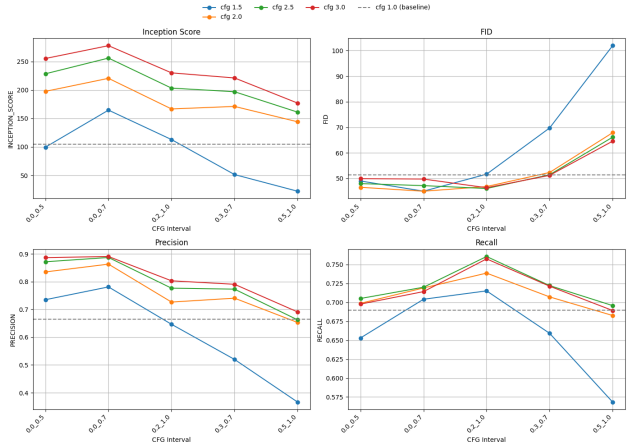

Figure 11. Effect of applying CFG during different sampling intervals at various strengths, evaluated on FID, IS, Precision, and Recall. The dashed line indicates the unguided baseline (CFG = 1.0).

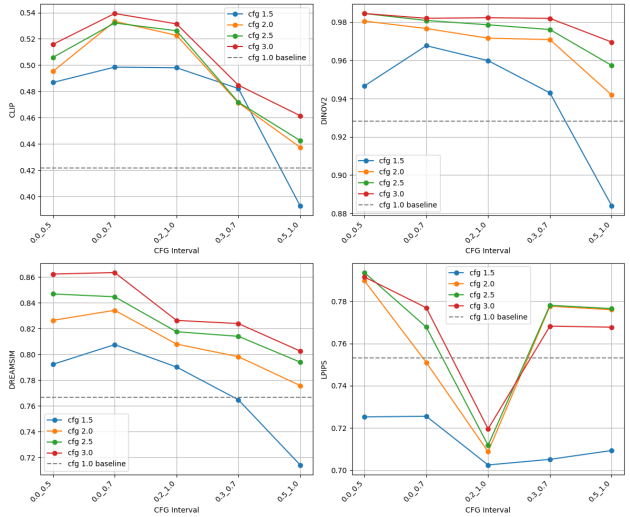

Figure 12. Effect of applying CFG during different sampling intervals at various strengths, evaluated on DreamSim, LPIPS, CLIP, and DINOv2. The dashed line shows the baseline without CFG applied.

5.3.4 CFG Interval

Previous work [29] suggests that CFG is most effective when applied in the middle of the diffusion process, with early guidance potentially harming diversity and late guidance being redundant. To expand this analysis more broadly, we evaluate four CFG strengths (1.5–3.0) across five sub-intervals of the sampling trajectory, and assess performance on both quality and diversity metrics. Results are shown in Figures 11 and 12, with each plot including a baseline of no CFG application. To contextualize these

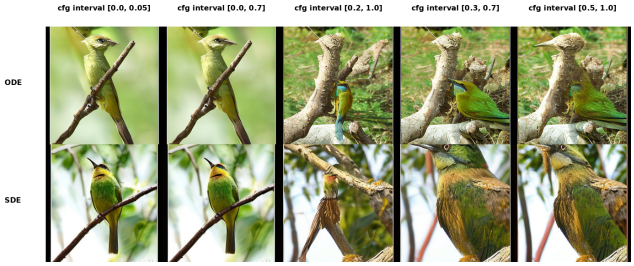

Figure 13. Qualitative comparison of ODE and SDE samples under different CFG intervals.

results, we compare them against our earlier findings (Figures 6 and 7) of CFG = 3.0 applied across the full interval, since it yielded the best performance under SDE sampling. In this comparison, only a few metrics improve under restricted guidance intervals. Specifically:

- **Recall** improves from 0.62 to 0.75 when guidance is applied over $[0.2, 1.0]$.
- **CLIP** increases from 0.48 to 0.54 over $[0.0, 0.7]$.
- **DreamSim** shows marginal gains from 0.85 to 0.86 in the $[0.0, 0.7]$ interval.

