Not All Prediction Targets Keep Training-Free Diffusion Guidance on the Manifold

Yunsung Lee[1⋆] and Hyeongmin Lee[2⋆†]

[1] MAUM.AI, Republic of Korea
sung@maum.ai
[2] Seoul National University of Science and Technology, Republic of Korea
hyeongmin.lee@seoultech.ac.kr

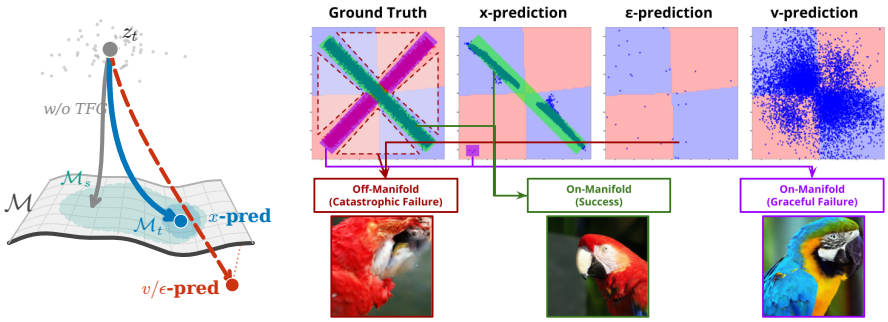
Abstract. Training-free guidance (TFG) steers a pretrained diffusion model toward a desired attribute at inference. To be effective, this guidance must be applied from the earliest, high-noise steps of sampling. Because its objective (a classifier or energy) is defined on clean images, ϵ- and v-prediction models must first estimate the clean image $\hat{x}$ from the noisy state at each step, and the accuracy of that estimate determines how easily guidance drifts off the data manifold. x-prediction, a recent alternative, outputs the clean image directly, removing this source of error even at high noise. This is our motivation. We provide a theoretical analysis of how each prediction target shapes this accuracy, and introduce guided-class FID (Child FID), a metric that exposes the manifold damage standard evaluation misses. Experiments on a new fine-grained bird benchmark and on style transfer confirm that x-prediction keeps guided samples on the manifold most reliably, making it the strongest foundation for training-free guidance. Code is available at https://github.com/ManLuML/on-manifold-tfg.

Keywords: Diffusion Models · Training-Free Guidance · Prediction Target · Flow Matching · Inference-Time Guidance

1 Introduction

Training-free guidance (TFG) [5, 34, 38] steers diffusion models [13, 31] toward desired properties without retraining, but strong guidance can push samples off the data manifold. The resulting *catastrophic failures* (collapsed, distorted images) are qualitatively different from a *graceful failure*, in which guidance misses the target class but the image remains a realistic sample from the learned distribution. A model whose worst case is graceful failure is fundamentally more dependable: even when guidance errs, the basic contract of generative modeling, producing plausible images, is preserved.

⋆ Equal contribution.
† Corresponding author.



(a) Guidance schematic **(b)** Crossed-lines (D=512) and ImageNet examples

Fig. 1: Prediction target determines whether training-free guidance stays on-manifold. (a) Without guidance, all targets denoise z_t onto source class $\mathcal{M}_s \subset \mathcal{M}$. With TFG, x-prediction slides along $\mathcal{M}$ to target class $\mathcal{M}_t$; v/ϵ-prediction departs $\mathcal{M}$. **(b)** Crossed-lines (D=512, top) and ImageNet (bottom): x-prediction preserves structure; ϵ-prediction collapses off-manifold with catastrophic artifacts.

Yet this distinction has gone unnoticed. Standard Validity (top-1 accuracy) rewards any sample the classifier accepts, whether on- or off-manifold, a blind spot shared by 15 of 17 recent TFG papers [33] (Appendix D). When prior work maximises Validity under strong guidance [38], it unknowingly selects off-manifold images fooling classifiers (akin to adversarial perturbations [24, 36]) rather than diverse samples of the target class. The evaluation does not merely fail to detect manifold departure; it actively encourages it.

We trace the difference between catastrophic and graceful failure to a design decision predating any guidance algorithm: the prediction target. Three targets have been proposed: ϵ-prediction (noise) [25], v-prediction (velocity) [3, 21], and x-prediction (clean data) [18]. Guidance operates on the clean-image estimate $\hat{x}$: ϵ- and v-prediction must recover it from the noisy state, whereas x-prediction outputs it directly. Under identical architecture and training, the three targets produce comparable generation quality, yet differ fundamentally in manifold preservation [18]: x-prediction succeeds where ϵ-prediction fails catastrophically. Because TFG computes its guidance gradient through this estimate ($\nabla_{z_t}\mathcal{E}(\hat{x})$), the fidelity of $\hat{x}$ directly controls guidance quality. We ask: *does the prediction target's influence on manifold quality, demonstrated at training time [18], extend to inference-time guidance?*

We prove that prediction targets create a strict hierarchy of error amplification (Proposition 1). The mechanism is the recovery formula: ϵ-prediction divides by t, amplifying errors by $(1-t)/t$, a factor that diverges at high noise ($t \to 0$). In contrast, v-prediction attenuates errors by a bounded $(1-t)$ factor, and x-prediction introduces no amplification at all. These per-step errors compound across sampling, causing ϵ-prediction's cumulative perturbation to diverge while x-prediction's remains bounded (Proposition 2). The gap further widens with

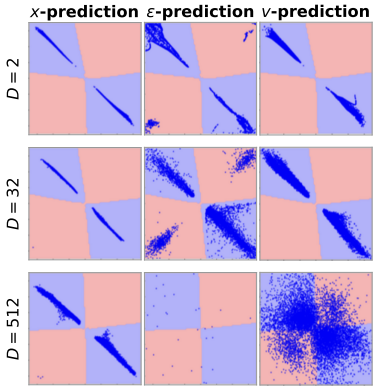


Fig. 2: Crossed-lines guided generation (s=10, 100 steps). ϵ-prediction collapses by D=32. Full grid in Appendix G.

