Guided Path Sampling: Steering Diffusion Models Back on Track with Principled Path Guidance

Anonymous Author(s)

Abstract

Iterative refinement methods based on a denoising-inversion cycle are powerful tools for enhancing the quality and control of diffusion models. However, their effectiveness is critically limited when combined with standard Classifier-Free Guidance (CFG). We identify a fundamental limitation: CFG's extrapolative nature systematically pushes the sampling path off the data manifold, causing the approximation error to diverge and undermining the refinement process. To address this, we propose Guided Path Sampling (GPS), a new paradigm for iterative refinement. GPS replaces unstable extrapolation with a principled, manifold-constrained interpolation, ensuring the sampling path remains on the data manifold. We theoretically prove that this correction transforms the error series from unbounded amplification to strictly bounded, guaranteeing stability. Furthermore, we devise an optimal scheduling strategy that dynamically adjusts guidance strength, aligning semantic injection with the model's natural coarse-to-fine generation process. Extensive experiments on modern backbones like SDXL and Hunyuan-DiT show that GPS outperforms existing methods in both perceptual quality and complex prompt adherence. For instance, GPS achieves a superior ImageReward of 0.79 and HPS v2 of 0.2995 on SDXL, while improving overall semantic alignment accuracy on GenEval to 57.45%. Our work establishes that path stability is a prerequisite for effective iterative refinement, and GPS provides a robust framework to achieve it.

Keywords

Text-to-image, Diffusion Models, Off-Manifold

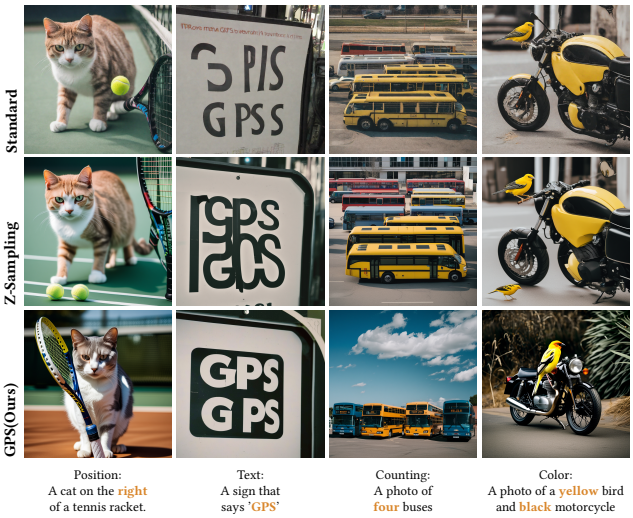

Figure 1: Visual comparison of GPS with baselines on challenging prompts. GPS demonstrates superior performance across tasks involving spatial positioning, text rendering, object counting, and attribute binding. Unlike Standard and Z-Sampling methods which suffer from artifacts or semantic misalignment, GPS maintains high fidelity and prompt adherence.

1 INTRODUCTION

Diffusion models have emerged as the dominant paradigm for high-fidelity signal generation [2, 16, 20, 22, 27], with their control largely specified through sampling techniques [17, 28]. The popular of these, which we term Z-sampling[1], leverages Classifier-Free Guidance (CFG)[11] to effectively steer generation. They found that the guidance gap between denoising and inversion could accumulate semantic information and the process of repeatedly applying an inversion-denoising cycle serves to maximize the integration of semantic information. While powerful, we find that Z-sampling suffers from a crucial "off-manifold" limitation [13]. Our analysis, supported by mathematical proof, shows that CFG's core "extrapolation" strategy causes guidance errors to systematically diverge, pushing the denoised estimate away from the true data manifold. This issue presents a primary bottleneck for achieving complex, detailed control [4, 9].

The consequences of this off-manifold problem are especially severe for advanced iterative refinement methods built upon the cycle. During Z-sampling, each CFG-guided denoising step introduces an error by pushing the estimate off-manifold. Because this estimate is no longer on the valid data distribution, the subsequent inversion step is inherently inaccurate [19]. This error is then fed back into the next iteration, where it is compounded by further CFG extrapolation . The continuous introduction and amplification of this error throughout the cycle not only causes visual artifacts and oversaturation but critically disrupts the invertibility of samplers like DDIM [28], ultimately rendering these powerful refinement techniques ineffective.

