Beyond Single-Shot Conditioning: Test-Time Condition Refinement for Diffusion-Based Image Restoration

Aiping Zhang[1] **Jiangang Wang**[1,2] **Shangquan Sun**[3] **Yuning Cui**[4]
Fan Li[5] **Renjing Pei**[5] **Wenqi Ren**[1*] **Xiaochun Cao**[1]

[1]Shenzhen Campus of Sun Yat-sen University
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[3]Nanyang Technological University [4]Technical University of Munich
[5]Huawei Noah's Ark Lab

{zhangaip7@mail2,renwq3@mail}.sysu.edu.cn

Abstract

Conditional diffusion models have emerged as a powerful paradigm for image restoration, leveraging pre-trained generative priors to recover high-fidelity details from degraded inputs. However, existing methods typically follow a static inference protocol, treating the degraded observation as a fixed control signal throughout sampling. This overlooks the potential of inference-time scaling, where guidance can be refined on the fly to improve output quality. Recent test-time optimization approaches that manipulate noise variables or denoising trajectories often compromise structural fidelity, while those targeting text embeddings lack the spatial precision needed for restoration. To address this, we propose Condition-Aware Test-Time Optimization (CATTO), a training-free method that iteratively refines the visual conditioning signal at inference time. CATTO performs reward-aligned condition refinement under the pre-trained prior to enhance perceptual quality. To avoid complex backpropagation, we design a gradient-free optimization strategy guided by a joint objective: maximizing a perceptual reward while enforcing trajectory- and condition-level consistency to explicitly preserve structural fidelity. This process is accelerated by optimizing within a low-dimensional frequency subspace and reusing reliable updates across nearby denoising steps. Experiments on image restoration benchmarks show that CATTO improves perceptual quality over strong diffusion-based baselines without updating model parameters.

1 Introduction

Image restoration (IR) [1, 2] aims to recover clean images from their degraded counterparts, serving as a fundamental task in computer vision. Recently, Diffusion Probabilistic Models (DPMs) [3, 4, 5, 6, 7, 8] have demonstrated strong capabilities in IR. Owing to their powerful generative priors, diffusion-based approaches effectively traverse the complex natural image manifold and synthesize visually plausible details [9, 10, 11, 12]. Given a degraded observation c, diffusion-based image restoration aims to recover a clean image x_0 by modeling the conditional distribution $p_\theta(x_0 \mid c)$, where c guides the denoising process toward solutions. Nevertheless, standard posterior sampling in diffusion-based IR often leads to distortions and artifacts, reflecting the inherent fidelity-perception trade-off [13, 9, 10, 14, 15]. Existing approaches commonly rely on retraining, architectural modifications,

*Corresponding author.

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

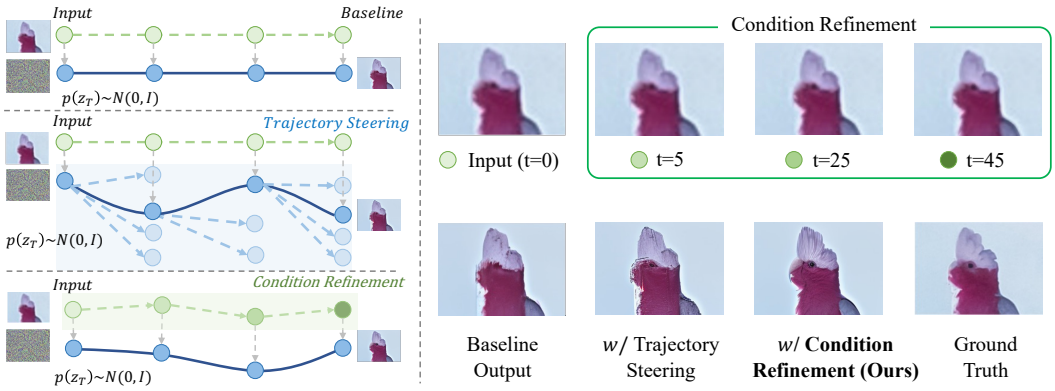

Figure 1: Motivation of condition-aware test-time optimization. Left: Comparison of different optimization targets during inference in diffusion-based IR. Right: Visual comparison. The top row visualizes the input image condition and our refined conditions at $t = 5$, $t = 25$, and $t = 45$. Ours better preserves input-consistent eye and feather structures.

or larger DPMs to alleviate this issue. However, they constrain the model to a static trade-off point that lacks flexibility for instance-specific restoration needs, in addition to incurring substantial training and deployment costs [9, 14, 10, 16]. In contrast, test-time optimization [17, 18, 19, 20, 21, 22] provides a training-free solution to improve the fidelity-perception trade-off by allocating additional computational budget at inference time.

Building on this line of work, existing test-time methods improve diffusion outputs by optimizing noise variables [19, 23] or by searching and steering denoising trajectories [17, 18, 24, 20]. While effective for reward-aligned generation, directly applying these methods to IR can yield visually improved results but with compromised structural fidelity, as shown in Fig. 1. A plausible explanation is that they optimize the sampling process toward perceptually preferred outputs, rather than explicitly enforcing structural consistency with the input, as shown in Appendix Fig. 5. Meanwhile, existing diffusion-based IR methods typically adopt a *single-shot conditioning* paradigm, where the degraded observation is encoded once as the visual condition and then kept fixed throughout sampling [11, 9, 10, 14]. In real-world restoration, however, such fixed conditioning can be unreliable because complex degradations often distort local structures and make the visual condition ambiguous [10]. As a result, imperfect guidance can persist throughout denoising and degrade restoration quality. Crucially, unlike raw noise variables or denoising trajectories, the degraded observation serves as the primary conditioning signal in IR and inherently encodes restoration-relevant information, including spatial structures and degradation cues [9, 10, 14]. Furthermore, because these image conditions are injected through dedicated conditioning pathways [11, 14, 10], optimizing them provides a structured, task-aligned interface that explicitly controls restoration guidance. Therefore, the above observations motivate us to shift the optimization target to the image condition itself, offering a more principled way to achieve structure-aware perceptual refinement without blindly disrupting structural integrity.

While prior text-to-image works have explored test-time optimization of text conditional embeddings [25, 26], such methods offer limited control over the fine-grained spatial structures critical for image reconstruction. To this end, we propose **C**ondition-**A**ware **T**est-**T**ime **O**ptimization (**CATTO**), a training-free and plug-and-play framework. It moves beyond *single-shot conditioning* by iteratively refining the input *image* condition throughout sampling. Specifically, to effectively navigate the fidelity-perception trade-off, CATTO formulates a synergistic condition refinement objective driven by three concurrent components: a perceptual reward to synthesize visually pleasing details, a trajectory consistency term to maintain prior-aligned denoising dynamics, and a condition fidelity term to explicitly preserve the structural content of the degraded observation. Rather than acting as post-hoc constraints, these components operate in tandem to guide condition refinement, ensuring the final restoration is both perceptually rich and structurally faithful.