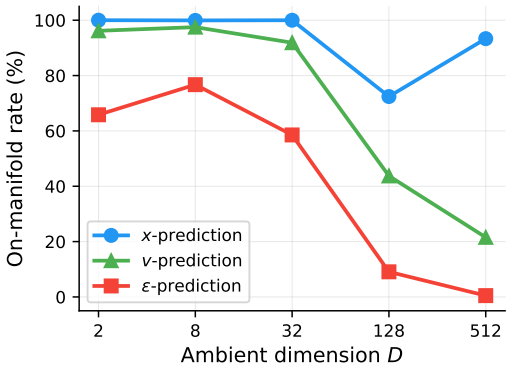


Fig. 3: On-manifold rate vs. ambient dimension (s=10). x-prediction holds >93% at D=512; v-prediction degrades to 21.5%; ϵ-prediction drops to 0.5%. Data from Tab. 10.

ambient dimension [15, 16] (Appendix A). Controlled ablations confirm this hierarchy: in crossed-lines experiments (identical architecture and training, varying only the prediction target across $D \in \{2, 8, 32, 128, 512\}$), x-prediction maintains high on-manifold rate while ϵ-prediction collapses (Fig. 1, Figs. 2 and 3).

Validating this hierarchy at real scale demands three evaluation advances absent from prior TFG work: (i) a fine-grained benchmark separating guidance and evaluation classifiers, (ii) a manifold-aware metric, and (iii) guidance-strength sweep plots replacing single-point comparisons. We construct a 143-species bird classification benchmark on ImageNet 256×256 and introduce Child FID (C-FID), FID between guided samples and the target species domain, sweeping guidance strength ρ across Pareto frontiers (Fig. 4). We evaluate four pretrained Diffusion Transformers at comparable quality (FID≈2) across all three prediction targets [3,18,21,25]. At matched classifier accuracy (≈26.6%), C-FID reveals a 5.2-point gap between x- and ϵ-prediction (**32.9** vs. **38.1**), manifold damage invisible to standard evaluation. Qualitative analysis shows ϵ-prediction achieving Validity via classifier-friendly patterns rather than diverse samples (Fig. 7). PixelFlow (v-prediction, pixel space) isolates prediction target as the decisive variable: its C-FID *reverses* under strong guidance while JiT's continues decreasing (Sec. 5).

Across controlled ablations and ImageNet-scale experiments alike, x-prediction yields the most stable behavior among the three targets for inference-time guidance. Among JiT variants (B/L/H), larger models achieve strictly better guidance Pareto frontiers, a *guidance scaling* effect in which capacity improves both generation quality and guidance responsiveness. These findings establish prediction target selection as a first-order design decision for inference-time control.

Contributions. This paper makes three contributions:

1. **Theoretical framework.** We prove a strict error amplification hierarchy across prediction targets and show these errors compound into divergent trajectory perturbation for ϵ-prediction while x-prediction's cumulative error remains bounded (Propositions 1 and 2).
2. **Manifold-aware evaluation protocol.** We introduce *(i)* a fine-grained bird classification benchmark (143 species, separate guidance and evaluation classifiers), *(ii)* Child FID (C-FID) to measure within-class realism, and *(iii)* guidance-strength Pareto sweeps replacing single-point comparisons.
3. **Empirical validation.** In crossed-lines ablations, x-prediction maintains >93% on-manifold rate while ϵ-prediction collapses to <1%. On ImageNet at matched Validity, C-FID reveals a 5.2-point gap between x- and ϵ-prediction; PixelFlow's C-FID reversal confirms prediction target, not operating space, as the decisive factor (Sec. 5).

2 Related Work

Prediction Targets in Diffusion Models. DDPM [13] established ϵ-prediction as the default; score-based models [35] gave an equivalent view via Tweedie's formula [8,30]. Salimans and Ho [32] introduced v-prediction for improved stability. State-of-the-art Diffusion Transformers achieve comparable quality across all three targets: DiT-XL [25] (ϵ), SiT-XL [21] (v), and JiT-G [18] (x). Prior work observed ϵ-prediction's training-time error amplification [11,16]; Jin and Wang [15] independently confirm this from a dimensionality perspective. We extend these training-time observations to inference-time guidance (Proposition 1).

Training-Free Guidance and Off-Manifold Departure. DPS [5] and LGD [34] apply gradient-based guidance for inverse and general problems. TFG [38] unifies prior methods [2, 39] via seven hyperparameters controlling mean/variance guidance and recurrence. All methods depend on clean data estimates $\hat{x}$: guidance computes $\nabla_{z_t}\mathcal{E}(\hat{x})$, so the fidelity of $\hat{x}$ determines both guidance quality and sample realism. Strong guidance produces degraded, off-manifold samples akin to adversarial perturbations [24, 36]. Theoretically, nonzero score error enables strong guidance to push samples off the data support [4], a failure mode confirmed for CFG [6] and extending to training-free methods. Our analysis (Sec. 3.2) shows that the prediction target determines the accuracy of the manifold-restoring force.

Guidance for Flow Matching and Evaluation. Feng *et al.* [9] derive Flow Matching guidance via velocity-field modifications; we adopt TFG's post-step correction for its unified DDPM–Flow Matching interface. Standard FID and classifier accuracy cannot distinguish on-manifold success from adversarial artifacts [29,33]: a concern borne out across 17 surveyed papers, most lacking manifold-aware metrics (Appendix D). We address this with guided-class FID (Child FID) and guidance-strength Pareto sweeps (Appendix C).

