ONE-STEP DIFFUSION INPAINTING VIA RECONSTRUCTION SPACE ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have shown strong performance in image inpainting by leveraging powerful generative priors, but their iterative denoising process often incurs substantial inference cost. One-step diffusion models provide a promising alternative, yet removing iterative denoising also removes the opportunity for progressive error correction, allowing prediction errors produced in a single step to more directly affect the final reconstruction. In this work, we investigate how prediction errors propagate into reconstruction errors under different prediction targets and, more importantly, whether lower theoretical error propagation translates into better downstream adaptation. Although velocity prediction has lower theoretical error amplification than noise prediction under the adopted noise schedule, controlled experiments show that replacing the prediction target used during pretraining with velocity prediction degrades downstream adaptation performance under the evaluated training budget. This reveals a discrepancy between theoretical error propagation and practical adaptation performance. Motivated by this observation, we propose reconstruction-space adaptation for one-step diffusion inpainting. Instead of changing the pretrained prediction target, the proposed method preserves noise prediction while extending supervision from the prediction space to the reconstructed latent representation and decoded image, thereby aligning downstream optimization more directly with the final restoration objective. Independent-sample averaging and a lightweight residual refiner are further used as complementary components to reduce sampling-induced output fluctuations and correct remaining image-space errors. Extensive experiments on Places2 and CelebA-HQ validate the proposed adaptation strategy, showing consistent reconstruction improvements under controlled adaptation settings and competitive inpainting performance.

1 INTRODUCTION

Image inpainting is a fundamental task in computer vision that aims to restore missing regions by exploiting textural, color, structural, and other contextual cues from known regions (Xiang et al., 2023), with broad applications in image editing (Li et al., 2026), remote sensing (Wu et al., 2025), and cultural heritage conservation (Lee & Choi, 2025).

Traditional image inpainting methods are mainly based on partial differential equations and patch-matching-based texture synthesis (Bertalmio et al., 2000; Criminisi et al., 2004). These methods typically rely on repetitive texture priors and can achieve satisfactory results when the missing regions contain highly repetitive textures or relatively simple structures. However, their limited ability to model complex structures substantially restricts their performance when dealing with large missing regions or structurally complex scenes. In recent years, large-scale data-driven deep learning methods have enabled inpainting models to automatically learn semantic, structural, and textural representations from data, significantly improving their ability to restore complex missing regions. Early deep learning-based inpainting methods were mainly built upon convolutional neural networks, which enlarged the receptive field and incorporated contextual information to learn mappings between missing and known regions (Pathak et al., 2016; Iizuka et al., 2017). Subsequently, generative adversarial networks, attention mechanisms, and Transformers were introduced into image inpainting, further enhancing the modeling of high-level semantics and long-range dependencies (Yu et al., 2018; Wan et al., 2021; Li et al., 2022; Zheng et al., 2022). More recently, diffusion mod-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

els (Ho et al., 2020) have further improved the inpainting quality of complex missing regions by leveraging powerful generative priors (Rombach et al., 2022; Xie et al., 2023; Ju et al., 2024). RePaint (Lugmayr et al., 2022) repeatedly incorporates information from known regions as conditional constraints during the reverse diffusion process, while StrDiffusion (Liu et al., 2024) introduces structural guidance into the diffusion process to improve semantic and structural consistency. Nevertheless, the substantial inference cost caused by multi-step iterative denoising remains a major challenge for diffusion-based image inpainting.

To reduce the inference cost of diffusion models, recent studies have begun to explore few-step and one-step diffusion models for image inpainting (Xie et al., 2025; Vu et al., 2026; Zhou et al., 2026). TurboFill (Xie et al., 2025) adapts the few-step text-to-image model DMD2 (Yin et al., 2024) to image inpainting through a dedicated inpainting adapter and a three stage adversarial training strategy, enabling high-quality inpainting with only a few diffusion steps. InverFill (Vu et al., 2026) further points out that random Gaussian initialization in few-step diffusion inpainting can lead to semantic misalignment between the generated content and the known background. It therefore employs a one-step inversion network to encode the input information into a semantically constrained initial latent representation, which is subsequently used with an existing few-step diffusion model for inpainting. These studies demonstrate that few-step and one-step diffusion models can substantially reduce the inference cost of image inpainting. However, as the number of diffusion steps decreases, the opportunities for iterative denoising and progressive error correction are also reduced. When inference is further compressed into a single step, prediction errors produced in that step can more directly affect the final reconstruction. Therefore, one-step diffusion inpainting should not only focus on inference efficiency, but also reconsider how pretrained diffusion models can be reliably adapted in the absence of subsequent error correction.

A key aspect of this adaptation problem lies in the choice of prediction target. Diffusion models can adopt different prediction targets, such as noise prediction, velocity prediction, and data prediction (Salimans & Ho, 2022), and prediction errors under different targets can exhibit substantially different behaviors when estimating or reconstructing the clean latent representation (Lee & Lee, 2026; Jin & Wang, 2026). Recent studies have analyzed these differences from perspectives such as training-free guidance and data geometry. For example, Lee & Lee (2026) investigated how different prediction targets affect clean-sample estimation errors and showed that, under high-noise conditions, they can exhibit substantially different error amplification behaviors. From the perspective of data dimensionality, Jin & Wang (2026) further studied the suitability of different prediction targets and proposed a unified framework for learning prediction targets. These studies indicate that prediction-target choice is not arbitrary, but primarily examine it from the perspectives of theoretical properties, generative training, or inference-time guidance. For diffusion models that have already undergone large-scale pretraining and are subsequently adapted to downstream tasks under a limited training budget, however, it remains unclear whether a prediction target with lower theoretical error propagation necessarily leads to better practical adaptation performance.

To examine this question, we first investigate how prediction errors propagate into reconstruction errors under different prediction targets. Our analysis shows that, under the noise schedule and fixed timestep adopted in this work, velocity prediction has lower theoretical error amplification than noise prediction. However, controlled experiments show that replacing the prediction target used during pretraining with velocity prediction degrades downstream adaptation performance under the evaluated training budget. This observation reveals a discrepancy between theoretical error propagation and practical adaptation performance: error-propagation characteristics describe how a given prediction error affects reconstruction, but do not determine how accurately a different prediction target can be learned during downstream adaptation. Therefore, theoretical error propagation alone is insufficient for selecting the prediction target when adapting a pretrained diffusion model.