To optimize this objective efficiently, CATTO combines a gradient-free optimizer with a compact condition search strategy. This design sidesteps challenges in existing reward-aligned inference methods, such as non-differentiable rewards [19] and costly backpropagation [27]. Accordingly, CATTO treats rewards as black-box scoring signals and estimates condition update directions via zero-

order optimization. Moreover, to reduce repeated exploration cost in existing inference-time methods [19, 18, 20, 21], CATTO performs optimization in a low-dimensional frequency subspace [28] and reuses reliable updates across nearby denoising steps through momentum caching.

Our main contributions are summarized as follows:

- We propose CATTO, a training-free, plug-and-play test-time framework that optimizes the image condition to refine restoration guidance, enabling structure-aware perceptual improvement while preserving structural fidelity.
- We design a fidelity-aware condition refinement objective that combines perceptual reward maximization with trajectory consistency and condition fidelity regularization, improving the fidelity-perception trade-off.
- We develop an efficient gradient-free optimizer for black-box perceptual rewards, together with a compact search strategy that restricts condition refinement to a low-dimensional frequency subspace and reuses updates across nearby denoising steps.
- Extensive experiments show that CATTO improves restoration quality across multiple tasks and baseline models, achieving a better balance between perceptual quality and structural fidelity.

2 Related Work

2.1 Diffusion Models for Image Restoration

Diffusion models have emerged as a dominant and versatile framework for image restoration [29]. Existing approaches can be broadly categorized into two groups. Unconditional methods [30, 31, 32, 33, 34, 35] reuse a pretrained unconditional diffusion model as a clean-image prior $p(x)$, incorporating measurements only at inference time by modifying the reverse sampling dynamics. Despite their strong transferability and zero-shot flexibility, these methods rely on accurate knowledge of the degradation process, which is often unavailable in real-world image restoration tasks [36]. In contrast, conditional methods [37, 9, 10, 38, 39, 40] learn restoration models conditioned on degraded observations, which helps better constrain the posterior and often improves stability and reconstruction quality. Still, restoration remains challenging under complex degradations, where the conditional information can be insufficient or ambiguous, thereby limiting performance.

2.2 Inference-time Optimization in Diffusion Models

Recent research improves test-time generation by trading additional inference-time computation for better task alignment and restoration quality. Broadly, these efforts follow two main strategies: directly optimizing or searching variables in the sampling process and scaling the denoising trajectory. DNO [19] and InitNO [23] directly optimize the noise for reward alignment or prompt-image consistency, while variational formulations [41, 42, 43] balance data fidelity with diffusion-prior regularization. Searching and steering methods instead improve sampling through trajectory exploration and budget reallocation. DSearch [18] performs dynamic tree search, Noise Trajectory Search [24] formulates noise selection as sequential decision making, and FK-steering [20] uses SMC-based particle resampling to favor high-reward trajectories. Other scaling methods [20, 44, 17] further allocate computation to candidate generation, verifier guidance, particle steering, or reweighting. For specialized IR methods, KDS [21] improves reconstruction by steering multi-particle samples, which biases the solution toward a high-density mode of the candidate-particle distribution. RFJS [22] uses external side information, such as reference images or textual descriptions, to guide trajectory search at test time, but such information is often unavailable in practice. Despite their flexibility, these methods can be computationally costly, and optimization in high-dimensional noise or latent spaces may lead to unstable updates, artifacts, or reward over-optimization.

3 Method

3.1 Preliminaries

Latent Diffusion Models [6]. Operating in the compressed latent space of a pre-trained VAE, an encoder $\mathcal{E}$ maps an image x to $z_0 = \mathcal{E}(x)$. The forward diffusion process then progressively injects

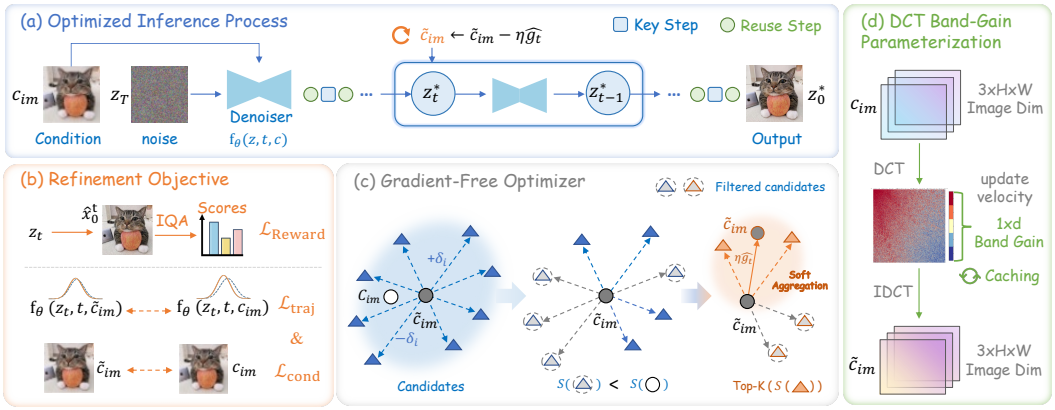

Figure 2: Overview of CATTO. **(a) Optimized Inference Process:** We treat the visual condition as an optimizable variable $\tilde{c}_{im}$. During sampling, $\tilde{c}_{im}$ is iteratively refined to improve reward and consistency and the pre-trained denoiser f_θ is frozen. **(b) Refinement Objective:** The objective balances three terms: perceptual reward maximization ($\mathcal{L}_{\text{reward}}$) on the estimated clean image, trajectory consistency ($\mathcal{L}_{\text{traj}}$) to adhere to the generative prior, and fidelity to the input condition ($\mathcal{L}_{\text{cond}}$) to the original input. **(c) Gradient-Free Optimizer:** To handle non-differentiable rewards, we use antithetic perturbations for gradient-free condition search. We use a two-stage scoring to filter candidates and aggregate the Top-K feasible samples to estimate the update direction. **(d) DCT Band-Gain Parameterization:** DCT band-gain caching with subspace-constrained optimization in a compact frequency space.

Gaussian noise into z_0 to produce z_t. A denoising network is trained to predict either the added noise or the velocity, parameterized as $\epsilon_\theta(z_t, t, c)$ or $v_\theta(z_t, t, c)$, where c denotes the conditioning signals.

Conditional Generation [45, 46]. In tasks such as image restoration [37, 9, 10, 38, 39, 40, 47], the network is conditioned on a visual input c_{im}, e.g., a degraded image. The denoising model is therefore parameterized as $\epsilon_\theta(z_t, t, c_{\text{im}})$ or $v_\theta(z_t, t, c_{\text{im}})$, where conditioning is primarily derived from the degraded visual input rather than an external textual condition [14, 10, 38, 40]. During inference, the visual condition c_{im} is kept fixed. Our objective is to learn an optimized condition $\tilde{c}_{\text{im}}$ that steers the generation toward higher perceptual rewards while preserving fidelity to the original input, thereby achieving a better fidelity–perception trade-off.