3 Method

3.1 Preliminaries

Flow Matching Formulation. Following JiT [18], we adopt the Flow Matching formulation [1, 19, 20]: $z_t = t \cdot x + (1-t) \cdot \epsilon$ with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $t \in [0, 1]$, where t=1 corresponds to clean data and t=0 to pure noise.

Prediction Targets. Three prediction targets have been proposed:

- ϵ-prediction: The network predicts noise $\epsilon_\theta(z_t, t) \approx \epsilon$
- v-prediction: The network predicts velocity $v_\theta(z_t, t) \approx v = x - \epsilon$
- x-prediction: The network predicts clean data $x_\theta(z_t, t) \approx x$

Given each prediction, clean data can be recovered as $\hat{x}^{(\epsilon)} = (z_t - (1-t)\epsilon_\theta)/t$, $\hat{x}^{(v)} = z_t + (1-t)v_\theta$, and $\hat{x}^{(x)} = x_\theta$ directly. Each formula estimates the posterior mean $\mathbb{E}[x \mid z_t]$, the flow matching analogue of Tweedie's formula [8], but through parameterizations with fundamentally different numerical stability (Appendix A).

3.2 Error Propagation in Clean Data Estimation

We analyze how prediction errors propagate to clean data estimates, the quantity that determines whether guided trajectories remain on the data manifold.

Proposition 1 (Error Amplification). *Let $\delta_\epsilon = \|\epsilon - \epsilon_\theta\|_2$, $\delta_v = \|v - v_\theta\|_2$, and $\delta_x = \|x - x_\theta\|_2$ be prediction errors for each target. The error in recovered clean data is:*

$$\left\|\hat{x}^{(\epsilon)} - x\right\|_2 = \frac{1-t}{t}\delta_\epsilon \tag{1}$$

$$\left\|\hat{x}^{(v)} - x\right\|_2 = (1-t)\delta_v \tag{2}$$

$$\left\|\hat{x}^{(x)} - x\right\|_2 = \delta_x \tag{3}$$

Proof (Proof sketch). By direct substitution of recovery formulas into the forward process. For ϵ-prediction: $\hat{x}^{(\epsilon)} = x + \frac{1-t}{t}(\epsilon - \epsilon_\theta)$. For v- and x-prediction: analogous. Full proof in Appendix B.

As $t \to 0$, ϵ-prediction's amplification diverges while v- and x-prediction remain bounded: a strict hierarchy in how prediction errors propagate to clean-data estimates.

Cumulative Trajectory Divergence. Proposition 1 bounds the error at a single timestep, but guided sampling involves many steps; whether trajectories stay on the data manifold depends on how these errors accumulate. Guided sampling is iterative: errors at step k corrupt the state for step k+1.

Proposition 2 (Cumulative Guidance Error). *Under guided Euler sampling with L_g-Lipschitz guidance, the cumulative perturbation for ϵ-prediction contains a $-\ln t_0$ term that diverges as $t_0 \to 0$, while the bound for x-prediction remains $\mathcal{O}(1-t_0)$. Full statement and proof in Appendix B.*

Early high-noise steps contribute disproportionately large errors under ϵ-prediction, corrupting the trajectory for all subsequent steps and driving systematic departure from the data manifold.

Remark 1 (Manifold Force Interaction). The score $\nabla_{z_t} \log p_t(z_t)$ decomposes into a denoising component and a *manifold force* [26] that pulls samples toward the data manifold $\mathcal{M}$. This restoring force is weakest near $t = 0$ (pure noise) and becomes dominant only as $t \to 1$ (clean data) (Appendix A). At the start of the reverse process, where this restoring force is weakest, ϵ-prediction's $\mathcal{O}(1/t)$ error amplification (Proposition 1) is simultaneously at its strongest, corrupting the clean data estimate and allowing guidance to overpower the weakened manifold force, driving samples off $\mathcal{M}$. For x-prediction, no such singularity exists, and guidance and manifold forces compose stably.

The amplification factors above are dimension-independent, but prior work [11, 15,16] has shown that prediction errors themselves scale with dimension: $\delta_\epsilon \sim \sqrt{D}$ while $\delta_x \sim \sqrt{d}$ with $d \ll D$. This base-error gap compounds with the amplification hierarchy; see Appendix A for a detailed analysis.

3.3 Implications for Training-Free Guidance

Guidance methods compute gradients $\nabla_{z_t} \mathcal{E}(\hat{x})$ where $\mathcal{E}$ is an energy function.

Theorem 1 (Gradient Stability). *For a Lipschitz energy function $\mathcal{E}$ with constant L, the guidance gradient bound scales as $\mathcal{O}(1/t)$ for ϵ-prediction, $\mathcal{O}(1)$ for v-prediction, and $\mathcal{O}(\|\mathbf{J}_{x_\theta}\|)$ for x-prediction. Full bounds and proof in Appendix B.*

The gradient bounds mirror the error hierarchy (assuming comparable network Jacobian norms across targets), indicating x-prediction yields the most stable guidance gradients among the three targets, particularly at early timesteps where global structure is determined.

From Error Amplification to Child FID. The error amplification hierarchy predicts a specific empirical signature. Cumulative trajectory perturbation (Proposition 2) means that samples departing the data manifold during early guided steps cannot re-enter the target class's natural distribution. We measure this through *guided-class FID* (Child FID): FID computed between guided samples of class y and real images of class y. A model achieving high Validity (classifier accuracy) but high C-FID produces adversarial-like successes: samples that fool the classifier without resembling real class members. The hierarchy predicts that ϵ-prediction enters this adversarial regime at lower guidance strengths than v- or x-prediction, a prediction we test directly in Sec. 5.

Table 1: ImageNet 256×256 pretrained models. All use the Diffusion Transformer architecture. *JiT-G/16 (2B) achieves FID 1.82; we use JiT-H for parameter comparability. †Includes VAE decoder (49M); under TFG the decoder is invoked at every denoising step for guidance in pixel space, making it an integral part of the generation pipeline. ‡Cascade total across all stages; not directly comparable to single-pass GFLOPs.