Motivated by this observation, we propose reconstruction-space adaptation for one-step diffusion inpainting. Instead of changing the pretrained prediction target, the proposed method extends supervision from the prediction space to the reconstructed latent representation and decoded image, thereby aligning downstream optimization more directly with the final restoration objective while retaining the prediction target used during pretraining. This allows downstream adaptation to constrain reconstruction quality directly without selecting a new prediction target solely according to its theoretical error-propagation properties. The resulting generator is further complemented by two practical components. Independent-sample averaging is used to reduce output fluctuations caused by

stochastic latent sampling, while a lightweight residual refiner corrects the remaining image-space errors.

Our main contributions are summarized as follows:

- We identify a discrepancy between theoretical error propagation and practical downstream adaptation in one-step diffusion inpainting. Controlled comparisons show that a prediction target with lower theoretical error amplification does not necessarily yield better adaptation performance when replacing the target used during pretraining.
- We propose reconstruction-space adaptation for one-step diffusion inpainting. The proposed method retains the pretrained prediction target while extending supervision to the reconstructed latent representation and decoded image, thereby aligning downstream optimization more directly with the final restoration objective.
- Extensive experiments on Places2 and CelebA-HQ validate the proposed adaptation strategy, showing consistent reconstruction improvements under controlled adaptation settings and competitive inpainting performance. Additional analyses characterize independent-sample averaging and residual refinement as complementary mechanisms for output stabilization and image-space correction.

2 METHOD

Multi-step diffusion inpainting progressively corrects prediction errors through iterative denoising over multiple steps. When inference is compressed into a single step, however, subsequent opportunities for progressive error correction are no longer available, allowing errors produced by a single prediction to propagate more directly to the final reconstruction. To investigate this issue, we first analyze how prediction errors propagate into reconstruction errors under different prediction targets and further examine empirically the relationship between theoretical error-propagation characteristics and practical downstream adaptation performance. Our analysis and experiments show that lower theoretical error amplification does not necessarily translate into better adaptation performance under a limited training budget.

Motivated by this observation, we propose reconstruction-space adaptation for one-step diffusion inpainting, as illustrated in Fig. 1. The overall framework consists of two training stages and an inference-time stabilization strategy. In Stage A, we retain the noise-prediction target used during pretraining and extend supervision from the prediction space to the reconstructed clean latent and its corresponding decoded image. LoRA and a conditional control branch are used for parameter-efficient adaptation of the pretrained one-step diffusion model, allowing downstream optimization to constrain final reconstruction quality more directly. In Stage B, the adapted diffusion generator is frozen and a lightweight residual refiner is trained to correct remaining systematic image-space errors. In addition, to reduce output fluctuations caused by stochastic latent sampling, we introduce an inference-time independent-sample averaging strategy that requires no additional training. Multiple independently sampled one-step predictions are averaged to reduce sampling variance, after which the same residual refiner is applied to obtain the final reconstruction.

2.1 TASK FORMULATION AND METHOD OVERVIEW

To establish a unified formulation for subsequent analysis and method design, we first formalize the one-step image inpainting task and clarify the information flow during training and inference. Let $x \in [0,1]^{3\times H\times W}$ denote a complete image and $m \in \{0,1\}^{1\times H\times W}$ a binary mask, where $m = 1$ indicates missing pixels. The visible image is defined as $x_{\text{vis}} = (1-m)\odot x$ and the inpainting condition is $c = [x_{\text{vis}}, m]$.

We consider a pretrained one-step diffusion generator whose prediction target during pretraining is noise prediction and adapt it to conditional image inpainting. As shown in Fig. 1, Stage A retains the pretrained noise-prediction target while adding direct supervision to the reconstructed clean latent and decoded image. After generator adaptation, Stage B freezes the generator and trains a lightweight residual refiner to correct remaining image-space errors. During inference, independent-sample averaging can be applied before residual refinement to reduce sampling-induced output fluctuations.

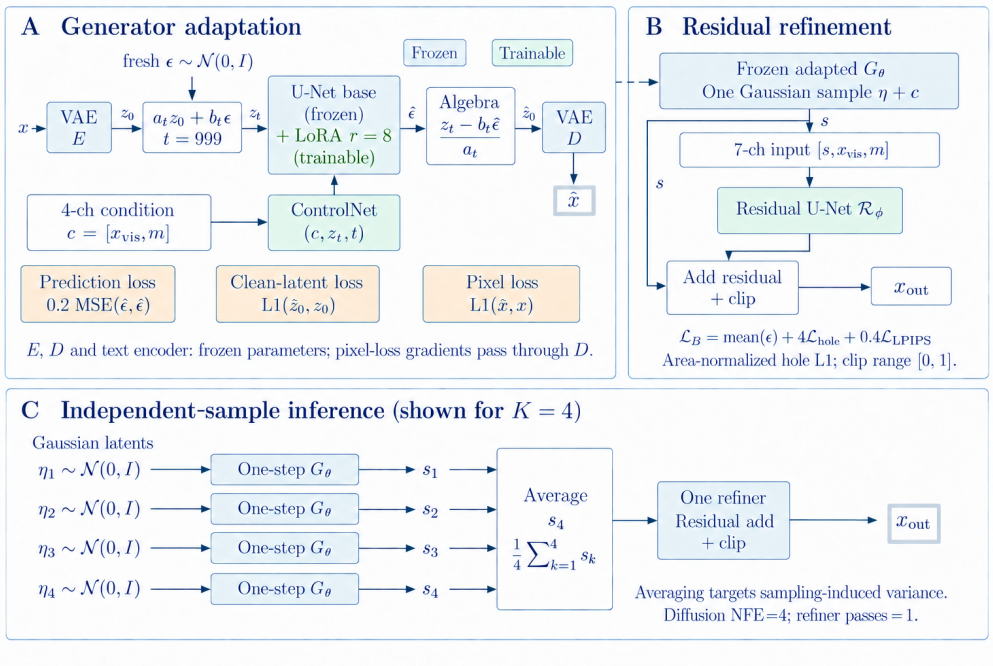

Figure 1: Overview of the proposed method. Stage A retains the pretrained noise-prediction target and trains ControlNet and LoRA using prediction, clean-latent, and decoded-image losses. The VAE decoder remains frozen, while gradients from the image reconstruction loss propagate through it. Stage B trains a lightweight residual refiner with the adapted generator frozen. At inference time, K independently sampled one-step decoded predictions are averaged before a single refiner pass and known-pixel composition. The illustration uses $K = 4$, corresponding to four independent denoiser evaluations rather than four sequential denoising steps.