To counteract CFG's error divergence, one could theoretically use high-order solvers [12], but this is computationally inefficient. We therefore propose a fundamentally different and efficient paradigm, guided by the geometric imperative to remain on the data manifold: **G**uided **P**ath **S**ampling (GPS). At its core, GPS re-engineers the denoising-inversion cycle with a manifold-constrained interpolation mechanism [6], replacing unstable extrapolation. This elegantly resolves the cumulative error problem. Drawing further inspiration from the coarse-to-fine nature of diffusion [5], GPS dynamically schedules the guidance strength, which drastically reduces artifacts and enhances semantic fidelity.

Our contributions are twofold. **Theoretically**, we prove that GPS stabilizes iterative refinement by constraining the approximation error within a strictly bounded convex hull, thereby preventing divergence. We also show that a monotonically increasing cosine schedule for guidance strength better simulates cognitive refinement. **Experimentally**, GPS demonstrates marked superiority over

standard and Z-sampling methods across multiple benchmarks. On SDXL [20], it achieves a leading ImageReward of 0.79 and HPS v2 of 0.2995, significantly surpassing the Z-sampling baseline. This superiority extends to the transformer-based Hunyuan-DiT [15], where GPS sets a new benchmark with an ImageReward of 0.97. Furthermore, on GenEval [7], GPS improves overall prompt alignment accuracy to 57.45%, with notable gains in complex compositional tasks. These results validate GPS as a more robust and effective solution for high-fidelity, controllable generation [26].

2 RELATED WORKS

Off-Manifold problems in diffusion models Recent studies have shown that CFG can introduce systematic *off-manifold* errors during the denoising process. Specifically, the linear extrapolation step in CFG pushes intermediate estimates away from the true data manifold $\mathcal{M}$ [3]. Consequently, iterative refinement schemes that rely on a *denoising–inversion* cycle accumulate these deviations, yielding visible artifacts and violating the invertibility guarantees of deterministic samplers such as DDIM.

While interpreting the sampling process as solving an ordinary or stochastic differential equation (ODE/SDE) [29] can partially mitigate the drift, these solvers incur a significant computational overhead that is impractical for interactive editing or real-time applications. To address this limitation, manifold-preserving techniques have been proposed: (i) *geometric projection* methods that explicitly project back onto $\mathcal{M}$; (ii) *energy-guided* samplers that penalize off-manifold deviations with learned energy functions [18, 25]; and (iii) *shortcut* algorithms that re-parameterize the sampling path to remain within a learned latent subspace [8]. Despite these advances, existing approaches either require auxiliary networks or rely on costly optimization loops.

3 METHODOLOGY

In this section, we first analyze the origin of systematic error in iterative samplers, then introduce a manifold constraint to resolve it, and finally derive our method **GPS**, based on theoretical analysis.

3.1 Preliminaries & Definitions

To formally ground our analysis, we first establish the key preliminaries and definitions and a core assumption about the data manifold.

3.1.1 Denoising Diffusion Implicit Models. DDIM introduce a deterministic sampling process that is also invertible. This process is defined by a pair of single-step mappings: a denoising operation $\mathcal{D}$ and its exact inverse $\mathcal{I}$.

The denoising map $\mathcal{D}$ computes a less noisy data sample $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ as:

$$\mathbf{x}_{t-1} = \mathcal{D}(\mathbf{x}_t) = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\mathbf{x}_t^{\omega}}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\,\mathbf{x}_t^{\omega}$$

Conversely, the inversion map $\mathcal{I}$ reconstructs the noisier sample $\mathbf{x}_t$ from $\mathbf{x}_{t-1}$:

$$\tilde{\mathbf{x}}_t = \mathcal{I}(\mathbf{x}_{t-1}) = \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_{t-1}}}\mathbf{x}_{t-1} + \left(\sqrt{1-\bar{\alpha}_t} - \frac{\sqrt{\bar{\alpha}_t(1-\bar{\alpha}_{t-1})}}{\sqrt{\bar{\alpha}_{t-1}}}\right)\mathbf{x}_{t-1}^{\omega}$$