Space	Model	Target	FID↓	IS↑	Params	GFLOPs
Pixel	PixelFlow [3]	v	1.98	282.1	677M	2909‡
	JiT-H/16* [18]	x	**1.86**	**303.4**	953M	182
Latent	DiT-XL/2 [25]	ϵ	2.27	278.2	724M†	119
	SiT-XL/2 [21]	v	2.06	277.5	724M†	119

Applying TFG to x-prediction. Under TFG [38], x-prediction simplifies guidance: $\hat{x} = x_\theta(z_t, t)$ directly, bypassing the unstable recovery formula. The full algorithm and latent-space details are in Appendix F.

4 Experiments

We evaluate whether x-prediction provides a better foundation for training-free guidance compared to ϵ- and v-prediction. Full experimental protocols and additional studies are in Appendix G.

4.1 Models

We use official pretrained checkpoints spanning three prediction targets and two operating spaces (Tab. 1); model sources, seeds, and the compute budget are in Appendix I. DiT-XL/2 (ϵ) and SiT-XL/2 (v) share identical architecture, parameters (675M diffusion model + 49M VAE decoder), training data, and latent space, isolating prediction target as the sole variable. JiT-H/16 (x, 953M) is our primary x-prediction model, chosen over the larger JiT-G (2B) for parameter-scale comparability. PixelFlow (v, pixel) provides a critical control: against SiT it isolates operating-space effects; against JiT it isolates prediction target within pixel space. ADM-G [7] (U-Net, ϵ-prediction, FID 4.59) is the only available pixel-space ϵ-prediction baseline (Appendix E). JiT model variants (B/L/H/G) are detailed in Appendix E.

Why models differ beyond prediction target. Li and He [18] showed that, under identical pixel-space transformer training, ϵ-prediction achieves FID 372.38 versus 8.62 for x-prediction, a 43× gap indicating that ϵ-prediction depends on latent-space compression to function competitively. Each model in our comparison therefore represents its prediction target's best achievable configuration; the architecture and space differences are consequences, not confounds, of prediction target choice (Appendix E).

4.2 Controlled Ablation: Crossed-Lines

To isolate prediction target effects from all other confounds, we train identical MLP-based flow matching models on a 2D crossed-lines dataset (two 1D line manifolds, $b{=}a$ and $b{=}{-}a$, with perpendicular Gaussian noise) embedded in ambient dimensions $D \in \{2, 8, 32, 128, 512\}$ via column-orthogonal projection (Appendix G).

Architecture and training. For each D, we train three residual MLPs (256 hidden, 5 blocks) differing only in prediction target, with flow matching for 500 epochs and identical hyperparameters; a separate MLP classifier per D provides the DPS signal (details in Appendix G).

Guidance and evaluation. We apply DPS with guidance strength $s{=}10$, using Euler sampling with 100 steps. Each condition generates 10,000 samples. We report **on-manifold rate**: the fraction of generated samples within perpendicular distance δ of the true 1D manifold, where δ is the 95th percentile of ground truth perpendicular distances (measurement details in Appendix G). By embedding the same 1D manifold ($d{=}1$) in progressively higher ambient dimensions, this ablation directly tests the dimension-dependent error scaling analyzed in Appendix A. Full experimental details and additional metrics are in Appendix G.

4.3 Fine-Grained Bird Classification Benchmark

We construct a hierarchical fine-grained classification benchmark to evaluate training-free guidance quality at the species level.

Construction. Starting from 30 ImageNet bird classes (e.g., `goldfinch`, `hummingbird`, `drake`), we identify fine-grained species from a 525-species bird classification dataset [27] (previously used for fine-grained guidance evaluation by Ye *et al.* [38]) that map to each parent, yielding 143 species nested within 30 parent classes (2–20 species per parent, mean 4.8; dataset details in Appendix G). This hierarchy naturally separates two levels of conditioning: classifier-free guidance (CFG) [14] steers toward the parent class using the model's own class conditioning, while gradient-based guidance (DPS [5]) steers each sample toward a specific species via an external fine-grained classifier[1]. We use separate classifiers for guidance and evaluation to avoid circular evaluation [33] (Appendix C).

Why fine-grained birds? Bird classes are already part of ImageNet's label space, so pretrained models can generate them without domain transfer. Species differ in subtle plumage, beak, and eye markings, requiring semantic shifts that make the gap between classifier-fooling artifacts and on-manifold guidance more visible. The hierarchical structure (parent class → species) naturally separates the roles of CFG and DPS, enabling controlled guidance-strength sweeps.

[1] Guidance: `https://huggingface.co/dennisjooo/Birds-Classifier-EfficientNetB2`; evaluation: `https://huggingface.co/chriamue/bird-species-classifier`

Scale. Each condition generates 64 samples per species across all 143 classes (9,152 images), with ρ swept across 5–10 values per model for six models in total.

4.4 Evaluation Protocol

Standard evaluation reports Validity and FID at a fixed guidance strength, but neither metric detects manifold departure. Validity rewards any image the classifier labels correctly, including off-manifold artifacts that fool the network. Parent FID (P-FID, against the full ImageNet reference) rises whether guidance degrades images or successfully shifts them toward a sub-class. A survey of 17 method papers reveals that manifold-aware metrics and guidance-strength sweeps remain uncommon: only two report manifold-aware metrics (Appendix D).

Child FID and guidance sweeps. We propose **Child FID** (C-FID): FID [12] between guided samples and the *target domain*, the bird species dataset (justified in Appendix C). Rising P-FID paired with falling C-FID signals successful guidance; both rising signals degradation. Rather than single-point comparisons, we sweep ρ and plot Pareto curves (P-FID vs. Validity, P-FID vs. C-FID; Fig. 4).