2.2 Prediction-Target Error Propagation

Without iterative error correction, prediction errors produced in a single diffusion step are propagated more directly to the reconstructed clean latent. To determine whether changing the prediction target can reduce this reconstruction sensitivity, we analyze the error-propagation characteristics of noise, velocity, and data prediction.

Let

$$\bar{\alpha}_t = \prod_{j=0}^{t} (1 - \beta_j), \qquad a_t = \sqrt{\bar{\alpha}_t}, \qquad b_t = \sqrt{1 - \bar{\alpha}_t}. \tag{1}$$

The clean latent z_0 is noised according to

$$z_t = a_t z_0 + b_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

For noise prediction, velocity prediction with $v = a_t \epsilon - b_t z_0$, and direct data prediction, the corresponding clean-latent estimates are (Salimans & Ho, 2022)

$$\hat{z}_0^{\epsilon} = \frac{z_t - b_t \hat{\epsilon}}{a_t}, \qquad \hat{z}_0^{v} = a_t z_t - b_t \hat{v}, \qquad \hat{z}_0^{\text{data}} = \hat{z}_0^{\text{net}}. \tag{3}$$

Let q denote a prediction target and $\hat{z}_0^q$ the clean latent reconstructed from that target. For a fixed timestep and a common forward-noising distribution, subtracting the exact reconstruction yields

$$\mathcal{E}_q(t) = \mathbb{E}\left[\|\hat{z}_0^q - z_0\|_{\text{RMS}}^2\right] = \kappa_q(t)^2 \, \mathbb{E}\left[\|\hat{q} - q\|_{\text{RMS}}^2\right], \tag{4}$$

where RMS denotes the root mean square over latent elements, and the expectation covers images, masks, VAE sampling, and forward noise. The corresponding reconstruction-error gains are

$$\kappa_\epsilon(t) = \frac{b_t}{a_t}, \qquad \kappa_v(t) = b_t, \qquad \kappa_{\text{data}}(t) = 1. \tag{5}$$

Under the 1000-step scaled-linear schedule used in our experiments, with β increasing from 0.00085 to 0.012, $t = 999$ gives $\kappa_\epsilon \approx 14.6$ and $\kappa_v \approx 1$. Thus, for target-space prediction errors of the same magnitude, velocity prediction has substantially lower theoretical reconstruction-error amplification than noise prediction. Crucially, these gains characterize only how a given prediction error is propagated through the reconstruction mapping; they do not determine how accurately a different prediction target can be learned during downstream adaptation. A smaller κ_q therefore does not by itself imply a better adapted predictor. Our experiments show that, for the evaluated pretrained backbone and training budget, replacing the prediction target used during pretraining with velocity prediction degrades adaptation performance (full-budget learning curves in Appendix A). This discrepancy motivates retaining the pretrained noise-prediction target and moving additional supervision toward the reconstruction space.

2.3 Reconstruction-Space Adaptation

The preceding analysis shows that lower theoretical error amplification does not necessarily translate into better downstream adaptation, suggesting that simply replacing the prediction target used during pretraining is insufficient. To improve adaptation while preserving the pretrained prediction target, we propose reconstruction-space adaptation, which extends supervision from the prediction space to the reconstructed clean latent and decoded image.

We instantiate the one-step generator using SD-Turbo (Sauer et al., 2025). A four-channel ControlNet (Zhang et al., 2023a) receives the inpainting condition c, while LoRA adapters (Hu et al., 2022) are applied to the diffusion U-Net. The base U-Net, VAE, and text encoder remain frozen, and the text prompt is empty. Only the ControlNet and LoRA parameters, collectively denoted by θ, are optimized.

During adaptation, z_0 is sampled from the frozen VAE posterior of the ground-truth image x, including the pretrained latent scaling factor. Fresh Gaussian noise is sampled and used to construct z_t at the fixed timestep $t^\star = 999$ according to Eq. equation 2. The generator predicts

$$\hat{\epsilon} = f_\theta(z_t, c, t^\star), \tag{6}$$

from which the reconstructed clean latent and decoded image are obtained as

$$\hat{z}_0 = \frac{z_t - b_{t^\star}\hat{\epsilon}}{a_{t^\star}}, \qquad \hat{x} = D(\hat{z}_0), \tag{7}$$

where D includes inverse latent scaling, VAE decoding, RGB conversion, and clipping to $[0, 1]$.

Let $\ell_1(u, v) = \text{mean}\,|u - v|$ and $\ell_2(u, v) = \text{mean}(u - v)^2$. The Stage A objective is

$$\mathcal{L}_A = \lambda_\epsilon \ell_2(\hat{\epsilon}, \epsilon) + \lambda_z \ell_1(\hat{z}_0, z_0) + \lambda_x \ell_1(\hat{x}, x), \qquad (\lambda_\epsilon, \lambda_z, \lambda_x) = (0.2, 1, 1). \tag{8}$$

The noise-prediction term retains supervision in the output space used during pretraining. The clean-latent term constrains the induced reconstruction before decoding, while the decoded-image term directly constrains the RGB reconstruction before known-pixel composition. Both reconstruction losses are averaged over all elements without additional hole-region weighting. The VAE decoder remains frozen, but gradients from the decoded-image loss propagate through it to the generator.

At the fixed timestep, Eq. equation 3 further implies

$$\ell_1(\hat{z}_0, z_0) = \kappa_\epsilon(t^\star)\, \ell_1(\hat{\epsilon}, \epsilon), \qquad \frac{\partial \hat{z}_0}{\partial \hat{\epsilon}} = -\kappa_\epsilon(t^\star) I. \tag{9}$$

Therefore, at a fixed timestep and under the same reduction, the clean-latent ℓ_1 term does not provide an independent prediction target. Instead, it changes the effective scale and loss geometry of the noise residual after mapping that residual into reconstruction space. In contrast, the decoded-image loss further constrains the nonlinear mapping from the reconstructed latent to RGB space. Their joint use allows adaptation to retain the pretrained noise-prediction target while directly constraining the reconstruction induced by that prediction.

2.4 Residual Image-Space Refinement

Although reconstruction-space adaptation aligns downstream optimization more directly with the final restoration objective, a single generator forward pass can still leave reconstruction errors in

image space. To correct these remaining errors without further modifying the adapted diffusion generator, we introduce a lightweight residual refiner that operates directly on the generated image.