3.2 Condition Refinement Method

Recent training-free inference-time strategies [21, 48, 18, 49, 19, 20] demonstrate that allocating additional computation during inference can improve diffusion outputs. These approaches typically operate on the sampling process by optimizing latent states z_t or steering denoising trajectories $\{z_t\}_{t=0}^{T}$. To tailor test-time optimization to IR, we instead optimize the visual condition c_{im}, which lies on a more structured manifold [50] and retains spatial guidance from the degraded observation. As illustrated in Fig. 2, CATTO iteratively refines the input image condition during sampling to provide adaptive restoration guidance without directly perturbing denoising states.

3.2.1 Fidelity-Aware Refinement Objective

We aim to derive a surrogate condition $\tilde{c}_{\text{im}}$ that maximizes a reward function $r(z_0)$ reflecting perceptual quality [20, 19]. To prevent reward hacking and preserve realism, we define the optimal target distribution $p_{\text{tar}}(z_0)$ as a trade-off between reward maximization and fidelity to the pretrained prior $p_{\text{pre}} = p_\theta(\cdot \mid c_{\text{im}})$ as:

$$p_{\text{tar}}(z_0) = \arg\max_p \Big(\mathbb{E}_{z_0 \sim p}[r(z_0)] - \alpha D_{\text{KL}}(p \| p_{\text{pre}}) \Big). \tag{1}$$

Because directly sampling from p_{tar} is intractable, we choose to optimize $\tilde{c}_{\text{im}}$ such that its induced distribution $p_\theta(\cdot \mid \tilde{c}_{\text{im}})$ approximates p_{tar}. Here, we associate the refined and original conditions with Gaussian densities $q_{\tilde{c}}(c)$ and $p_c(c)$, centered at $\tilde{c}_{\text{im}}$ and c_{im}, respectively. The variable c denotes

a condition sampled from these densities, while $\tilde{c}_{\text{im}}$ remains the optimization variable. We define the joint distributions $Q_{\tilde{c}}(z_{0:T}, c) = q_{\tilde{c}}(c)p_\theta(z_{0:T} \mid c)$ and $P(z_{0:T}, c) = p_c(c)p_\theta(z_{0:T} \mid c_{\text{im}})$. To regularize this optimization, we minimize their joint KL divergence:

$$\min_{\tilde{c}_{\text{im}}} D_{\text{KL}}\Big(Q_{\tilde{c}}(z_{0:T}, c) \,\Big\|\, P(z_{0:T}, c)\Big). \tag{2}$$

Using the KL chain rule and the Markov property of diffusion models [51], and assuming a shared condition-independent terminal distribution, we decompose this joint KL as:

$$D_{\text{KL}}(Q_{\tilde{c}} \| P) = \sum_{t=1}^{T} \mathbb{E}_{c\sim q_{\tilde{c}}} \mathbb{E}_{z_t \sim p_\theta(z_t|c)} \left[D_{\text{KL}}(p_\theta(z_{t-1} \mid z_t, c) \,\|\, p_\theta(z_{t-1} \mid z_t, c_{\text{im}})) \right] + D_{\text{KL}}(q_{\tilde{c}} \| p_c). \tag{3}$$

Here, we approximate the expectation over c by evaluating the trajectory term at its mean $\tilde{c}_{\text{im}}$, and use an empirical distance $\mathcal{D}(\tilde{c}_{\text{im}}, c_{\text{im}})$ as a surrogate for the condition-level KL term to enforce fidelity.

To seamlessly unify discrete (e.g., DDPM [4]/DDIM [3]) and continuous (e.g., Flow Matching [5]) generative paradigms, we formulate the reverse transition via a generalized *unified predictor* $\mathbf{f}_\theta$:

$$p_\theta(z_{t-1} \mid z_t, c) = \mathcal{N}\big(b_t(z_t) + \lambda_t \mathbf{f}_\theta(z_t, t, c), \sigma_t^2 \mathbf{I}\big). \tag{4}$$

For continuous models, we discretize time into sampling steps. When $\sigma_t = 0$, the update is deterministic. For non-degenerate transitions with $\sigma_t > 0$, the step-wise KL divergence exactly reduces to a weighted L_2 distance between predictors:

$$\ell_t = w_t \|\mathbf{f}_\theta(z_t, t, \tilde{c}_{\text{im}}) - \mathbf{f}_\theta(z_t, t, c_{\text{im}})\|_2^2, \tag{5}$$

where $w_t = \lambda_t^2/(2\sigma_t^2)$ for $\sigma_t > 0$, and the predictor $\mathbf{f}_\theta$ is instantiated according to the specific sampling formulation (see Tab. 1). For deterministic updates, we use the same L_2 distance as a regularizer with finite positive weights w_t. This concisely unifies both the noise prediction discrepancy and the vector field discrepancy.

Combining the reward, the decomposed trajectory penalties, and the explicit distance relaxation for the optimized condition, the global objective becomes:

$$\max_{\tilde{c}_{\text{im}}} \mathbb{E}_{z_0\sim p_\theta(z_0|\tilde{c}_{\text{im}})}[r(z_0)] - \alpha \sum_{t=1}^{T} \bar{\ell}_t - \beta \mathcal{D}(\tilde{c}_{\text{im}}, c_{\text{im}}), \tag{6}$$

where $\bar{\ell}_t = \mathbb{E}_{z_t\sim p_\theta(z_t|\tilde{c}_{\text{im}})}[\ell_t]$ denotes the expected trajectory penalty at step t. In practice, optimizing across the entire trajectory is memory-intensive. Therefore, we approximate the expectation via single-sample Monte Carlo rollouts and perform per-step optimization. By estimating the clean target $\hat{z}_0^t$ (e.g., via Tweedie's formula) at each timestep, we obtain a tractable, memory-efficient objective:

$$\max_{\tilde{c}_{\text{im}}} r(\hat{z}_0^t) - \alpha \ell_t - \beta \mathcal{D}(\tilde{c}_{\text{im}}, c_{\text{im}}). \tag{7}$$

Table 1: Parameter mapping for unified trajectory decomposition.

Model	**Predictor** $\mathbf{f}_\theta$	**Weighting** $w_t = \frac{\lambda_t^2}{2\sigma_t^2}$
DDPM	Noise ϵ	$\frac{1-\alpha_t}{2\alpha_t(1-\bar{\alpha}_{t-1})}$
DDIM	Noise ϵ	$\lambda_t^2/2\sigma_t^2$
Flow Matching	Velocity v	$(\Delta t)^2/2\sigma_t^2$

Motivated by Eq. 7, at sampling step t, we obtain the refined condition $\tilde{c}_{\text{im}}$ by optimizing a minimal reward-consistency-regularization objective. The step objective is defined as:

$$\mathcal{L}_t = \mathcal{H}_t\big(\mathcal{L}_{\text{reward}}, \mathcal{L}_{\text{traj}}, \mathcal{L}_{\text{cond}}\big), \tag{8}$$

where the three terms correspond, respectively, to reward maximization on the pixel-level prediction, trajectory consistency under fixed pre-trained dynamics, and fidelity to the input condition. We regard the latter two terms as complementary measures of consistency, assessing deviations in the predicted trajectory and the conditioning signal, respectively. $\mathcal{H}_t$ denotes the two-stage scoring strategy described in Sec. 3.2.2.