Inference setup. We standardize all models to NFE$\approx$100 using each model's native sampler: DiT uses 100-step DDPM, SiT and JiT use 50-step Heun, and PixelFlow uses 30-step$\times$4-stage Euler (NFE=120). Latent models use the ema VAE decoder. These settings reduce NFE from each model's published optimum to enable fair comparison; full configurations are in Appendix G (Tab. 13).

Guidance. We apply DPS [5], adding $\rho\nabla_{z_t} \log p(y \mid \hat{x})$ at each denoising step with x-space corrections. Latent models (DiT, SiT) require VAE decoder passes at each guidance step for pixel-space gradients; this overhead is absent for pixel-space models (JiT, PixelFlow). Full configuration details are in Appendix G.

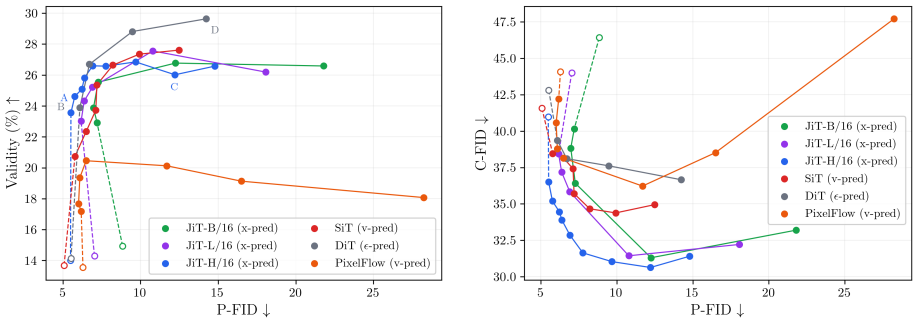
5 Results

5.1 Crossed-Lines Ablation

The crossed-lines toy experiment (Fig. 1b, Fig. 2) isolates the effect of prediction target in a controlled setting where architecture and training are identical. As ambient dimension increases from D=2 to D=512, the hierarchy predicted by Proposition 1 emerges (Fig. 3): x-prediction maintains 93.3% on-manifold rate at D=512, v-prediction degrades to 21.5%, and ϵ-prediction collapses to 0.5%, consistent with $\sqrt{D}$ error scaling (Appendix A). Full metrics and the complete visualization grid are in Appendix G (Tab. 10 and Fig. 11).

5.2 Fine-Grained Bird Classification

Fig. 4 presents guidance-strength sweeps across six models spanning three prediction targets and two operating spaces.



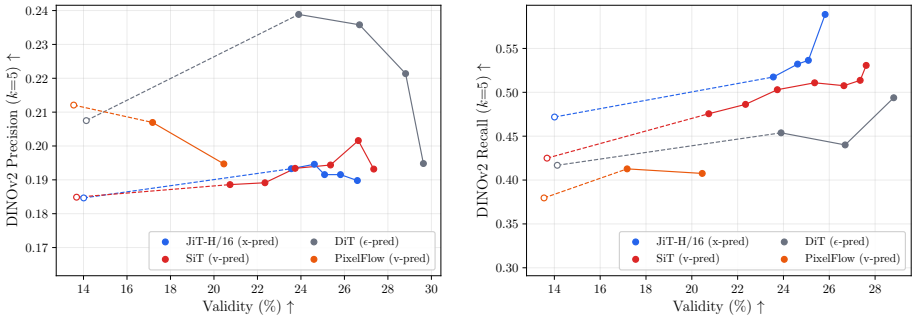
(a) P-FID vs. Validity **(b)** P-FID vs. C-FID

Fig. 4: Guidance-quality Pareto frontiers on fine-grained bird classification. Each curve traces one model across guidance strengths ρ; open markers denote CFG-only baselines (ρ=0), dashed segments connect to the first guided setting. Each data point represents 9,152 generated images (143 species × 64 samples). **(a)** P-FID vs. Validity. Labels A–D mark the operating points visualized in Fig. 7. **(b)** P-FID vs. C-FID.

Validity alone is misleading. All models show increasing Validity with ρ (Fig. 4a), but the quality cost differs substantially across targets. At matched Validity (≈26.6%), JiT-H (ρ=3, P-FID 6.9) and DiT (ρ=0.1, P-FID 6.7) appear equivalent. Under stronger guidance, DiT reaches 29.6% Validity but at P-FID 14.2 (2.6× its baseline), while JiT maintains low P-FID throughout. Single-point comparisons obscure this divergence.

Child FID reveals manifold fidelity. Fig. 4b disambiguates the P-FID trade-off. At matched Validity (≈26.6%), JiT-H achieves C-FID 32.9 versus DiT's 38.1 and SiT's 34.7, a 5.2-point gap between x- and ϵ-prediction at identical classifier confidence. DiT's trajectory is revealing: from ρ=0.1 to ρ=0.5, Validity gains come with stagnant C-FID and collapsing P-FID, a characteristic signature of adversarial-like guidance predicted by Proposition 1. Qualitative inspection confirms this pattern: DiT's guided samples concentrate on a narrow set of visual templates, whereas JiT produces diverse compositions across the same species (Fig. 7).

Mode collapse. We quantify the diversity loss with DINOv2 Precision and Recall [23] (Fig. 5). Mode collapse corresponds to high Precision with low Recall: a model that produces a narrow but realistic subset covers the target distribution poorly. DiT (ϵ) follows this pattern, with Precision peaking at 0.24 while Recall stays around 0.49, whereas JiT-H (x) has lower Precision but higher Recall (up to 0.59). This accounts for the high Precision of ϵ-prediction: Precision measures only the realism of generated samples, not coverage of the target distribution, and Recall shows that DiT covers less of it. The behavior is consistent with the error amplification hierarchy (Proposition 1): under guidance, ϵ-prediction concentrates samples on a narrow set of classifier-activating features and reduces the support that x-prediction retains.