Let $G_\theta(u, c)$ denote the decoded one-step generator mapping at $t^\star$. After Stage A, G_θ is frozen and a residual U-Net R_ϕ is trained. Given a sampled generator output

$$s = G_\theta(\eta, c), \qquad \eta \sim \mathcal{N}(0, I), \tag{10}$$

the refiner takes the seven-channel concatenation $[s, x_{\text{vis}}, m]$ as input and predicts an image-space residual. Each refiner training example uses one independently sampled generator output.

The refined prediction is

$$F_\phi(s, c) = m \odot \text{clip}_{[0,1]}(s + R_\phi([s, x_{\text{vis}}, m])) + x_{\text{vis}}, \tag{11}$$

which preserves all known pixels exactly. For a mini-batch, let

$$e = |F_\phi(s, c) - x| \,. \tag{12}$$

We optimize the refiner with

$$\mathcal{L}_B = \text{mean}(e) + 4 \, \frac{\sum_{i,r,h,w} m_{i,h,w} e_{i,r,h,w}}{3 \max\Big(\sum_{i,h,w} m_{i,h,w}, 1\Big)} + 0.4 \, \mathcal{L}_{\text{LPIPS}}(F_\phi(s, c), x) \,. \tag{13}$$

where i indexes images and r indexes RGB channels.

2.5 Independent-Sample Averaging for Stable Inference

In addition to systematic reconstruction errors, one-step diffusion inpainting remains stochastic because different Gaussian latent samples can produce different restoration results for the same condition. To suppress these sampling-induced fluctuations without retraining the generator or refiner, we introduce independent-sample averaging at inference time.

For a fixed condition c, we independently sample

$$\eta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I), \qquad k = 1, \ldots, K, \tag{14}$$

and obtain

$$s_k = G_\theta(\eta_k, c), \qquad \bar{s}_K = \frac{1}{K} \sum_{k=1}^{K} s_k, \qquad x_{\text{out}} = F_\phi(\bar{s}_K, c). \tag{15}$$

Conditioned on c, independent generator outputs satisfy

$$\text{Cov}(\bar{s}_K \mid c) = \frac{1}{K} \text{Cov}(s_1 \mid c). \tag{16}$$

3 Experiments

3.1 Experimental setup

Datasets. Following the protocol of (Quan et al., 2022), we construct the Places2 (Zhou et al., 2018) training and test sets by randomly selecting 2,000 training images and 100 test images from each of the first 20 categories, yielding 40,000 training images and 2,000 test images. For CelebA-HQ, we randomly select 28,000 images for training and 2,000 images for testing. Training masks are drawn from the Quick Draw Irregular Mask Dataset. For testing, we use the irregular masks introduced by (Liu et al., 2018), comprising six mask-coverage intervals: $(0.01, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.4]$, $(0.4, 0.5]$, and $(0.5, 0.6]$, with 2,000 masks per interval. Each test image is evaluated with one mask from each interval, resulting in 12,000 image and mask pairs per benchmark. All images and masks are resized to 256×256.

Implementation detail. Experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU. We use SD-Turbo as the pretrained backbone and adapt it with a four-channel ControlNet and rank-8 LoRA adapters at the fixed timestep $t = 999$. The base U-Net, VAE, and text encoder remain frozen, and the text prompt is empty. Stage A uses AdamW with a batch size of 16 and a learning rate of 2×10^{-5}. The weights for noise-prediction, clean-latent reconstruction, and decoded-image reconstruction losses are set to 0.2, 1.0, and 1.0, respectively. In Stage B, the adapted generator is frozen, and the residual refiner is trained using AdamW with a batch size of 8 and a learning rate of 10^{-4}. Unless otherwise specified, inference averages $K = 4$ independently sampled one-step predictions before applying the refiner once, corresponding to four denoiser evaluations and one refinement pass. A second-backbone probe based on SDXL-Turbo is provided in Appendix B

Compared Methods and Evaluation Metrics. We compare our method with TFill (Zheng et al., 2022), W-Net (Zhang et al., 2023b), SCAT (Zuo et al., 2023), StrDiffusion (Liu et al., 2024), GraphFill (Verma et al., 2024), FDM (Park & Jung, 2025), and APLRL (Cao et al., 2025), using their publicly released pretrained weights. On each benchmark, all compared methods are evaluated on the same test images and masks, and their outputs are scored using a common metric implementation. We report PSNR, SSIM, L1 error, LPIPS, and FID. Higher PSNR and SSIM and lower L1, LPIPS, and FID indicate better performance. Hole only PSNR is additionally reported in the component ablations to assess reconstruction quality within missing regions.

3.2 Comparison experiments

Quantitative experiment. Tables 1 and 2 report quantitative comparisons on Places2 and CelebA-HQ. On Places2, our method achieves the best PSNR, SSIM, and L1 scores, reaching 26.91 dB, 0.8727, and 0.0218, respectively. Compared with the strongest baseline on these reconstruction-oriented metrics, it improves PSNR by 1.18 dB and SSIM by 0.0199, while reducing L1 error from 0.0245 to 0.0218. On CelebA-HQ, our method also obtains the highest PSNR and SSIM, with 30.04 dB and 0.9075, demonstrating that the reconstruction improvements generalize across datasets. Although the best LPIPS and FID values are achieved by other methods, our method remains competitive on these perceptual metrics, indicating a favorable balance between reconstruction accuracy and perceptual quality. The reconstruction gains are statistically significant under a paired bootstrap test (Appendix C) and stable across three full-pipeline training seeds (Appendix D). Zero-shot transfer to 512×512 inputs and out-of-distribution mask shapes are evaluated in Appendix E.

Table 1: Quantitative comparison on Places2. ↑ denotes that higher is better; ↓ denotes that lower is better. Bold denotes the best result.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	L1↓
TFill	25.04	0.8454	0.1192	11.51	0.0263
W-Net	25.59	0.8501	0.1244	12.02	0.0245
SCAT	25.23	0.8263	**0.1067**	11.72	0.0326
StrDiffusion	24.99	0.8426	0.1195	**10.95**	0.0275
GraphFill	25.73	0.8528	0.1140	11.33	0.0250
FDM	23.40	0.8192	0.1526	11.91	0.0347
Ours	**26.91**	**0.8727**	0.1247	11.25	**0.0218**

Table 2: Quantitative comparison on CelebA-HQ. ↑ denotes that higher is better; ↓ denotes that lower is better. Bold denotes the best result.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	L1↓
TFill	28.02	0.8802	0.0701	8.77	0.0186
W-Net	29.46	0.8970	**0.0584**	8.25	**0.0149**
GraphFill	27.96	0.8769	0.0818	9.86	0.0197
APLRL	29.67	0.8926	0.0602	**7.99**	0.0153
Ours	**30.04**	**0.9075**	0.0656	8.79	0.0153

Qualitative experiment. Figures 2 and 3 present qualitative comparisons on CelebA-HQ and Places2, respectively. On CelebA-HQ, the proposed method produces visually coherent facial structures and preserves fine details around the eyes, mouth, hair, and facial contours, while reducing noticeable distortions and local artifacts observed in some competing methods. On Places2, our method also demonstrates stable reconstruction across diverse scenes, maintaining better structural continuity and texture consistency in large irregular missing regions. In particular, the restored content is more naturally integrated with the surrounding known regions, with fewer obvious semantic inconsistencies, texture discontinuities, and color artifacts in the shown examples. These qualitative

results are consistent with the quantitative improvements and indicate that the proposed method can achieve reliable reconstruction while preserving overall visual realism.