Here, $\mathbf{x}_t$ denotes the latent state at timestep t, while $\mathbf{x}_t^{\omega}$ represents the noise estimate predicted by the U-Net backbone [24] under the classifier-free guidance scale ω. The term $\bar{\alpha}_t$, defined as $\prod_{i=1}^{t}(1-\beta_i)$, characterizes the cumulative noise schedule, where β_i governs the variance of the Gaussian noise injected at each forward step.

3.1.2 Zigzag Sampling. For $t = T, \ldots, T-K$, alternate:

$$\textbf{Zig:}\quad \mathbf{x}_{t-1} = \mathcal{D}\big(\mathbf{x}_t \mid c, \omega_h\big),$$
$$\textbf{Zag:}\quad \tilde{\mathbf{x}}_t = \mathcal{I}\big(\mathbf{x}_{t-1} \mid c, \omega_l\big).$$

Z-sampling refines the sampling path by repeatedly alternating between a "Zig" step with a high guidance scale ω_h and a "Zag" step with a low guidance scale ω_l. This iterative process is applied for the initial timesteps before switching to a standard denoising procedure to obtain the final clean image.

Definition 3.1 (Semantic Information Gain). We define the *semantic information gain term* $\boldsymbol{\tau}_1(t)$ as the difference between the noise estimates:

$$\boldsymbol{\tau}_1(t) = \mathbf{x}_t - \tilde{\mathbf{x}}_t. \tag{1}$$

This gain is scheduled to be proportional to the difference in guidance scales, denoted by δ_ω, such that:

$$\boldsymbol{\tau}_1(t) \propto \delta_\omega \quad \text{where} \quad \delta_\omega = \omega_1 - \omega_2. \tag{2}$$

Definition 3.2 (Approximation Error and its Decomposition). The single-step approximation error $\boldsymbol{\tau}_2(t)$ in Z-Sampling is defined and decomposed as:

$$\boldsymbol{\tau}_2(t) = \tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{t-1} = \boldsymbol{\tau}_{\text{manifold}}(t) + \boldsymbol{\tau}_{\text{local}}(t). \tag{3}$$

The two components are defined as:

- **Local Discretization Error** ($\boldsymbol{\tau}_{\text{local}}$): The ideal error between two on-manifold points:
$$\boldsymbol{\tau}_{\text{local}}(t) := \tilde{\mathbf{x}}_t^{on} - \mathbf{x}_{t-1}^{on} \tag{4}$$
- **Systematic Manifold-Offset Error** ($\boldsymbol{\tau}_{\text{manifold}}$): The error induced by the guidance mechanism:
$$\boldsymbol{\tau}_{\text{manifold}}(t) := (\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_t^{on}) - (\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}^{on}) \tag{5}$$

Definition 3.3 (Guidance Mechanisms). We distinguish between two guidance mechanisms. **Off-Manifold Guidance** uses extrapolation ($\omega > 1$) , pushing the estimate $\mathbf{x}_t^{\omega}$ off the on-manifold path. In contrast, our **Manifold-Constrained Guidance** uses interpolation ($\lambda \in [0, 1]$), ensuring the estimate $\mathbf{x}_t^{\lambda}$ remains on the path. The respective estimates are:

$$\mathbf{x}_t^{\omega} = (1-\omega)\mathbf{x}_t^{\phi} + \omega\mathbf{x}_t^{c} \tag{6}$$
$$\mathbf{x}_t^{\lambda} = (1-\lambda)\mathbf{x}_t^{\phi} + \lambda\mathbf{x}_t^{c} \tag{7}$$

3.2 Guided Path Sampling

The cumulative effect of the approximation error $\boldsymbol{\tau}_2(t)$ directly determines the final image quality. We argue that standard CFG's use of *Off-Manifold Guidance* is the root cause of performance degradation in iterative samplers like Z-Sampling. This continuous off-manifold guidance introduces an ineliminable systematic error, ultimately causing the cumulative error series to diverge.