(a) Validity vs. Precision **(b)** Validity vs. Recall

Fig. 5: Precision and Recall under guidance on fine-grained bird classification. DINOv2 k-NN Precision and Recall traced over the DPS ρ-sweep; open markers denote CFG-only baselines (ρ=0), dashed segments connect to the first guided setting. **(a)** ϵ-prediction (DiT) attains the highest Precision (fidelity), while **(b)** x-prediction (JiT) attains the highest Recall (coverage). The joint pattern of high Precision and low Recall is the mode-collapse signature of ϵ-prediction. Full six-model curves are in Fig. 14 (Appendix H).

Prediction target vs. operating space. PixelFlow (v-prediction, pixel space) provides a critical control for JiT (x-prediction, pixel space). Despite sharing the same operating space, PixelFlow exhibits a substantially worse Pareto frontier: its C-FID initially improves from 44.1 (ρ=0) to 36.2 (ρ=2) but then *increases* to 47.7 at strong guidance (Fig. 4b), signaling manifold departure. In contrast, JiT-H's C-FID continues decreasing throughout the sweep, reaching 30.6 at ρ=8. This indicates that the prediction target, rather than the operating space, is the primary determinant of guidance robustness in this comparison.

Latent space models. DiT and SiT operate in a 32×32 VAE latent space, requiring decoder passes at each guidance step. SiT (v-prediction) dominates DiT (ϵ-prediction), consistent with bounded error attenuation (Proposition 1), but both are dominated by JiT in C-FID. The VAE's 8× downsampling may further limit fine-grained guidance resolution, though we do not isolate this factor.

Scaling with model capacity. Among JiT variants (B/L/H), larger models achieve strictly better Pareto frontiers: JiT-H reaches C-FID 30.6 versus JiT-B's 31.3 at comparable P-FID. Increased capacity yields higher-fidelity $\hat{x}$ estimates and more accurate guidance gradients, a *guidance scaling* effect.

Connection to theory. The C-FID evidence supports the error amplification hierarchy (Propositions 1 and 2 and Theorem 1): samples departing $\mathcal{M}$ at early steps cannot return to realistic distributions, inflating C-FID even when the classifier is fooled. x-prediction tolerates aggressive guidance with less manifold degradation. The same ordering holds under other gradient-based methods (LGD [34], FreeDoM [39]) and on a second fine-grained domain, a 34-species

butterfly benchmark (Appendix G, Figs. 8 and 9): guidance quality follows the prediction target, not the specific method or domain. Additional ablations are in Appendix G.

5.3 Style Transfer

To test whether the prediction target hierarchy extends beyond classification, we apply DPS to style transfer: guiding generation toward a target visual style via CLIP Gram matrix matching [10]. We use CLIP ViT-B/16 [28] for guidance and evaluate with Gram Distance (CLIP ViT-B/32; lower = stronger style match) and Content Accuracy (DeiT-Small [37] top-1 accuracy on the original ImageNet class; higher = better content preservation). Four WikiArt[2] styles are evaluated across 100 ImageNet classes (400 images per model/ρ).

Fig. 6 presents Gram Distance vs. Content Accuracy Pareto frontiers. All models trade content preservation for style fidelity as ρ increases, but the degradation rate varies considerably across prediction targets. Overall, the differences between models are less pronounced than in fine-grained classification (Sec. 5.2).

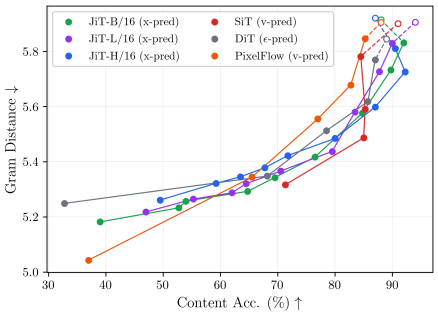


Fig. 6: Style transfer: Gram Distance vs. Content Accuracy. Curves trace models as ρ increases (right to left). Open markers = CFG-only baselines. Lower-right is preferred.

ϵ- and v-prediction collapse under strong style guidance. DiT's Content Accuracy drops from 89% (ρ=0) to 1.5% (ρ=10), effectively random, while achieving Gram Distance 5.30. The low Gram Distance is meaningless when images no longer depict recognizable content. PixelFlow (v-prediction, pixel space) follows a similar pattern: at ρ=4, it achieves the lowest Gram Distance of any model (5.04) but at only 37% Content Accuracy, where images lose semantic coherence.

x-prediction preserves content over a wider guidance range. At high Content Accuracy (>80%), JiT-H achieves better Gram Distance than DiT and PixelFlow: at comparable Content Accuracy (≈80%), JiT-H (ρ=10, Gram Distance 5.48) outperforms DiT (ρ=0.5, Gram Distance 5.62) and PixelFlow (ρ=1, Gram Distance 5.56). JiT-H further reaches Content Accuracy 49.5% at ρ=50 (Gram Distance 5.26), matching DiT's best Gram Distance while retaining meaningful content fidelity.

[2] `https://www.wikiart.org/`

SiT is competitive at moderate guidance. SiT (v-prediction, latent) achieves a competitive Pareto frontier at moderate ρ: Gram Distance 5.49 at Content Accuracy 85.0% (ρ=1). At stronger guidance (ρ=4), SiT already drops to 71% Content Accuracy while JiT-H retains 87%; by ρ=10, SiT falls to 38% while JiT-H retains 80%. The Pareto frontiers of SiT and JiT overlap in the moderate-guidance regime and diverge at the extremes.