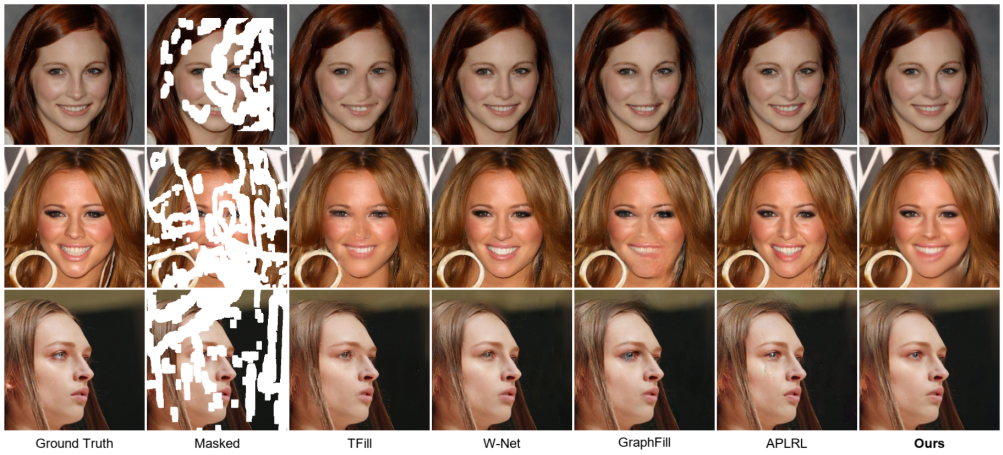

Figure 2: Qualitative comparison on CelebA-HQ.

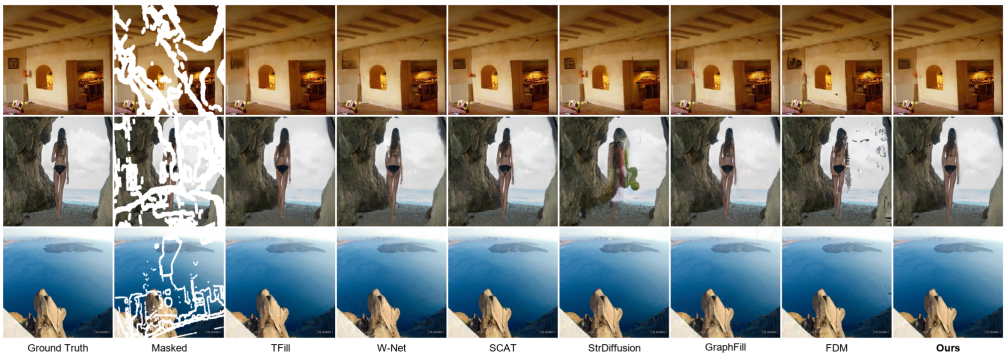

Figure 3: Qualitative comparison on Places2.

3.3 Ablation Experiments

Table 3: Stage A objective and target study on the Places2. ϵ represents noise prediction, the pretraining target; v represents velocity prediction; z_0 denotes direct clean-latent prediction. 20 epochs for recipe screening and 46 epochs matching the full Stage A schedule. *latent term scaled by $1/\kappa_\epsilon$.

Target	Pred.	Latent	Image	Epochs	Val. PSNR
ϵ	✓			20	25.58
ϵ	✓	✓		20	25.81
ϵ	✓	✓	✓	20	26.05
ϵ		✓	✓	20	25.80
ϵ $1/\kappa_\epsilon$	✓	✓*		20	25.66
v	✓	✓		20	23.43
v	✓	✓	✓	20	23.82
z_0	✓	✓		20	25.60
ϵ	✓	✓	✓	46	26.66
v	✓	✓		46	25.43
v	✓	✓	✓	46	25.49

Reconstruction-space supervision. Table 3 shows that reconstruction-space supervision consistently improves the pretrained noise-prediction model, increasing validation PSNR from 25.58 to 26.05 dB. In contrast, switching to velocity prediction yields lower performance under both 20 and 46 epoch. These results support retaining the pretrained prediction target while extending supervision into the reconstruction space.

Residual refinement. Table 4 shows that residual refinement substantially improves the generator output, increasing PSNR from 25.803 to 26.905 dB and reducing LPIPS from 0.1717 to 0.1247.

The generic refiner already provides most of the gain, while area normalization slightly improves reconstruction fidelity. The final configuration provides a balanced trade-off between reconstruction accuracy and perceptual quality.

Table 4: Stage B refinement ablation on Places2.

Configuration	PSNR	SSIM	LPIPS	Hole PSNR
Generator only	25.803	0.8520	0.1717	19.40
Generic refiner	26.885	0.8722	0.1207	20.48
Area-normalized	26.920	0.8736	0.1497	20.52
Area-normalized + LPIPS (final)	26.905	0.8727	0.1247	20.50

Sampling stability and efficiency. Table 5 shows that increasing K consistently reduces pre-refinement output variance, closely following the expected $1/K$ trend. Sample averaging provides modest gains in PSNR and SSIM, while inference time increases nearly linearly with K. We therefore use $K = 4$ as a practical trade-off between output stability and computational cost. The measured variance-reduction curve is shown in Appendix F.

Table 5: Inference-time sample-averaging study on Places2. V_K/V_1 denotes the pre-refinement hole-pixel variance relative to $K = 1$. Results are measured using the full pipeline on a single RTX 4090 with batch size 1.