3.2.2 Gradient-Free Optimizer

Backpropagating through the full diffusion trajectory and IQA reward models is computationally prohibitive. For efficient, training-free inference optimization, we introduce a zero-order (gradient-free) refinement. At each key step t, our goal is to minimize the local objective $\mathcal{L}_t$ by selecting candidates that maintain or improve both the perceptual reward and the consistency.

We estimate the surrogate gradient via an antithetic Evolution Strategy (ES) [52]. We first generate $2N$ candidates $\tilde{c}_{\text{im}}^{(i)} = \tilde{c}_{\text{im}} + \Delta_i$ using symmetric Gaussian perturbations $\Delta = \{\pm\sigma\delta_1, \ldots, \pm\sigma\delta_N\}$, where σ is the exploration scale. To promote step-by-step improvement, we first evaluate the candidates and keep any that result in a higher prioritized score S (the negative of the corresponding loss) than the baseline condition, *i.e.*, $S(\tilde{c}_{\text{im}}^{(i)}) > S(c_{\text{im}})$. From the remaining valid candidates, we select the top-K elites [53] to constitute the subset $\mathcal{M}_K$ with the best complementary scores. Reward and consistency can respectively serve as the primary scoring criterion, with the other used as the complementary criterion. The descent direction $\hat{g}_t$ is then computed as a temperature-scaled, softmax-weighted sum of these elites:

$$\hat{g}_t = -\sum_{i \in \mathcal{M}_K} \omega_i \Delta_i, \quad \text{where} \quad \omega_i \propto \exp\big(S(\tilde{c}_{\text{im}}^{(i)})/\tau\big), \tag{9}$$

where the temperature τ smooths the weights, mitigating variance from noisy objectives and diffusion stochasticity. We directly update the condition as: $\tilde{c}_{\text{im}} \leftarrow \tilde{c}_{\text{im}} - \eta\hat{g}_t$, where η is the learning rate.

3.2.3 Compact Search Strategy

Although the proposed zero-order refinement avoids gradient computation, directly searching over high-dimensional image conditions and evaluating IQA rewards at every denoising step remains costly [19, 27]. We therefore introduce a compact search strategy that combines frequency-domain condition parameterization with momentum-based update reuse.

Global DCT Band-Gain Parameterization. Instead of optimizing pixels, we transform the input LQ condition c_{im} into DCT coefficients and optimize a low-dimensional band-gain vector $\gamma \in \mathbb{R}^d$, where d is the number of gain parameters. The updated gains modulate the corresponding frequency bands, and the refined condition is reconstructed by inverse DCT:

$$\tilde{c}_{\text{im}} = \text{DCT}^{-1}[\mathcal{G}(\gamma) \odot \text{DCT}(c_{\text{im}})], \tag{10}$$

where $\mathcal{G}(\gamma)$ maps the gain parameters to a frequency-domain gain map, and $\odot$ denotes element-wise multiplication. This low-dimensional parameterization reduces the search space and constrains optimization to structured changes.

Momentum Caching and Conditional Reuse. To avoid performing IQA-guided ES at every step, we introduce a momentum-based caching mechanism, as shown in Algorithm 1. We maintain a cache state $\mathcal{S} = \{\gamma, v\}$, where γ is the current DCT band-gain parameter serving as the search anchor, and v is its direction momentum. At a scheduled *key-step* $t \in \mathcal{T}_{\text{key}}$, CATTO first obtains perturbed candidates around γ and evaluates them from the same latent state. The two-stage scoring determines the update direction $\hat{g}_t$, while the obtained candidate latents are combined with the same weights. The updated band-gain parameter γ is then used to update v. We carry the updated γ forward, allowing consecutive key-steps to start from the parameter obtained at the preceding step. At *reuse-steps*, CATTO skips ES sampling and IQA evaluation, and extrapolates γ using the cached momentum with decay coefficient μ. The reused condition is obtained by applying DCT modulation with γ to the fixed input condition c_{im}. This reduces redundant reward evaluations while preserving temporally consistent condition refinement.

Algorithm 1: Momentum-Based Conditional Reuse

Require: Input condition c_{im}, initial latent z_T, key-step set $\mathcal{T}_{\text{key}}$, learning rate η, momentum coefficient μ

$z \leftarrow z_T, \quad \gamma \leftarrow 0, \quad v \leftarrow 0$
for each sampling step t **do**
  if $t \in \mathcal{T}_{\text{key}}$ **then**
    $\{\gamma^{(i)}\} \leftarrow \text{Perturb}(\gamma)$
    $\tilde{c}_{\text{im}}^{(i)} \leftarrow \text{DCT}^{-1}\left[\mathcal{G}(\gamma^{(i)}) \odot \text{DCT}(c_{\text{im}})\right]$
    $\{z^{(i)}, s_i\} \leftarrow \text{Rollout\&Score}_t(z, \{\tilde{c}_{\text{im}}^{(i)}\})$
    $(\mathcal{S}_t, \{\omega_i\}) \leftarrow \text{TwoStageScore}(\{s_i\})$
    $\hat{g}_t \leftarrow \sum_{i \in \mathcal{S}_t} \omega_i (\gamma - \gamma^{(i)})$
    $\gamma^+ \leftarrow \gamma - \eta\hat{g}_t$
    $z \leftarrow \sum_{i \in \mathcal{S}_t} \omega_i z^{(i)}$
    $v \leftarrow \mu v + (1-\mu)(\gamma - \gamma^+), \quad \gamma \leftarrow \gamma^+$
  else
    $\gamma \leftarrow \gamma - v, \quad v \leftarrow \mu v$
    $\tilde{c}_{\text{im}} \leftarrow \text{DCT}^{-1}[\mathcal{G}(\gamma) \odot \text{DCT}(c_{\text{im}})]$
    $z \leftarrow \text{SamplerStep}_t\big(z, f_\theta(z, t, \tilde{c}_{\text{im}})\big)$
  end if
end for

4 Experiments

Implementation Details. Unless otherwise specified, CATTO and all reward-based inference-time optimization baselines use the same no-reference IQA metric for reward evaluation, with

Table 2: Real-world image super-resolution comparison with DDIM and DDPM on DRealSR [60].