In contrast, quality-related metrics deteriorate under interval-restricted guidance. FID worsens significantly, rising from 19.5 to over 45, while IS drops from 390 to 220. Precision (0.92 to 0.90), DINOv2 (0.99 to 0.98, and LPIPS (0.82 to 0.79) also show slight but not significant decreases under interval-restricted guidance. Notably, the worst performance across nearly all metrics occurs when guidance is applied only at the end of sampling ($[0.5, 1.0]$), indicating that late guidance harms image quality the most. Figure 13 visualizes these observations: while differences between ODE and SDE samples are subtle for many CFG intervals, later guidance (e.g., $[0.2, 1.0]$, $[0.5, 1.0]$) visibly degrades structural coherence: birds appear distorted or unnaturally fused with their backgrounds. These results support the metric-based findings, as well as the hypothesis from [29], that late-stage guidance may not only be redundant but actively harmful for sample fidelity.

These results suggest that interval-restricted CFG can enhance certain forms of diversity, particularly training distribution coverage (Recall) and perceptual/semantic diversity (CLIP, DreamSim). However, these gains may come at the cost of sample fidelity and realism, as reflected in the drop in IS and FID. Still, almost all interval-restricted CFG configurations outperform the unguided baseline (CFG = 1.0), indicating that selective guidance is more beneficial than applying no guidance at all.

While our findings partially align with [29], which identified $[0.3, 0.7]$ as optimal for FID, we observe that early-to-mid intervals better preserve or enhance diversity metrics. This discrepancy may arise from differences in model

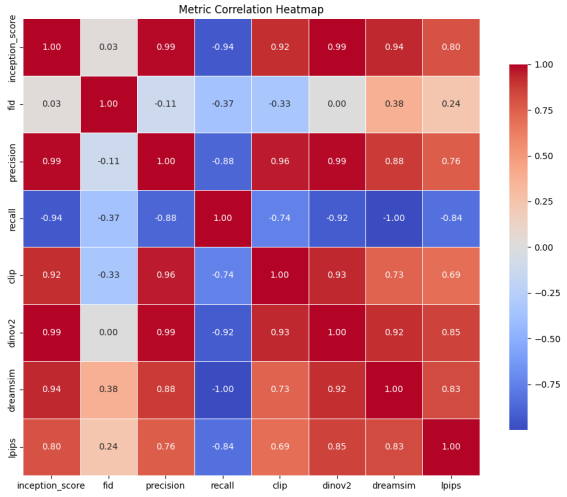

Figure 14. Pearson correlation between all metrics.

architecture, sampling method (ODE vs. SDE), or evaluation scope. Moreover, whereas their study used a single CFG value and focused solely on FID, we explore multiple CFG scales and a wider metric set. Further investigation is needed to validate these findings.

5.4. Quality–Diversity Trade-Off

To analyze the interaction between different evaluation metrics, we compute the Pearson correlation coefficients, visualized in Figure 14. The strength of relationships is interpreted following standard guidelines from [36].

Within quality metrics, IS and Precision exhibit near-perfect correlation ($\rho = 0.99$), indicating highly consistent relationships. However, both metrics show only weak correlation with FID: IS correlates at $\rho = 0.03$ and Precision at $\rho = -0.11$. This suggests that FID behaves quite differently from the other two metrics in our setting. These findings raise doubts about FID's alignment with standard notions of perceptual quality and support its reconsideration as a possible indicator of another diversity dimensions. For example, FID moderately correlates with DreamSim ($\rho = 0.38$). We conclude that relying solely on FID can produce misleading conclusions.

Within diversity metrics, strong positive correlations are found among DINOv2, DreamSim, and LPIPS (e.g., DINOv2 & DreamSim: $\rho = 0.92$; DreamSim & LPIPS: $\rho = 0.83$), indicating that these metrics capture similar semantic and perceptual diversity signals. CLIP also correlates well with these (e.g., CLIP & DINOv2: $\rho = 0.73$), reinforcing its role as a semantic diversity measure.