K (=NFE)	Gen. PSNR	Final PSNR	SSIM	LPIPS	V_K/V_1	Time (ms)
1	25.689	26.870	0.8719	0.1221	1.0000	107.4
2	25.762	26.892	0.8724	0.1235	0.5020	209.2
4	25.801	26.904	0.8727	0.1247	0.2517	413.8
8	25.819	26.910	0.8728	0.1256	0.1255	818.9

3.4 Prediction error propagation and target discrepancy

As derived in Section 2.2, at $t = 999$, velocity prediction has substantially lower theoretical reconstruction-error amplification than noise prediction ($\kappa_v \approx 1$ versus $\kappa_\epsilon \approx 14.6$). However, Table 3 shows that this theoretical advantage does not translate into better downstream adaptation. Under the 20 epoch, velocity prediction reaches only 23.82 dB with latent and image reconstruction supervision, compared with 26.05 dB for noise prediction, and the gap remains after 46 epochs (25.49 versus 26.66 dB). Gradient, curvature, and representation-level diagnostics of this mismatch are reported in Appendix G. These results confirm that theoretical error propagation alone is insufficient to determine practical adaptation performance, motivating our strategy of retaining the pretrained noise-prediction target while extending supervision into the reconstruction space.

4 Conclusion

In this work, we investigate one-step diffusion inpainting from the perspective of prediction-target error propagation and downstream adaptation. Our analysis shows that lower theoretical reconstruction-error amplification does not necessarily translate into better practical adaptation performance for a pretrained diffusion model. Based on this observation, we propose reconstruction-space adaptation, which retains the pretrained noise-prediction target while extending supervision to the reconstructed clean latent and decoded image, thereby aligning downstream optimization more directly with the final restoration objective. A lightweight residual refiner and independent-sample averaging are further introduced as complementary components to correct remaining image-space errors and reduce sampling-induced output fluctuations. Extensive experiments on Places2 and CelebA-HQ demonstrate consistent reconstruction improvements under controlled adaptation settings and competitive inpainting performance. These results highlight the importance of considering both theoretical error propagation and pretrained adaptation behavior when designing efficient one-step diffusion inpainting methods.

AI use statement

Large language models (LLMs) were used only for language polishing, manuscript consistency and error checking, and assisting with the organization and documentation of the accompanying open-source GitHub code package. They were not used to generate research ideas, develop methodology,

analyze results, or formulate scientific conclusions. All AI-assisted content was reviewed and revised by the authors. The use of LLMs did not affect the main methodology, contributions, or originality of this work, and the authors assume full responsibility for the final manuscript.

ETHICS STATEMENT

This work focuses on image inpainting and does not involve human-subject experiments or new data collection. All experiments are conducted using existing public benchmark datasets, including Places2 and CelebA-HQ, together with publicly available mask datasets. The proposed method is intended for general-purpose image restoration and completion; however, as with other image generation and editing techniques, it could potentially be misused to create misleading or manipulated visual content. We therefore encourage responsible use of the method and appropriate disclosure when generated or restored images are used in contexts where authenticity is important.

REPRODUCIBILITY STATEMENT

We place a strong emphasis on reproducibility. We will provide the complete source code of our method, including Stage A reconstruction-space adaptation, Stage B residual refinement, and independent-sample averaging at inference time.

REFERENCES

Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 417–424, 2000.

Feilong Cao, Qijin Xu, and Hailiang Ye. Adaptive prior and long-range dependency-based learners for image inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11): 10742–10755, 2025.

Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR 2022*, April 2022.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36:1 – 14, 2017.

Qing Jin and Chaoyang Wang. Revisiting diffusion model predictions through dimensionality. *arXiv preprint arXiv:2601.21419*, 2026.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pp. 150–168. Springer, 2024.

Jungmin Lee and Jongwon Choi. Visual design from cultural heritage through inpainting-based restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4491–4496, 2025.

Yunsung Lee and Hyeongmin Lee. Not all prediction targets keep training-free diffusion guidance on the manifold. *arXiv preprint arXiv:2607.00647*, 2026.

Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10748–10758, 2022. doi: 10.1109/CVPR52688.2022.01049.

Yaowei Li, Yuxuan Bian, Xuan Ju, Zhaoyang Zhang, Junhao Zhuang, Ying Shan, Yuexian Zou, and Qiang Xu. Brushedit: All-in-one image inpainting and editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):11746–11757, 2026. doi: 10.1109/TPAMI.2026.3695011.

Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 89–105, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01252-6.

Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8038–8047. IEEE, 2024.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11451–11461. IEEE, 2022.

Kyungri Park and Woohwan Jung. Improving detail in pluralistic image inpainting with feature dequantization. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 680–689. IEEE, 2025.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Weize Quan, Ruisong Zhang, Yong Zhang, Zhifeng Li, Jue Wang, and Dong-Ming Yan. Image inpainting with local and global refinement. *IEEE Transactions on Image Processing*, 31:2405–2420, 2022. doi: 10.1109/TIP.2022.3152624.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 87–103, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73016-0.

Shashikant Verma, Aman Sharma, Roopa Sheshadri, and Shanmuganathan Raman. Graphfill: Deep image inpainting using graphs. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4984–4994. IEEE, 2024.

Duc Vu, Kien Nguyen, Trong-Tung Nguyen, Ngan Nguyen, Phong Nguyen, Khoi Nguyen, Cuong Pham, and Anh Tran. Inverfill: One-step inversion for enhanced few-step diffusion inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25677–25687, June 2026.

Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4672–4681, 2021. doi: 10.1109/ICCV48922.2021.00465.

Yinhu Wu, Junping Zhang, and Dongyang Liu. Predictive filtering integrated generative remote sensing hyperspectral image inpainting. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025. doi: 10.1109/LGRS.2025.3526177.

Hanyu Xiang, Qin Zou, Muhammad Ali Nawaz, Xianfeng Huang, Fan Zhang, and Hongkai Yu. Deep learning for image inpainting: A survey. *Pattern Recognition*, 134:109046, 2023. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2022.109046.

Liangbin Xie, Daniil Pakhomov, Zhonghao Wang, Zongze Wu, Ziyan Chen, Yuqian Zhou, Haitian Zheng, Zhifei Zhang, Zhe Lin, Jiantao Zhou, and Chao Dong. Turbofill: Adapting few-step text-to-image model for fast image inpainting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7613–7622, 2025. doi: 10.1109/CVPR52734.2025.00713.

Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22428–22437. IEEE, 2023.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5505–5514, 2018. doi: 10.1109/CVPR.2018.00577.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023a. doi: 10.1109/ICCV51070.2023.00355.