Method	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	ARNIQA↑	CLIPIQA↑	MANIQA↑	NIQE↓	LIQE↑
LDM [6] (DDIM)	28.64	0.7781	0.3441	50.84	0.5766	0.5956	0.3317	7.313	2.641
BoN [17](CVPR'25)	28.75†	0.7782†	0.3430†	51.16†	0.5865†	**0.6159**†	0.3392†	7.278†	2.741†
FK [20](ICML'25)	28.80†	0.7799†	0.3409†	50.49	0.5825†	0.5904	0.3327†	7.379	2.677†
DSearch [18](arxiv'25)	28.70†	0.7826†	0.3391†	50.61	0.5882†	0.6013†	0.3429†	7.330	2.644†
KDS [21](NeurIPS'25)	28.75†	0.7787†	0.3418†	50.45	0.5817†	0.5921	0.3314	**7.272**†	2.680†
CATTO (Ours)	**28.86**†	**0.7937**†	**0.3374**†	**51.67**†	**0.6073**†	0.6043†	**0.3458**†	7.636	**2.822**†
SeeSR [10] (DDPM)	28.07	0.7684	0.3174	65.08	0.6885	0.6903	0.5128	6.381	4.127
BoN [17](CVPR'25)	28.11†	0.7675	0.3221	65.75†	0.6964†	0.7197†	0.5216†	6.624	4.162†
FK [20](ICML'25)	28.19†	0.7684	0.3203	65.04	0.6939†	0.7098†	0.5092	6.720	4.102
DSearch [18](arxiv'25)	28.08†	0.7674	0.3246	65.46†	0.6998†	0.7216†	0.5273†	6.482	4.171†
KDS [21](NeurIPS'25)	28.10†	0.7683	**0.3072**†	64.32	0.6926†	0.6427	0.4757	**6.056**†	3.947
CATTO (Ours)	**28.21**†	**0.7698**†	0.3291	**66.09**†	**0.7008**†	**0.7258**†	**0.5279**†	6.461	**4.180**†
DiffBIR [14] (DDPM)	25.38	0.6035	0.4829	67.30	0.6653	0.7102	0.5617	6.734	4.117
BoN [17](CVPR'25)	25.27	0.5937	0.4899	67.86†	0.6720†	0.7544†	0.5775†	6.725†	4.263†
FK [20](ICML'25)	25.38	0.6015	0.4883	67.57†	0.6729†	0.7376†	0.5668†	6.899	4.171†
DSearch [18](arxiv'25)	25.33	0.6129†	0.4730†	**68.32**†	**0.6761**†	0.7550†	0.5750†	9.530	**4.340**†
KDS [21](NeurIPS'25)	25.61†	0.5940	0.5087	65.92	0.6514	0.7002	0.5696†	6.466†	4.015
CATTO (Ours)	**25.62**†	**0.6130**†	**0.4713**†	67.88†	0.6730†	**0.7590**†	**0.5794**†	**6.032**†	4.220†

† denotes improvement over the corresponding backbone baseline.

Table 3: Comparison of different inference-time methods on image inpainting, photo-realistic face restoration, and low-light enhancement. LDM [6] is used with DDIM sampling, while FlowIE [57] and FLUX-IR [58] are based on rectified-flow formulations.

Method	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	ARNIQA↑	CLIPIQA↑	MANIQA↑	NIQE↓	LIQE↑
LDM [6] / Inpainting	21.70	0.8419	0.1342	66.17	0.6560	0.5771	0.4202	**3.317**	**3.836**
BoN [17](CVPR'25)	21.75†	0.8424†	0.1337†	66.20†	0.6561†	**0.5882**†	0.4221†	3.326	3.834
FK [20](ICML'25)	21.74†	0.8425†	0.1338†	66.17	0.6552	0.5763	0.4220†	3.325	3.832
DSearch [18](arxiv'25)	21.74†	0.8426†	0.1340†	66.20†	0.6554	0.5823†	0.4223†	3.336	**3.836**
KDS [21](NeurIPS'25)	**21.79**†	**0.8430**†	0.1337†	66.15	0.6559	0.5748	**0.4225**†	3.334	3.827
CATTO (Ours)	21.74†	0.8419	**0.1329**†	**66.41**†	**0.6563**†	0.5852†	**0.4225**†	3.328	**3.836**
FlowIE [57] / Face Restoration	25.63	0.6863	0.2238	71.01	0.6773	0.6716	0.5738	5.338	4.664
BoN [17](CVPR'25)	25.72†	0.6864†	**0.2169**†	71.44†	0.6866†	0.6777†	0.5598	**4.693**†	4.673†
FK [20](ICML'25)	25.77†	0.6862	0.2193†	71.58†	0.6826†	0.6905†	0.5648	4.932†	4.694†
DSearch [18](arxiv'25)	25.96†	0.6945†	0.2212†	71.09†	0.6790†	0.6851†	0.5797†	5.326†	4.691†
KDS [21](NeurIPS'25)	25.99†	0.6948†	0.2217†	71.21†	0.6788†	0.6733†	0.5735	5.275†	4.656
CATTO (Ours)	**26.19**†	**0.6978**†	0.2281	**71.95**†	**0.6888**†	**0.7102**†	**0.5877**†	5.463	**4.719**†
FLUX-IR [58] / Low-light	21.97	0.8251	0.1142	73.46	0.7204	0.5131	0.4770	4.046	4.631
BoN [17](CVPR'25)	21.55	0.8197	0.1163	73.57†	0.7227†	**0.5608**†	0.4861†	3.932†	4.623
FK [20](ICML'25)	21.70	0.8237	0.1149	73.16	0.7119	0.4939	0.4651	4.026†	4.455
DSearch [18](arxiv'25)	22.15†	0.8226	0.1089†	73.73†	0.7219†	0.5435†	0.4820†	**3.874**†	4.645†
KDS [21](NeurIPS'25)	21.70	0.8238	0.1150	73.12	0.7108	0.4893	0.4644	4.042†	4.440
CATTO (Ours)	**22.60**†	**0.8253**†	**0.1085**†	**73.77**†	**0.7232**†	0.5566†	**0.4874**†	4.033†	**4.663**†

† denotes improvement over the corresponding backbone baseline.

CLIPIQA [54] as the default. When integrating these methods, we retain the default configurations of the underlying restoration backbones.

CATTO employs a two-stage scoring strategy for candidate selection. Motivated by the coarse-to-fine nature of diffusion denoising [55, 56], we prioritize consistency at early optimization key steps: candidates with worse consistency than the baseline condition are discarded, and the remaining candidates are ranked by IQA reward. At later key steps, we prioritize reward by reversing these roles, filtering candidates by IQA reward relative to the baseline condition and ranking the remaining candidates by consistency score. This schedule emphasizes structural preservation early in denoising and perceptual refinement at later stages.

Backbones and Compared Methods. We consider LDM [6], DiffBIR [14], SeeSR [10], FlowIE [57] and FLUX-IR [58] as restoration backbones. The compared inference-time optimization methods include Best-of-N (BoN) [17], DSearch [18], Feynman–Kac steering (FK) [20], and Kernel Density Steering (KDS) [21]. To ensure fair comparisons within each backbone, we follow prior work [17, 59] in matching inference-time compute across methods, setting the budget to approximately 200 network function evaluations (NFEs) per method. Each NFE corresponds to one network forward pass.



Figure 3: Visual comparison of different inference-time methods across different baselines.