In contrast, Recall is negatively correlated with all other diversity metrics (e.g., Recall & DreamSim: $\rho = -1.00$), suggesting that it captures an orthogonal notion of diversity, specifically, distributional coverage rather than perceptual

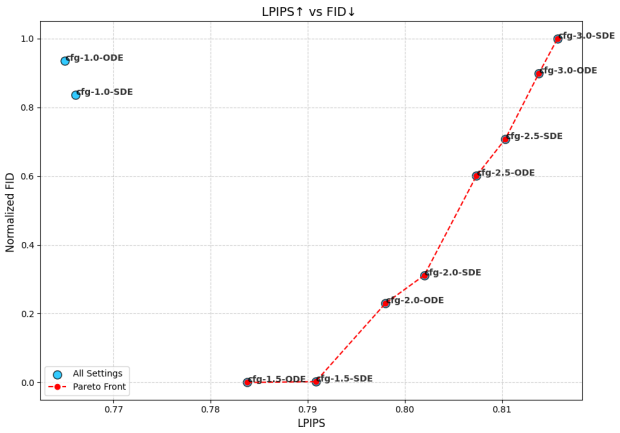

Figure 15. Trade-off between sample fidelity (FID ↓) and perceptual diversity (LPIPS ↑) across all sampling configurations. The Pareto front in red reveals optimal trade-off settings.

or semantic variety. This divergence supports the need to treat Recall as a fundamentally different dimension of diversity.

Across quality and diversity metrics, we observe an interesting bridge: IS correlates strongly with DINOv2 ($\rho = 0.99$), DreamSim ($\rho = 0.94$), and LPIPS ($\rho = 0.80$), and negatively with Recall ($\rho = -0.94$). Precision follows a similar pattern. These results challenge the common assumption of a strict quality–diversity trade-off: in our setting, perceived quality improves alongside semantic and perceptual diversity. Only Recall consistently moves in the opposite direction. Well-known trade-offs from prior work, such as Precision vs. Recall [10, 29] and IS vs. FID [4, 57], are also reflected here.

These findings reinforce patterns observed throughout this paper, especially from Section 5.3.1. IS, Precision, and semantic diversity metrics (e.g., DINOv2, DreamSim) tend to improve with CFG strength, while Recall consistently penalizes the same configurations, signaling increased risk of mode collapse. FID, meanwhile, shows weak alignment with all other metrics and should be interpreted with caution, especially at lower sample sizes, as discussed in Section 5.2. Overall, these results highlight the need for a nuanced, goal-dependent evaluation strategy. Metrics such as CLIP or LPIPS do not align clearly with either quality or coverage but capture perceptual variations in between. Relying on any single metric may obscure key trade-offs, reinforcing our recommendation to adopt a multi-metric approach based on the specific evaluation objective.

To further investigate one specific metric interaction in detail, we illustrated LPIPS vs. FID in Figure 15. The Pareto front (dotted red line) reveals a clear trade-off: settings that improve perceptual diversity (high LPIPS) tend to worsen FID (high FID). Notably, SDE sampling consistently achieves higher LPIPS across configurations, and increasing CFG further improves LPIPS at the cost of FID, replicating trends discussed in Section 5.3.1.

This final trade-off plot highlights three core insights from our study:

- No single configuration optimizes all metrics simultaneously; trade-offs are inevitable.
- Diversity and quality are multifaceted, requiring multiple, complementary metrics to evaluate meaningfully.
- Extensive exploration of configurations (e.g., sampling strategy, CFG scale, noise schedule) is essential, as no single setting is universally superior.

6. Implications

Diversity is not one-dimensional. The assumption that CFG harms diversity is only partially supported: while Recall, as a proxy for distributional coverage, decreases with increasing guidance, metrics capturing perceptual (LPIPS) and semantic (CLIP, DreamSim, DINOv2) diversity improve consistently. This suggests that guidance does not simply reduce variability but reshapes it, from global mode coverage toward finer-grained within-class variety. Thus, CFG may even act as a diversity shaper, not merely a restrictor. This reframes the quality–diversity trade-off as a redistribution across diversity dimensions.