Ruisong Zhang, Weize Quan, Yong Zhang, Jue Wang, and Dong-Ming Yan. W-net: Structure and texture interaction for image inpainting. *IEEE Transactions on Multimedia*, 25:7299–7310, 2023b. doi: 10.1109/TMM.2022.3219728.

Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. Bridging global context interactions for high-fidelity image completion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11502–11512, 2022. doi: 10.1109/CVPR52688.2022.01122.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

Qinming Zhou, Chenxi Sun, Deyang Kong, Junhao He, Xiangheng Tang, Peike Yu, Haotian Wu, Leilei Cao, and Linfeng Zhang. Osor: One-step diffusion inpainting for effect-aware object removal. In *European Conference on Computer Vision*, pp. 367–384. Springer, 2026.

Zhiwen Zuo, Lei Zhao, Ailin Li, Zhizhong Wang, Zhanjie Zhang, Jiafu Chen, Wei Xing, and Dongming Lu. Generative image inpainting with segmentation confusion adversarial training and contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3888–3896, 2023.

A Full-Budget Parameterization Study

Section 2.2 reports that velocity prediction trails noise prediction under the 20-epoch screening budget. To check whether this ordering is an artifact of the training budget, both velocity arms were further trained for the full 46-epoch Stage A schedule. As shown in Fig. 4, the velocity arms improve steadily but remain at 25.43 dB (with latent supervision) and 25.49 dB (with latent and image supervision), versus 26.66 dB for the noise-prediction arm at the same budget, a gap of 1.17–1.23 dB. Two controls support reading this gap as a budget-robust adaptation cost rather than an optimization artifact: halving the learning rate costs the noise-prediction arm only 0.64 dB, and non-finite-loss skip events occur only for the velocity arms. The velocity curves have not fully asymptoted at 46 epochs, so the reported gap should be read as a large budget-robust lower bound under the evaluated conditions rather than a proof of structural impossibility.

B Second-Backbone Probe: SDXL-Turbo

To examine whether the conclusions depend on the specific backbone, we repeat the central comparison on SDXL-Turbo with LoRA adapters under the same protocol at a 20-epoch budget (Fig. 5). With the uncalibrated loss weights used for SD-Turbo, the reconstruction-supervised arm trails the

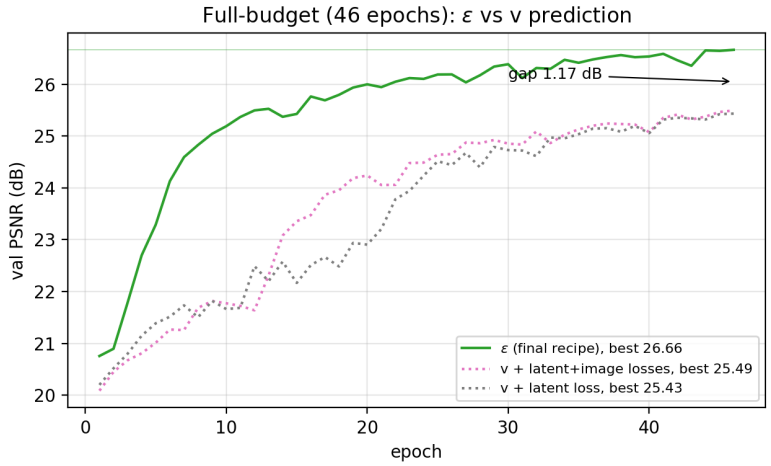

Figure 4: Validation PSNR over the full 46-epoch Stage A schedule for the noise-prediction arm and the two velocity-prediction arms. The ordering established at the 20-epoch screening budget persists at the full budget.

noise-only arm by 0.18 dB (24.27 versus 24.45 validation PSNR). A weight-mismatch analysis isolates the cause: the SDXL VAE latent scale is $1.4\times$ larger, so the clean-latent term initially dominates the loss; with the latent weight calibrated to 0.2, the ordering reverses at equal budget (22.28 versus 21.85 dB at 3 epochs), and the uncalibrated arm reaches parity at 6 epochs (23.59 versus 23.57 dB). The reconstruction-gain analysis of Section 2.2 is schedule- and backbone-independent and transfers directly, while the loss-recipe transfer requires weight calibration across backbones.

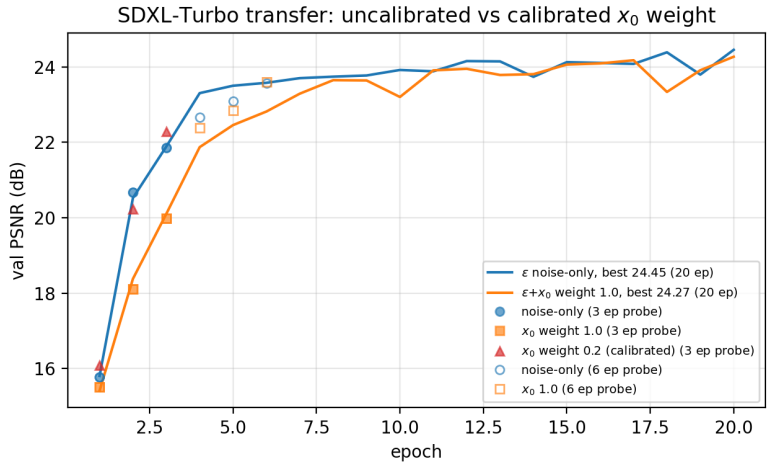

Figure 5: SDXL-Turbo transfer probe. Solid lines: 20-epoch arms with uncalibrated weights. Markers: short-budget probes isolating the latent-weight mismatch (B-series), including the calibrated-weight arm (triangles) and extended-budget arms (hollow markers).

C Statistical Significance of the Main Margins

We assess the significance of the Places2 margins with a paired bootstrap test (10,000 resamples over the 12,000 image–mask pairs). Table 6 reports the PSNR margin of our method over each compared method with 95% confidence intervals; all intervals exclude zero. SSIM and L1 margins are likewise significant with the expected signs, while LPIPS margins show the opposite sign for most compared methods, consistent with the fidelity–perception balance discussed in Section 3.2. FID has no per-image decomposition and is therefore outside the bootstrap protocol.

Table 6: Paired bootstrap PSNR margins (Ours and compared methods) on Places2 with 95% confidence intervals, 10,000 resamples over 12,000 pairs.