Algorithm 1 GPS

Input: Text prompt c, Denoising operation $\mathfrak{D}$, Inversion operation $\mathfrak{I}$, denoising guidance $\lambda_1 \in [0,1]$, inversion guidance scheduling function $\lambda_2(t)$, total inference steps T, self-reflection steps K.
Output: Clean image $\mathbf{x}_0$.
Sample Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$.
for $t = T$ **to** 1 **do**
  if $t > T - K$ **then**
    $\mathbf{x}'_{t-1} \leftarrow \mathfrak{D}(\mathbf{x}_t, c, \lambda_1)$
    $\lambda_{2,t} \leftarrow \lambda_2(t)$
    $\tilde{\mathbf{x}}_t \leftarrow \mathfrak{I}(\mathbf{x}'_{t-1}, c, \lambda_{2,t})$
    $\mathbf{x}_t \leftarrow \tilde{\mathbf{x}}_t$
  end if
  $\mathbf{x}_{t-1} \leftarrow \mathfrak{D}(\mathbf{x}_t, c, \lambda_1)$
end for
return $\mathbf{x}_0$

Theorem 3.4 (Error Divergence of Z-Sampling). *Assume:*

- *The noise prediction function is twice continuously differentiable near the data manifold.*

Then for Z-Sampling with CFG scale $\omega > 1$, the cumulative inversion error diverges:

$$\sum_{t=1}^{T} \|\boldsymbol{\tau}_2(t)\| \to \infty \quad \text{as } T \to \infty.$$

Proof. Let $\mathbf{d}(\mathbf{x}_t) := \mathbf{x}_t^c - \mathbf{x}_t^\phi$. At step t, the off-manifold perturbation magnitude is:

$$\|\boldsymbol{\delta}_{t-1}\| \approx \sqrt{\bar{\alpha}_{t-1}}\,(\omega - 1)\,\|\mathbf{d}(\mathbf{x}_t)\|.$$

Assuming the manifold has non-vanishing curvature represented by the tensor $\mathcal{H}$, the second-order manifold error satisfies:

$$\|\boldsymbol{\tau}_{\text{manifold}}(t)\| \approx \tfrac{1}{2}\|\mathcal{H}(\xi)[\boldsymbol{\delta}_{t-1}, \boldsymbol{\delta}_{t-1}]\| \geq \kappa \|\boldsymbol{\delta}_{t-1}\|^2,$$

where $\kappa > 0$ relates to the manifold curvature. Thus,

$$\|\boldsymbol{\tau}_{\text{manifold}}(t)\| \gtrsim \bar{\alpha}_{t-1}(\omega - 1)^2 \|\mathbf{d}(\mathbf{x}_t)\|^2.$$

Since the guidance term $\mathbf{d}(\mathbf{x}_t)$ represents a semantic direction independent of the step size Δt, this error term is $O(1)$ with respect to T. Consequently, summing this non-vanishing error over T steps leads to divergence:

$$\sum_{t=1}^{T} \|\boldsymbol{\tau}_2(t)\| \geq \sum_{t=1}^{T} c = \Omega(T) \to \infty.$$

$$\mathfrak{D}(\mathbf{x}_t) = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\mathbf{x}_t^\lambda}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\,\mathbf{x}_t^\phi \tag{8}$$

$$\mathfrak{I}(\mathbf{x}_{t-1}) = \sqrt{\bar{\alpha}_t}\left(\frac{\mathbf{x}_{t-1} - \sqrt{1-\bar{\alpha}_{t-1}}\,\mathbf{x}_{t-1}^\lambda}{\sqrt{\bar{\alpha}_{t-1}}}\right) + \sqrt{1-\bar{\alpha}_t}\,\mathbf{x}_{t-1}^\phi \tag{9}$$

To resolve the error divergence issue, we propose **Guided Path Sampling (GPS)**. Its core idea is to adopt the *Manifold-Constrained Guidance* defined in Section 3.1, which fundamentally eliminates the systematic manifold offset error. We then replace part of the guided noise with unconditional noise, as defined in equations $\mathfrak{D}(\mathbf{x}_t)$ and $\mathfrak{I}(\mathbf{x}_{t-1})$. The complete procedure is detailed in Algorithm 1. The stability of this approach is guaranteed by our first core theorem.