Consistent failure modes across tasks. Although the quantitative gap between models is smaller than in fine-grained classification, the qualitative failure modes are consistent: DiT's guided samples exhibit mode collapse and loss of background detail, while JiT-H preserves compositional diversity (Fig. 10 in Appendix G). This suggests that the error amplification hierarchy (Proposition 1) manifests across guidance tasks, even when the aggregate metrics show smaller differences. The Gram matrix guidance signal captures aggregate texture statistics rather than fine-grained spatial details, which may explain the reduced quantitative separation.

Additional experiments. We also evaluate DPS on two inverse problems (Gaussian deblurring and 4× super-resolution), where x-prediction again achieves the best perceptual quality (LPIPS) across all models. Full results and discussion are in Appendix G (Sec. G.9).

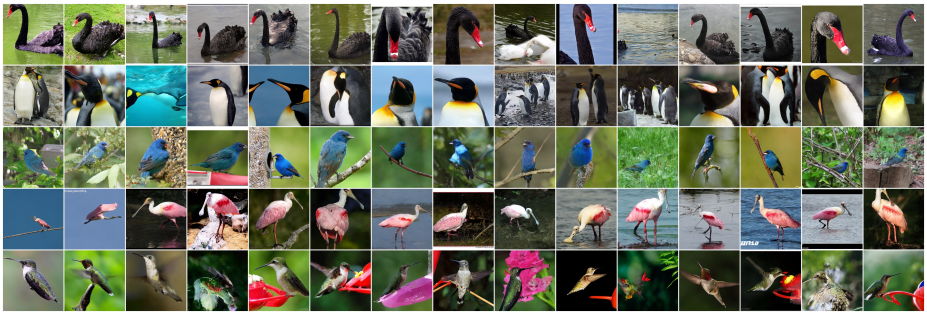
5.4 Qualitative Analysis

Fig. 7 presents randomly drawn (not cherry-picked) guided samples from JiT-H (x-prediction) and DiT (ϵ-prediction) at two guidance regimes: moderate (ρ chosen for matched P-FID≈6; (a),(b)) and strong ((c),(d)), across five visually distinct bird species. Corresponding visualizations for SiT and PixelFlow are in Appendix G (Fig. 13).

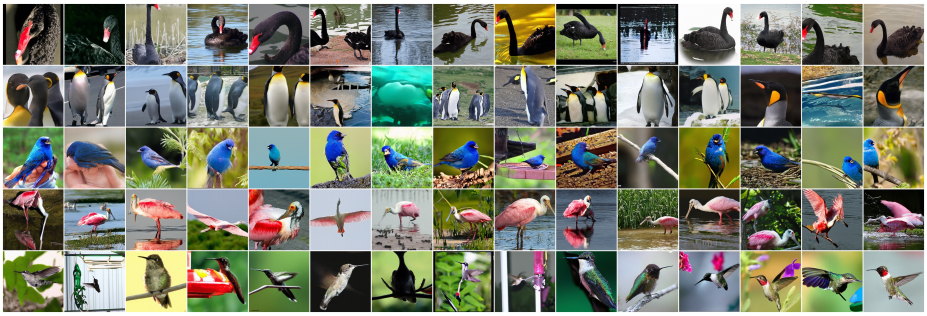
Off-manifold degradation is visible but not catastrophic for x-prediction. Under strong guidance, JiT-H (Fig. 7c) shows visible quality degradation (posterization and reduced fine detail) but maintains diverse poses, varied backgrounds, and recognizable species-specific features. DiT (Fig. 7d) achieves 29.6% Validity (higher than JiT-H's 26.0%), yet its samples exhibit pronounced mode collapse: many images share similar poses, framing, and uniform dark backgrounds, particularly visible in Black Swan and Painted Bunting.

Mode concentration as a failure signature. The samples make concrete the mode collapse quantified in Sec. 5.2 (Fig. 5): the low Recall measured there corresponds to the narrow set of poses and backgrounds DiT repeats here. C-FID registers the same effect: DiT's C-FID (36.7) remains worse than JiT-H's (30.6) despite higher Validity. The samples show how it appears: DiT's ϵ-prediction reuses a small set of classifier-friendly patterns rather than covering each species' variation, whereas JiT-H retains diverse poses and compositions.

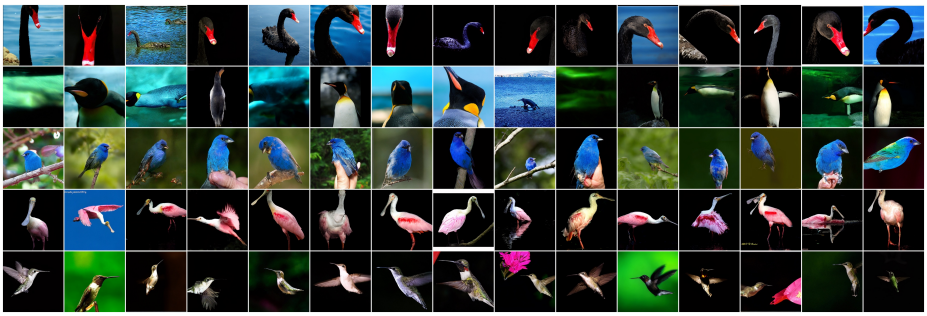
(a) JiT-H (x, ρ=1.5): P-FID 6.2, C-FID 34.5, Val. 25.1% (point A in Fig. 4a)



(b) DiT (ϵ, ρ=0.05): P-FID 6.1, C-FID 39.4, Val. 23.9% (point B in Fig. 4a)



(c) JiT-H (x, ρ=8): P-FID 12.2, C-FID 30.6, Val. 26.0% (point C in Fig. 4a)



(d) DiT (ϵ, ρ=0.5): P-FID 14.2, C-FID 36.7, Val. 29.6% (point D in Fig. 4a)

Fig. 7: Guided generation on five bird species (15 random samples each, no cherry-picking). (a),(b): moderate guidance (matched P-FID≈6); (c),(d): strong guidance. Rows: Black Swan, Emperor Penguin, Painted Bunting, Roseate Spoonbill, Ruby-throated Hummingbird.