Compared method	ΔPSNR (dB)	95% CI
GraphFill	$+1.17$	$[+1.16, +1.19]$
W-Net	$+1.31$	$[+1.29, +1.34]$
SCAT	$+1.67$	$[+1.63, +1.72]$
TFill	$+1.87$	$[+1.84, +1.89]$
StrDiffusion	$+1.91$	$[+1.86, +1.97]$
FDM	$+3.51$	$[+3.48, +3.54]$

Table 7: Full-pipeline test PSNR (dB) across three training seeds on Places2 (12,000 pairs each).

Seed	42	43	44
PSNR	26.905	26.876	26.894

D Robustness across Training Seeds

The full pipeline (Stage A, refiner, and evaluation) was repeated with three training seeds on Places2. Test PSNR varies by only 0.012 dB across seeds (Table 7), over an order of magnitude below the smallest claimed margin in Table 1, indicating that the reported improvements are not seed-dependent.

E Generalization beyond the Training Distribution

We evaluate the Places2-trained pipeline in two transfer settings without any fine-tuning (Fig. 6). First, zero-shot inference at 512×512, double the training resolution, retains full quality: 27.14 dB PSNR, 0.8924 SSIM, 0.1492 LPIPS, and 10.67 FID, versus 26.91 dB / 0.8727 / 0.1247 / 11.25 at 256×256, indicating that the one-step control pathway generalizes across resolutions. Second, under out-of-distribution large rectangular masks (a single contiguous box per image instead of irregular brush strokes), performance drops to 20.32 dB PSNR, 0.7980 SSIM, 0.2448 LPIPS, and 16.56 FID. Large contiguous holes remain the hardest setting for the one-step design, and we report this boundary explicitly rather than tuning toward it.

F Sampling Variance across K

Figure 7 shows the measured pre-refinement hole-pixel variance of the averaged output as a function of K, normalized by the $K = 1$ variance. The curve closely follows the theoretical $1/K$ trend of Eq. equation 16, confirming that independent-sample averaging suppresses sampling-induced fluctuations as predicted, with diminishing returns beyond $K = 4$.

G Objective-Mismatch Diagnostics

To characterize why velocity prediction adapts poorly despite its lower theoretical error amplification, we probe the converged noise-prediction model at $t = 999$ (Fig. 8). At the noise-prediction solution, the velocity loss exhibits a $559\times$ larger gradient norm than the noise loss (1.21 versus 2.2×10^{-3}) and a $13\times$ sharper Hessian top eigenvalue (14.6 versus 1.1, Lanczos power iteration with 100 Hessian-vector products), indicating that the velocity objective is far from stationary at the converged noise-prediction solution. Moreover, mid-block features of identically configured noise- and velocity-trained models are nearly orthogonal (linear CKA $= 0.003$), indicating that the two prediction targets induce essentially different learned representations. These diagnostics support the interpretation that the adaptation gap reflects a structural mismatch between the velocity objective and the noise-prediction pretraining, rather than a training-budget artifact.

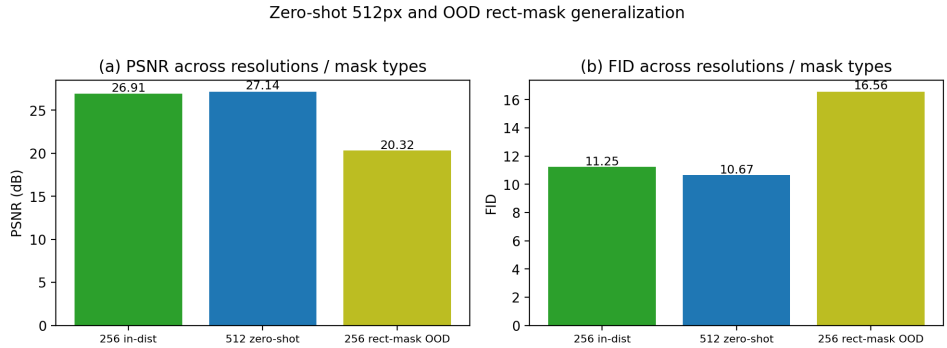

Figure 6: Zero-shot 512×512 and out-of-distribution rectangular-mask evaluation of the Places2-trained pipeline (no fine-tuning in either setting).

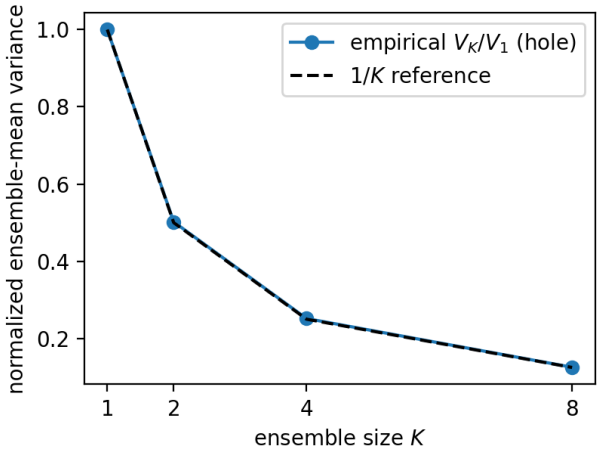

Figure 7: Measured pre-refinement output variance relative to $K = 1$ as a function of the number of averaged independent samples K, compared with the theoretical $1/K$ curve.

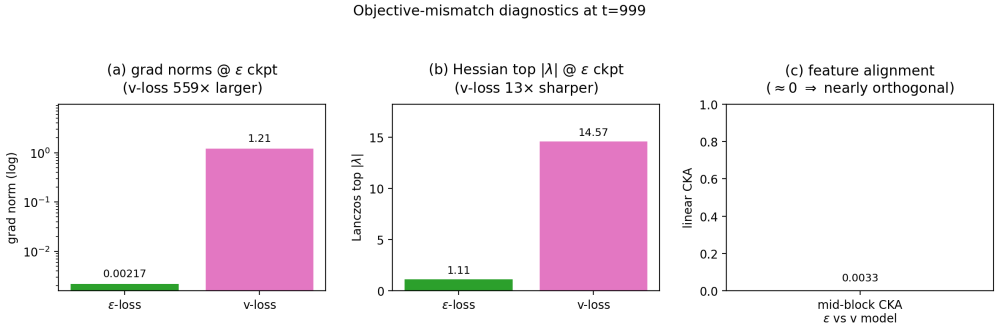

Figure 8: Objective-mismatch diagnostics at $t = 999$. (a) Gradient norms and (b) Hessian top eigenvalues of the noise and velocity losses at the converged noise-prediction solution; (c) linear CKA between mid-block features of noise- and velocity-trained models.