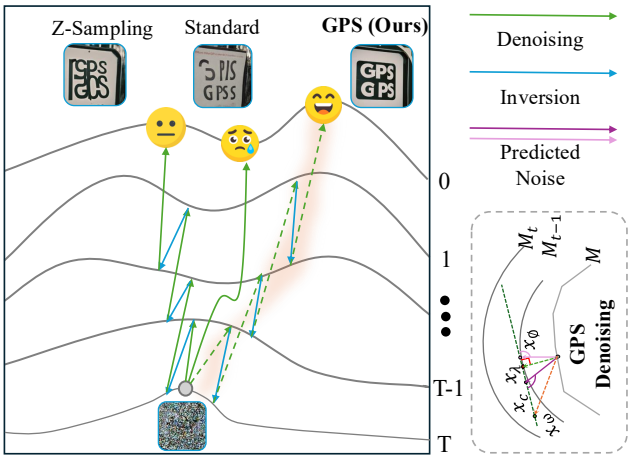

Figure 2: Schematic illustration of GPS. Unlike standard methods that extrapolate off the manifold (red dotted line), GPS employs a manifold-constrained interpolation (green solid line) during the zigzag cycle, ensuring the sampling trajectory remains stable and errors remain bounded.

Theorem 3.5 (Error Boundedness of GPS). *Let the noise predictions be bounded. For GPS employing manifold-constrained guidance (interpolation), the cumulative approximation error $\sum_{t=1}^{T} \|\boldsymbol{\tau}_2(t)\|$ is* ***strictly bounded****, ensuring sampling stability.*

Proof. Let $\mathbf{x}_t^\phi$ and $\mathbf{x}_t^c$ denote the unconditional and conditional noise predictions, respectively, assumed to be bounded in magnitude by a constant M.

Recall that standard CFG (Eq. 6) employs extrapolation with $\omega > 1$, which amplifies the deviation:

$$\|\mathbf{x}_t^\omega\| = \|(1-\omega)\mathbf{x}_t^\phi + \omega \mathbf{x}_t^c\| = \|\mathbf{x}_t^\phi + \omega(\mathbf{x}_t^c - \mathbf{x}_t^\phi)\|.$$

As ω increases, the norm $\|\mathbf{x}_t^\omega\|$ grows linearly with ω, potentially becoming unbounded and pushing the trajectory off-manifold.

In contrast, GPS (Eq. 7) employs interpolation with $\lambda \in [0,1]$:

$$\mathbf{x}_t^\lambda = (1-\lambda)\mathbf{x}_t^\phi + \lambda \mathbf{x}_t^c.$$

By the Triangle Inequality, the magnitude of the guided noise is strictly bounded by the convex hull of the component predictions:

$$\|\mathbf{x}_t^\lambda\| \leq (1-\lambda)\|\mathbf{x}_t^\phi\| + \lambda\|\mathbf{x}_t^c\| \leq \max(\|\mathbf{x}_t^\phi\|, \|\mathbf{x}_t^c\|) \leq M.$$

Consequently, the manifold offset error $\boldsymbol{\tau}_{\text{manifold}}(t)$ does not diverge. The total cumulative error is dominated by the local discretization error, which is bounded for the finite time horizon T:

$$\sum_{t=1}^{T} \|\boldsymbol{\tau}_2(t)\| \leq \sum_{t=1}^{T} (C \cdot \Delta t) = C \cdot T\Delta t = O(1).$$

Thus, the error series remains bounded, guaranteeing algorithmic stability.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our model on two complementary benchmarks. For assessing human-perceived **aesthetic quality**, we use the first 100 prompts from Pick-a-Pic [14]. For quantitatively measuring **compositional accuracy** (e.g., object count and position), we use GenEval. This dual evaluation provides a holistic view of our model's capabilities.