Practical implication. Low Validity with preserved image quality (JiT-H at strong guidance) is preferable to high Validity with degraded diversity (DiT): a user can retry generation for the correct class, but cannot recover from mode collapse or manifold departure. This asymmetry further motivates C-FID over Validity as the primary evaluation metric for training-free guidance.

6 Conclusion

Whether training-free guidance fails gracefully (missing the target but producing a realistic image) or catastrophically (collapsing into off-manifold artifacts) depends on the prediction target. The mechanism is the recovery formula: ϵ-prediction divides by t, amplifying errors unboundedly at high noise; v-prediction incurs bounded amplification; x-prediction incurs none. On ImageNet, C-FID supports this hierarchy with a 5.2-point gap at matched Validity (manifold damage invisible to standard evaluation), and PixelFlow's C-FID reversal under strong guidance isolates prediction target as the decisive factor. The hierarchy extends to style transfer, establishing x-prediction as the target that keeps guidance failures graceful.

The fine-grained bird benchmark and C-FID protocol we introduce let any ImageNet-scale diffusion model be immediately tested for manifold-aware guidance quality, and can serve as a practical diagnostic for evaluating new models.

Limitations. Our comparison involves models differing in architecture and operating space beyond prediction target; no single comparison perfectly isolates it, and the conclusion instead rests on the convergence of five independent control experiments (Appendix E), of which the latent-space DiT vs. SiT pair and the pixel-space PixelFlow vs. JiT pair are the most controlled within each space. The theoretical analysis assumes Lipschitz energy functions and well-trained models. We evaluate gradient-based TFG methods (DPS, LGD, FreeDoM), not the full TFG parameter space, and all experiments use 256×256 resolution; scaling behavior at higher resolutions remains to be tested.

Future Work. Inference-time scaling methods [17, 22], which are search-based and SMC approaches that compound guidance across many candidates, depend critically on $\hat{x}$ quality; the prediction target hierarchy should govern their sample efficiency. Video and text-to-image generation present higher-dimensional sequential settings where error amplification compounds across frames, and our dimension scaling analysis predicts even larger gaps between prediction targets.

Not all prediction targets keep guided samples on the manifold, but x-prediction does. For practitioners building inference-time control pipelines, it is the robust foundation.

Acknowledgements This research was supported by Seoul National University of Science and Technology.

References

1. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. Journal of Machine Learning Research (2023)
2. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
3. Chen, S., Ge, C., Zhang, S., Sun, P., Luo, P.: Pixelflow: Pixel-space generative models with flow. arXiv preprint arXiv:2504.07963 (2025)
4. Chidambaram, M., Gatmiry, K., Chen, S., Lee, H., Lu, J.: What does guidance do? A fine-grained analysis in a simple setting. In: Advances in Neural Information Processing Systems (2024)
5. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: International Conference on Learning Representations (2023)
6. Chung, H., Kim, J., Park, G.Y., Nam, H., Ye, J.C.: CFG++: Manifold-constrained classifier free guidance for diffusion models. In: International Conference on Learning Representations (2025)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
8. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**(496), 1602–1614 (2011). `https://doi.org/10.1198/jasa.2011.tm11181`
9. Feng, R., Yu, C., Deng, W., Hu, P., Wu, T.: On the guidance of flow matching. In: International Conference on Machine Learning (2025)
10. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
11. Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., Guo, B.: Efficient diffusion training via min-SNR weighting strategy. In: International Conference on Computer Vision (2023)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems **30** (2017)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Jin, Q., Wang, C.: Revisiting diffusion model predictions through dimensionality. arXiv preprint arXiv:2601.21419 (2026)
16. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: Advances in Neural Information Processing Systems (2022)
17. Kim, S., Kim, M., Park, D.: Test-time alignment of diffusion models without reward over-optimization. In: International Conference on Learning Representations (2025)
18. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
19. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (2023)

20. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (2023)
21. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision (2024)
22. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., Xie, S.: Inference-time scaling for diffusion models beyond scaling denoising steps. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
23. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International Conference on Machine Learning (2020)
24. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. In: International Conference on Machine Learning (2022)
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: International Conference on Computer Vision (2023)
26. Pidstrigach, J.: Score-based generative models detect manifolds. Advances in Neural Information Processing Systems **35**, 35852–35865 (2022)
27. Piosenka, G.: 525 bird species – image classification. Kaggle (2023), `https://www.kaggle.com/datasets/gpiosenka/100-bird-species`, cC0: Public Domain. HuggingFace mirror: `https://huggingface.co/datasets/chriamue/bird-species-dataset`. Accessed 30 June 2026
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
29. Räisä, O., van Breugel, B., van der Schaar, M.: Position: All current generative fidelity and diversity metrics are flawed. In: International Conference on Machine Learning (2025)
30. Robbins, H.: An empirical Bayes approach to statistics. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 157–163. University of California Press (1956)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
32. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022)
33. Shen, Y., Jiang, X., Wang, Y., Yang, Y., Han, D., Li, D.: Understanding and improving training-free loss-based diffusion guidance. In: Advances in Neural Information Processing Systems (2024)
34. Song, J., Zhang, Q., Yin, H., Mardani, M., Liu, M.Y., Kautz, J., Chen, Y., Vahdat, A.: Loss-guided diffusion models for plug-and-play controllable generation. In: International Conference on Machine Learning. vol. 202, pp. 32483–32498 (2023)
35. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
36. Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

37. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
38. Ye, H., Lin, H., Han, J., Xu, M., Liu, S., Liang, Y., Ma, J., Zou, J., Ermon, S.: Tfg: Unified training-free guidance for diffusion models. In: Advances in Neural Information Processing Systems (2024)
39. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. In: International Conference on Computer Vision (2023)