Benchmarks and metrics. For the evaluation of real-world image super-resolution, we use the pre-cropped test image pairs provided by StableSR [9], with 128×128 LR images from RealSR [61] and DRealSR [60], following the evaluation protocol of SeeSR [10]. For other restoration tasks, we use the first 100 images of CelebA-Test [62] for face restoration, LOL [63] for low-light enhancement, and the first 1,000 images of Places365 [64] for image inpainting. We evaluate restoration quality using full-reference metrics, including PSNR, SSIM [65], and LPIPS [66], and no-reference metrics, including MUSIQ [67], CLIPIQA [54], MANIQA [68], ARNIQA [69], NIQE [70], and LIQE [71].

4.1 Evaluation on Image Restoration

Real-World Image Super-Resolution. Tab. 2 shows that, on DRealSR [60], CATTO improves nearly all metrics over the corresponding backbone baselines. Crucially, the gains are not limited to reward-related perceptual metrics (CLIPIQA), but generally extend to other no-reference quality metrics (such as MANIQA, MUSIQ, and LIQE) and full-reference metrics (PSNR and SSIM), indicating improved visual quality without compromising content fidelity. In contrast, other reward-aligned sampling-process optimization methods (BoN, DSearch, FK) mainly enhance perceptual quality, while KDS primarily improves fidelity-oriented metrics. Tab. 8 in the Appendix further shows consistent results on RealSR [61].

Visual comparisons in Fig. 3 confirm these gains. CATTO effectively suppresses high-frequency artifacts in SeeSR [10] and corrects details in DiffBIR [14], ensuring better alignment with the

ground truth. These results validate that our consistency constraints successfully prevent over-optimization while refining local details. Appendix Fig. 5 further corroborates this through spatial- and frequency-domain analyses.

Broader IR Tasks. We further validate CATTO on a broader range of image restoration tasks (face restoration [62], inpainting [64], and low-light enhancement [63]). As shown in Tab. 3, integrating CATTO generally improves the reference metrics PSNR and SSIM across the evaluated tasks, and it also yields consistent gains on most no-reference perceptual metrics such as MUSIQ and ARNIQA. Qualitatively, as demonstrated in Fig. 3, CATTO recovers faithful details while suppressing hallucinated textures, producing outputs that look more natural and better match the reference. For example, when built on the FlowIE [57] for face restoration, CATTO produces more natural and faithful facial components with richer face textures.

Runtime & Computational Complexity. As shown in Tab. 7, CATTO offers a competitive balance between inference speed and memory usage. It requires 12.73 s per image, compared with 17.26 s for DSearch [18] and 16.72 s for FK [20], while maintaining a comparable memory footprint of 7.33 GB. CATTO concentrates candidate evaluation at scheduled key steps and reuses cached updates between them, limiting repeated search and IQA evaluation. Its gradient-free formulation also avoids storing activations for backpropagation through the denoising trajectory and reward models. Together, these design choices support efficient condition optimization with manageable memory requirements.

4.2 Ablation Study

We conduct a component-wise ablation study and a hyperparameter ablation study on the DRealSR dataset to validate the effectiveness of CATTO. The results are reported in Tabs. 4, 5, 6 and Fig. 4.

Impact of optimization objectives. We examine the contribution of each component in the composite objective (Eq. 8). After candidates are filtered against the reference condition, the objective guides elite selection. As shown in Tab. 4, selecting solely by $\mathcal{L}_{\text{reward}}$ (row a) prioritizes perceptual quality but provides weaker reconstruction fidelity than the full configuration. Adding trajectory consistency $\mathcal{L}_{\text{traj}}$ (row b) or condition fidelity $\mathcal{L}_{\text{cond}}$ (row c) improves PSNR while maintaining strong perceptual performance. These terms provide complementary guidance: trajectory consistency limits deviations from the pretrained dynamics, while condition fidelity anchors refinement to the input observation. Combining all three terms (row d) achieves the best PSNR and NIQE while retaining competitive perceptual scores, supporting the effectiveness of their joint regularization in balancing perceptual quality and reconstruction fidelity.

Impact of DCT and reuse. As shown in Tab. 5, we evaluate the effectiveness of DCT-subspace optimization and momentum caching & reuse (C.R.). Row (a) is the pixel-space baseline without reuse. Enabling C.R. alone in row (b) achieves the best PSNR and CLIPIQA, indicating that update directions found at key steps can remain useful during subsequent denoising steps. Enabling DCT alone in row (c) also improves PSNR and CLIPIQA over the baseline, supporting the use of structured band-gain perturbations for condition search. The two modules also operate in a shared parameter space: DCT defines the coordinates in which candidates are explored, and C.R. carries update directions forward in those same coordinates. With both enabled, our final setting in row (d) achieves the best NIQE and MUSIQ and near-best PSNR. The isolated settings demonstrate the contribution of each module, while their combination provides strong restoration quality with fewer repeated candidate evaluations between key steps.

Impact of IQA metric rewards. Tab. 6 examines how the reward choice affects CATTO on DRealSR. Using CLIPIQA yields the highest PSNR and CLIPIQA scores while maintaining competitive results on the other reported metrics. MUSIQ as the reward produces the highest MUSIQ and LIQE scores, but its PSNR and CLIPIQA scores are lower. ARNIQA provides an intermediate trade-off, whereas NIQE alone gives lower PSNR, CLIPIQA, and LIQE scores than CLIPIQA-based optimization. These results show that the reward choice changes the balance among quality metrics; no single choice leads on every measure. We therefore use CLIPIQA alone as the reward in the main experiments. Additional comparisons across inference-time scaling methods are provided in Appendix Tab. 10.

Impact of Hyperparameters. We analyze the effects of the number of perturbation size N and the elite set size K in Fig. 4. Increasing perturbation size usually improves both PSNR and CLIPIQA, as exploring more perturbation directions increases the likelihood of finding better candidates. Increasing K improves PSNR but reduces CLIPIQA. Aggregating more candidates favors agreement among

Table 4: Quantitative results of different settings for optimization objectives.

Methods			DRealSR				
R	T	C	PSNR↑	NIQE↓	MUSIQ↑	CLIPIQA↑	LIQE↑
SeeSR [10]			28.07	**6.381**	65.08	0.6903	4.127
(a) ✓			28.11	6.485	66.05	0.7254	4.195
(b) ✓	✓		28.15	6.545	66.28	**0.7289**	4.190
(c) ✓		✓	28.13	6.584	**66.39**	0.7285	**4.212**
(d) ✓	✓	✓	**28.21**	6.461	66.09	0.7258	4.180

Table 5: Quantitative results of different settings for DCT and caching & reuse (C.R.).

Methods			DRealSR				
DCT	C.R.		PSNR↑	NIQE↓	MUSIQ↑	CLIPIQA↑	LIQE↑
SeeSR [10]			28.07	**6.381**	65.08	0.6903	4.127
(a)			28.10	6.862	66.02	0.7261	**4.202**
(b)		✓	**28.22**	6.736	66.00	**0.7264**	4.174
(c)	✓		28.16	6.795	65.91	0.7256	4.187
(d)	✓	✓	28.21	6.461	**66.09**	0.7258	4.180

Table 6: Quantitative results of using different IQA rewards.