Metrics We evaluate text-image alignment using CLIP Score [10], and complement it with HPS v2 [30] and ImageReward (IR) [31], two learned metrics trained on extensive human preference judgments to capture subjective quality.

Diffusion Models We employ different diffusion models as the generation backbone in our experiments. For SD2.1 [23], SDXL [21], and Hunyuan-DiT [15], we perform 50 denoising steps. We set $\omega = 5.5$, $\lambda_1 = 0.5$ and use $\lambda_{2,t}$ that increases from 0.1 to 0.3 using a cosine function in SDXL/SD2.1, and $\omega = 6.0$ in Hunyuan-DiT, aligning with the default recommended values. Finally, the zigzag operation is executed along the entire path ($K = T - 1$).

4.2 Main Results

Table 1: Comparative results on the Pick-a-Pic benchmark.

	Method	CLIP ↑	HPS v2 ↑	IR ↑
SDXL	Standard	0.710	0.2899	0.64
	Z-Sampling	0.719	0.2980	0.75
	GPS	**0.723**	**0.2995**	**0.79**
SD-2.1	Standard	0.681	0.2541	-0.54
	Z-Sampling	0.696	0.2686	-0.25
	GPS	**0.702**	**0.2709**	**-0.18**
Hunyuan-DiT	Standard	0.712	0.2915	0.92
	Z-Sampling	0.724	0.3012	0.94
	GPS	**0.730**	**0.3056**	**0.97**

Table 2: Comparative results on GenEval with SDXL.

Metric	Standard	Z-Sampling	GPS (ours)
Single Obj.	97.50%	**100.00%**	**100.00%**
Two Obj.	69.70%	74.75%	**76.77%**
Count.	33.75%	46.25%	**48.75%**
Colors	86.71%	**87.23%**	86.17%
Pos.	10.00%	10.00%	**11.00%**
Color Attr.	18.00%	**24.00%**	22.00%
Overall	52.52%	57.04%	**57.45%**

As presented in Table 1 and 2, GPS consistently outperforms both Standard sampling and Z-Sampling methods across varying benchmarks and model architectures.

On the Pick-a-Pic benchmark (Table 1), GPS demonstrates superior performance in text-image alignment (CLIP) and human preference metrics (HPS v2, IR). Notably, on the SDXL backbone, GPS achieves an ImageReward of **0.79** and HPS v2 of **0.2995**, surpassing the strong Z-Sampling baseline. This superiority extends to different architectures, including the older SD-2.1 and the Transformer-based Hunyuan-DiT, where GPS sets a new state-of-the-art with an ImageReward of **0.97**. These results indicate that maintaining the manifold structure effectively enhances both semantic fidelity and aesthetic quality.

Table 2 further details the fine-grained semantic capabilities on SDXL using the GenEval benchmark. GPS achieves the highest **Overall** score of **57.45%**. Crucially, significant gains are observed in complex compositional tasks, such as **Counting** (48.75% vs. 46.25% for Z-Sampling) and **Two Object** generation (76.77% vs. 74.75%). This suggests that our stable iterative refinement better accumulates semantic details and spatial structures for complex prompts, validating the effectiveness of our manifold constraints.

4.3 Ablation Study

Table 3: Ablation of the inversion scheduler $\lambda_{2,t}$ on SDXL.

Scheduler	CLIP ↑	HPS v2 ↑	IR ↑
Constant (0.1)	0.711	0.2983	0.72
Constant (0.3)	0.714	0.2986	0.73
Sigmoid (0.1→0.3)	0.719	0.2993	0.74
Linear (0.1→0.3)	0.721	0.2994	0.75
Cos (0.1→0.3)	**0.723**	**0.2995**	**0.79**
Cos (0.3→0.1)	0.717	0.2992	0.74
Cos (0.1→0.3→0.1)	0.710	0.2983	0.72

Our ablation study on the inversion guidance scheduler (Table 3) provides strong empirical backing for our theory. We evaluated various strategies for scheduling the guidance scale $\lambda_{2,t}$, ranging from fixed `Constant` values to dynamic schedules like `Cos (0.1→0.3)`. The results across all metrics (CLIP, HPS v2, and IR) reveal two key findings.