CFG schedules act as implicit inductive biases. Applying guidance during early diffusion steps influences not only realism but semantic convergence: we observed that early-to-mid CFG (e.g., $[0.0, 0.7]$) improves semantic alignment without over-regularizing outputs. Late-stage guidance, by contrast, appears to constrain only local structure and leads to unnatural compositions.

Flow-matching solvers like SiT interact strongly with guidance. Compared to prior findings in diffusion models, the poor performance of FID under interval-based guidance in SiT reveals a key insight: flow-matching models may rely more heavily on continuous conditioning across the trajectory. The stronger impact of late-step omission in SiT vs. standard DDPMs hints at tighter coupling between the interpolant trajectory and conditioning signal.

Semantic diversity and realism may not trade off in the right configuration. Strong positive correlations between IS, DreamSim, and CLIP challenge the assumption that diversity must come at the cost of realism. In our experiments, the highest-performing settings on perceptual diversity metrics also exhibited strong Precision and IS scores. Hence, the quality–diversity trade-off is not universal, but depends on the choice of diversity metric and guidance regime. This

suggests that under certain configurations, semantic diversity and visual fidelity are mutually reinforcing rather than antagonistic.

These findings further highlight the importance of aligning evaluation strategies with task goals. If broad coverage of the training distribution is desired, excessive guidance may hinder performance. In contrast, when perceptual richness or semantic variation is prioritized, stronger guidance can be advantageous. Thus, careful metric selection is essential, and multiple perspectives should be considered to capture the full complexity of generative diversity.

6.1. Limitations

Our study is based on a single pretrained checkpoint, SiT-XL/2. While SiT is representative of recent flow-matching diffusion models, the results may not generalize to other architectures, such as DiT, ADM, or SDXL, especially those trained under different denoising objectives (e.g., noise prediction vs. velocity prediction).

Regarding sampling configurations, several improvements could further optimize performance. For example, the fixed number of 250 sampling steps may be too high for this setup, potentially leading to over-smoothing and degraded output. A reduced step count (e.g., 100 or fewer) could yield similar results with lower computational cost and is worth exploring in follow-up studies. Also, our experiments are restricted to class-conditional image generation on 10-classes. While this setup enables controlled comparisons, it may not reflect the behavior of CFG and sampling schedules in more complex generative tasks.

Additionally, our analysis focuses on three dimensions of diversity: distributional, perceptual, and semantic. Other forms of diversity are not explicitly captured. Metrics beyond those used in this study (see Section 2.5) could introduce new perspectives on model behavior. This work is intended as an initial step toward highlighting the multidimensional nature of diversity in generative models.

7. Conclusion

In this work, we investigated how sampling configurations influence diversity and quality in conditional image generation. Our results show that diversity is multidimensional: semantic, perceptual, and distributional aspects behave differently across settings. SDE sampling consistently outperforms ODE, and adaptive noise schedules (sigma, increasing–decreasing) yield better outcomes than constant schedules. Stronger CFG improves semantic and perceptual diversity but harms distributional coverage, a trade-off that can be mitigated via interval-restricted guidance. CFG should therefore not be seen purely as a constraint, but as a tool that reshapes diversity. Eventually, no single configuration is optimal: the best strategy depends on the target objective, reinforcing the need for goal-specific and multi-metric evaluation.

Future research should validate these findings across a broader range of models, also transferring to newer architectures such as PixArt [6] and Flux [30] for generalizability. Another promising direction is to explore diversity-enhancing techniques like diversity distillation [18], which applies the base model for early timesteps before transitioning to a distilled model. Understanding how such methods impact different diversity dimensions would offer interesting insights.