Methods	DRealSR					
Rewards	PSNR↑	NIQE↓	MUSIQ↑	CLIPIQA↑	ARNIQA↑	LIQE↑
SeeSR [10]	28.07	6.381	65.08	0.6903	0.6885	4.127
CLIPIQA	**28.21**	6.461	66.09	**0.7258**	0.7008	4.180
MUSIQ	28.11	6.547	**67.60**	0.7038	0.7020	**4.239**
ARNIQA	28.16	6.399	66.42	0.7054	**0.7157**	4.189
NIQE	28.14	**5.944**	66.12	0.6920	0.6964	4.129

Table 7: Comparison of time and memory under different methods.

Methods	**Time** (s)	**Memory** (GB)
SeeSR [10]	4.13	6.86
BoN [17]	9.30	7.55
KDS [21]	8.76	6.86
FK [20]	16.72	7.55
DSearch [18]	17.26	7.25
CATTO (Ours)	12.73	7.33

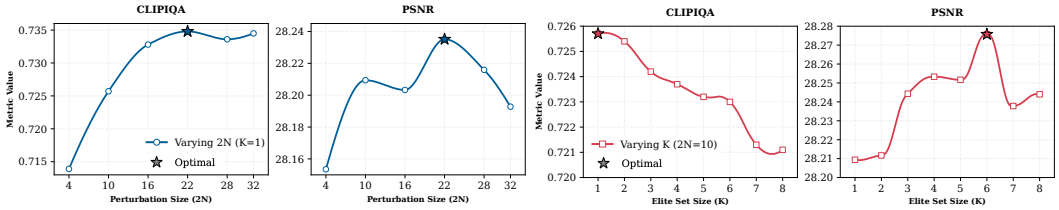

Figure 4: Impact of perturbation size N and elite set size K.

their predictions, yielding a more conservative update that improves fidelity but can weaken the perceptual gains of the highest-reward candidates. Considering the limited gains at larger budgets and the balance between perceptual quality, reconstruction fidelity, and efficiency, we adopt perturbation size $2N = 10$ and $K = 1$ as the default configuration.

5 Conclusion and Limitations

In this paper, we present Condition-Aware Test-Time Optimization (CATTO), a training-free and plug-and-play framework that redefines the interaction between degraded observations and diffusion priors. Moving beyond the conventional single-shot conditioning paradigm, CATTO iteratively refines the image condition during the reverse sampling process to navigate the fidelity-perception trade-off. By leveraging a gradient-free optimizer and a compact search strategy in the frequency subspace, CATTO bypasses the need for costly backpropagation and complex reward differentiation, ensuring reward optimization efficiency while maintaining structural consistency. Extensive evaluations demonstrate that CATTO is a plug-and-play solution that enhances perceptual quality across various baseline DPMs with manageable computational overhead.

Like other inference-time optimization methods, CATTO trades additional inference-time computation for improved restoration quality, inevitably introducing extra sampling-time overhead. Its performance also partially depends on the alignment between black-box rewards and restoration quality. Future work will explore more advanced task-adaptive rewards and combine latent-space reward proxies to improve both performance and efficiency.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (No. U24B20175, No. 62322216, No. 62625602), Scientific Research Innovation Capability Support Project for Young Faculty (No. SRICSPYF-ZY2025012), Guangdong S&T Program (No. 2026B0101090001), Shenzhen Science and Technology Program (No. KQTD20221101093559018, SYSRD20250529113401002), and the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (No. GML-KF-26-02).

References

[1] Jingwen Su, Boyan Xu, and Hujun Yin. A survey of deep learning approaches to image restoration. *Neurocomputing*, 487:46–65, 2022.

[2] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement: A comprehensive survey. *International Journal of Computer Vision*, 133(11):8078–8108, 2025.

[3] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

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

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

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

[8] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

[9] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.

[10] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 25456–25467, 2024.

[11] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *IEEE International Conference on Computer Vision*, pages 13095–13105, 2023.

[12] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European conference on computer vision*, pages 196–212. Springer, 2024.

[13] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

[14] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, pages 430–448. Springer, 2024.

[15] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv:2409.17058*, 2024.

[16] Guy Ohayon, Tomer Michaeli, and Michael Elad. Posterior-mean rectified flow: Towards minimum MSE photo-realistic image restoration. In *International Conference on Learning Representations*, 2025.

[17] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2523–2534, 2025.

[18] Xiner Li, Masatoshi Uehara, Xingyu Su, Gabriele Scalia, Tommaso Biancalani, Aviv Regev, Sergey Levine, and Shuiwang Ji. Dynamic search for inference-time alignment in diffusion models. *arXiv preprint arXiv:2503.02039*, 2025.

[19] Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. *International Conference on Machine Learning*, 2025.

[20] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *International Conference on Machine Learning*, 2025.

[21] Yuyang Hu, Kangfu Mei, Mojtaba Sahraee-Ardakan, Ulugbek S Kamilov, Peyman Milanfar, and Mauricio Delbracio. Kernel density steering: Inference-time scaling via mode seeking for image restoration. *Advances in Neural Information Processing Systems*, 2025.

[22] Mahdi Farahbakhsh, Vishnu Teja Kunde, Dileep Kalathil, Krishna Narayanan, and Jean-Francois Chamberland. Inference-time search using side information for diffusion-based image reconstruction. *arXiv preprint arXiv:2510.03352*, 2025.

[23] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[24] Vignav Ramesh and Morteza Mardani. Test-time scaling of diffusion models via noise trajectory search. *arXiv preprint arXiv:2506.03164*, 2025.

[25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[26] Taehoon Kim, Henry Gouk, and Timothy Hospedales. Test-time alignment of text-to-image diffusion models via null-text embedding optimisation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 43611–43620, 2026.

[27] Xiefan Guo, Miaomiao Cui, Liefeng Bo, and Di Huang. Shortft: Diffusion model alignment via shortcut-based fine-tuning. In *IEEE International Conference on Computer Vision*, pages 678–687, 2025.

[28] Nasir Ahmed, T Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 100(1):90–93, 1974.

[29] Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

[30] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

[31] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

[32] Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

[33] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[34] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[35] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[36] Xinran Qin, Yuhui Quan, Zhuojie Chen, and Hui Ji. Robust unsupervised deep learning for nonblind image deconvolution with inaccurate kernels. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):16960–16974, 2025.

[37] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023.

[38] Yuhong Zhang, Hengsheng Zhang, Xinning Chai, Zhengxue Cheng, Rong Xie, Li Song, and Wenjun Zhang. Diff-restorer: Unleashing visual prompts for diffusion-based universal image restoration. *arXiv preprint arXiv:2407.03636*, 2024.

[39] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for multi-task image restoration. *arXiv preprint arXiv:2310.01018*, 2023.

[40] Bingchen Li, Xin Li, Yiting Lu, Ruoyu Feng, Mengxi Guo, Shijie Zhao, Li Zhang, and Zhibo Chen. Promptcir: Blind compressed image restoration with prompt learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6442–6452, 2024.

[41] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023.