First, dynamic scheduling consistently surpasses constant schedules, with the `Cos (0.1→0.3)` strategy achieving the highest scores (e.g., **0.79** IR and **0.2995** HPS v2). Second, and most importantly, the results directly validate our proposed **"coarse-to-fine" refinement principle**: monotonically increasing schedules for $\lambda_{2,t}$ yield the best performance. This confirms that gradually strengthening the manifold constraint is the optimal strategy. Conversely, schedules that violate this principle by decreasing the scale (`Cos (0.3→0.1)`) or being non-monotonic (`Cos (0.1→0.3→0.1)`) fail to maintain this benefit, leading to a noticeable degradation in both semantic alignment and perceptual quality.

5 CONCLUSION

We identified that standard extrapolative CFG pushes sampling paths off-manifold, causing error divergence. We introduced GPS, which replaces extrapolation with manifold-constrained interpolation, transforming the divergent error of methods like Z-Sampling into a provably convergent process. Furthermore, we proposed an optimal, monotonically increasing guidance schedule to align semantic injection with the model's coarse-to-fine generation . Our

experiments show GPS significantly improves perceptual quality and semantic alignment. **The key takeaway is that path stability is a prerequisite for effective iterative refinement.** Future work will focus on extending GPS to stochastic samplers and exploring learned scheduling functions.

References

[1] Lichen Bai et al. 2024. Zigzag Diffusion Sampling: Diffusion Models Can Self-Improve via Self-Reflection. *arXiv preprint arXiv:2412.10891* (2024).

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Yam D, Dominik andSV, and Robin Rombach. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127* (2023).

[3] Huanran Chen et al. 2025. Your Diffusion Model is Secretly a Certifiably Robust Classifier. *arXiv preprint arXiv:2402.02316* (2025).

[4] Minghao Chen et al. 2023. Training-Free Layout Control with Cross-Attention Guidance. *arXiv preprint arXiv:2304.03373* (2023).

[5] Jooyoung Choi, Jungbeom Kim, Hunho Myeong, Seyoung Jeong, Youngjung Ko, and Sungroh Choi. 2022. Perception Prioritized Training of Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[6] Hyungjin Chung et al. 2024. CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models. *arXiv preprint arXiv:2406.08070* (2024).

[7] Dhruba Ghosh et al. 2023. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. *arXiv preprint arXiv:2310.11513* (2023).

[8] Yutong He et al. 2024. Manifold Preserving Guided Diffusion. In *International Conference on Learning Representations (ICLR).*

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626* (2022).

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv:2104.08718 [cs.CV] https://arxiv.org/abs/2104.08718

[11] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.*

[12] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems (NeurIPS).*

[13] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793* (2022).

[14] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.*

[15] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. 2024. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. arXiv:2405.08748 [cs.CV] https://arxiv.org/abs/2405.08748

[16] Shanchuan Lin, Anran Wang, and Hongxiang Yang. 2024. SDXL-Lightning: Progressive Adversarial Diffusion Distillation. *arXiv preprint arXiv:2402.13929* (2024).

[17] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In *Advances in Neural Information Processing Systems (NeurIPS).*

[18] Morteza Mardani et al. 2023. Guided diffusion as likelihood-based energy. In *International Conference on Machine Learning (ICML).*

[19] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[20] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV] https://arxiv.org/abs/2307.01952

[22] Zipeng Qi et al. 2024. Layered Rendering Diffusion Model for Controllable Zero-Shot Image Synthesis. *arXiv preprint arXiv:2311.18435* (2024).

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] https://arxiv.org/abs/2112.10752

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI).*

[25] Lior Rout et al. 2023. Perfusion: A parameter-efficient and generalizable approach for single-image text-to-image personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS).*

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML).* 2256–2265.

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations (ICLR).*

[29] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR).*

[30] Xiaoshi Wu et al. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv:2306.09341* (2023).

[31] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. arXiv:2304.05977 [cs.CV] https://arxiv.org/abs/2304.05977