References

[1] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023. 2

[2] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022. 2

[3] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 1

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 7, 11

[5] Allen Chang, Matthew C Fontaine, Serena Booth, Maja J Matarić, and Stefanos Nikolaidis. Quality-diversity generative sampling for learning with synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19805–19812, 2024. 1, 5

[6] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 12

[7] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020. 2, 3

[8] Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102*, 2023. 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 5, 6

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 3, 4, 5, 6, 7, 11

[11] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 3

[12] Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3029–3039, 2025. 1, 4, 5, 6

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 4, 5

[14] Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024. 4, 5

[15] Koulischer Felix, Deleu Johannes, Demeester Thomas, and Ambrogioni Luca. Feedback guidance of diffusion models. *arXiv preprint arXiv:2506.06085*, 2025. 1, 3, 4

[16] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022. 3

[17] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 3

[18] Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. *arXiv preprint arXiv:2503.10637*, 2025. 12

[19] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018. 5

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 2, 3, 4

[22] Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1085–1094, 2022. 5

[23] Tim Kaiser, Nikolas Adaloglou, and Markus Kollmann. The unreasonable effectiveness of guidance for diffusion models. *arXiv preprint arXiv:2411.10257*, 2024. 1, 3

[24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 2

[25] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024. 1, 3

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5, 6

[27] Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, et al. Shielded diffusion: Generating novel and diverse images using sparse repellency. In *Forty-second International Conference on Machine Learning*. 1, 5, 6

[28] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 3, 4, 5, 6

[29] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Proc. NeurIPS*, 2024. 1, 3, 4, 5, 6, 7, 8, 9, 10, 11

[30] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 12

[31] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 5

[32] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[33] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024. 2

[34] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018. 5, 6

[35] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024. 2, 3, 6, 7, 8

[36] Natarajan Meghanathan. Assortativity analysis of real-world network graphs based on centrality metrics. *Comput. Inf. Sci.*, 9(3):7–25, 2016. 10

[37] Milvus. What is the difference between sampling diversity and sample fidelity?, 2025. 1

[38] Badr Moufad, Yazid Janati, Alain Durmus, Ahmed Ghorbel, Eric Moulines, and Jimmy Olsson. Conditional diffusion models with classifier-free gibbs-like guidance. *arXiv preprint arXiv:2505.21101*, 2025. 1, 3, 4, 6, 7

[39] Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023. 1, 4

[40] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021. 3

[41] Natsuo Okamoto, Seitaro Shinagawa, and Satoshi Nakamura. Image diversity evaluation metrics correlated with human subjectivity and prediction of image diversity in text-to-image synthesis. *Transactions of the Japanese Society for Artificial Intelligence*, 39(6):E–O35, 2024. 4

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 6

[43] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3, 5, 6

[44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[45] A Polyak, A Zohar, A Brown, A Tjandra, A Sinha, A Lee, A Vyas, B Shi, CY Ma, CY Chuang, et al. Movie gen: A cast of media foundation models, 2025. *URL https://arxiv.org/abs/2410.13720*, page 51, 2024. 4, 5

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 6

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[48] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023. 1, 3, 4, 5, 6, 7

[49] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 4

[50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[51] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[52] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 769–778, 2023. 1

[53] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021. 2

[54] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1

[55] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2

[56] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017. 1

[57] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023. 1, 3, 4, 5, 6, 7, 11

[58] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[59] Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pages 1–10. IEEE, 2020. 1

[60] Gerrit van den Burg and Chris Williams. On memorization in probabilistic deep generative models. *Advances in Neural Information Processing Systems*, 34:27916–27928, 2021. 1

[61] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[62] Mariia Zameshina, Olivier Teytaud, and Laurent Najman. Diverse diffusion: Enhancing image diversity in text-to-image generation. *arXiv preprint arXiv:2310.12583*, 2023. 5

[63] Marvin Zammit, Antonios Liapis, and Georgios N Yannakakis. Seeding diversity into ai art paper type. 2022. 4

[64] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 3

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 586–595, 2018. 3

[66] Weitong Zhang, Chengqi Zang, Liu Li, Sarah Cechnicka, Cheng Ouyang, and Bernhard Kainz. Stability and generalizability in sde diffusion models with measure-preserving dynamics. *Advances in Neural Information Processing Systems*, 37:81606–81644, 2024. 2, 5

[67] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017. 5