[42] Badr Moufad, Yazid Janati, Lisa Bedin, Alain Oliviero Durmus, Randal Douc, Eric Moulines, and Jimmy Olsson. Variational diffusion posterior sampling with midpoint guidance. In *International Conference on Learning Representations*, 2025.

[43] Sojin Lee, Dogyun Park, Inho Kong, and Hyunwoo J Kim. Diffusion prior-based amortized variational inference for noisy inverse problems. In *European Conference on Computer Vision*, pages 288–304. Springer, 2024.

[44] Jaihoon Kim, Taehoon Yoon, Jisung Hwang, and Minhyuk Sung. Inference-time scaling for flow models via stochastic generation and rollover budget forcing. *arXiv preprint arXiv:2503.19385*, 2025.

[45] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision*, pages 3836–3847, 2023.

[46] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.

[47] Jiangang Wang, Qingnan Fan, Qi Zhang, Haigen Liu, Yuhang Yu, Jinwei Chen, and Wenqi Ren. Hero-sr: One-step diffusion for super-resolution with human perception priors. *arXiv:2412.07152*, 2024.

[48] Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimisation. In *International Conference on Learning Representations*, 2025.

[49] Mihir Prabhudesai, Tsung-Wei Ke, Alex Li, Deepak Pathak, and Katerina Fragkiadaki. Diffusion-tta: Test-time adaptation of discriminative models via generative feedback. *Advances in Neural Information Processing Systems*, 36:17567–17583, 2023.

[50] Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

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

[52] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[53] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

[54] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563, 2023.

[55] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8038–8047. IEEE, 2024.

[56] Bingchen Li, Xin Li, Jiaqi Xu, Jiaming Guo, Wenbo Li, Renjing Pei, and Zhibo Chen. Test-time preference optimization for image restoration. In *AAAI Conference on Artificial Intelligence*, volume 40, pages 5973–5981, 2026.

[57] Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, and Jiwen Lu. Flowie: Efficient image enhancement via rectified flow. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13–22. IEEE, 2024.

[58] Zhiyu Zhu, Jinhui Hou, Hui Liu, Huanqiang Zeng, and Junhui Hou. Learning efficient and effective trajectories for differential equation-based image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[59] Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Y Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. In *International Conference on Learning Representations*, volume 2026, pages 39133–39166, 2026.

[60] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European Conference on Computer Vision*, pages 101–117. Springer, 2020.

[61] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *IEEE International Conference on Computer Vision*, pages 3086–3095, 2019.

[62] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[63] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[64] 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, 2017.

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

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[67] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *IEEE International Conference on Computer Vision*, 2021.

[68] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022.

[69] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *IEEE Winter Conference on Applications of Computer Vision*, pages 189–198, 2024.

[70] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[71] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023.

Appendix

This supplementary material provides further details for the main paper. It covers the following aspects:

- **Section A: Experiment Details**: Introduction of hyperparameter settings.
- **Section B: Method Details**: Covering the derivation of reward-aligned condition refinement and details of the optimization objective.
- **Section C: More Experimental Results**: Presenting more visualization results, including analytic plots and qualitative examples for image super-resolution, face restoration, low-light enhancement, and inpainting to further demonstrate the efficacy of CATTO.

A Experimental Details

A.1 Hyperparameter Setting

Unless otherwise specified, all compared methods use a common inference budget of approximately 200 NFEs. The number of candidates or particles is adjusted according to each method's sampling procedure to match this budget. For CATTO, we set the exploration scale to $\sigma = 0.005$ for early steps and $\sigma = 0.02$ for late steps, and retain $K = 1$ candidate for the update. The zero-order update learning rate is $\eta = 1.0$.

B Method Details

B.1 Derivation of Sec. 3.2.1

Since direct sampling from the optimal target is intractable, we approximate it by optimizing the condition $\tilde{c}_{\text{im}}$. The KL regularizer effectively constrains the generative trajectory of the optimized condition to remain close to the original. Using the joint distributions $Q_{\tilde{c}}(z_{0:T}, c) = q_{\tilde{c}}(c)p_\theta(z_{0:T} \mid c)$ and $P(z_{0:T}, c) = p_c(c)p_\theta(z_{0:T} \mid c_{\text{im}})$ defined in the main text, we write:

$$\mathcal{L}_{\text{reg}} = D_{\text{KL}}\Big(Q_{\tilde{c}}(z_{0:T}, c) \,\Big\|\, P(z_{0:T}, c)\Big). \tag{11}$$

By applying the chain rule of KL divergence and leveraging the Markov property of diffusion models,

$$p_\theta(z_{0:T} \mid c) = p(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1} \mid z_t, c), \tag{12}$$

we can decompose this joint objective. Note that, for models with a shared condition-independent terminal distribution, the term corresponding to $p(z_T)$ cancels out. The decomposition proceeds as follows:

$$\begin{aligned}
\mathcal{L}_{\text{reg}} &= \mathbb{E}_{(z_{0:T},c)\sim Q_{\tilde{c}}}\left[\log \frac{q_{\tilde{c}}(c)\prod_{t=1}^{T} p_\theta(z_{t-1} \mid z_t, c)}{p_c(c)\prod_{t=1}^{T} p_\theta(z_{t-1} \mid z_t, c_{\text{im}})}\right] \\
&= \mathbb{E}_{c\sim q_{\tilde{c}}}\left[\log \frac{q_{\tilde{c}}(c)}{p_c(c)}\right] + \sum_{t=1}^{T} \mathbb{E}_{c\sim q_{\tilde{c}}}\mathbb{E}_{z_t\sim p_\theta(z_t|c)}\left[\mathbb{E}_{z_{t-1}\sim p_\theta(\cdot|z_t,c)}\left[\log \frac{p_\theta(z_{t-1} \mid z_t, c)}{p_\theta(z_{t-1} \mid z_t, c_{\text{im}})}\right]\right] \\
&= D_{\text{KL}}(q_{\tilde{c}} \| p_c) + \sum_{t=1}^{T} \mathbb{E}_{c\sim q_{\tilde{c}}}\mathbb{E}_{z_t\sim p_\theta(z_t|c)}\left[D_{\text{KL}}(p_\theta(z_{t-1} \mid z_t, c) \,\|\, p_\theta(z_{t-1} \mid z_t, c_{\text{im}}))\right],
\end{aligned} \tag{13}$$

where the first term acts as a fidelity constraint on the condition itself. By the data processing inequality, this joint KL upper-bounds the KL between the induced z_0 distributions. As in the main text, we approximate the expectation over c in the trajectory term by evaluating it at $\tilde{c}_{\text{im}}$, and use $\mathcal{D}(\tilde{c}_{\text{im}}, c_{\text{im}})$ as a surrogate for the condition-level KL.

To quantify the discrepancy between the optimized trajectory (under condition $\tilde{c}_{\text{im}}$) and the original trajectory (under c_{im}), we model the reverse transitions as Gaussian distributions with a shared

condition-independent variance $\sigma_t^2 > 0$. The mean $\mu_t(c)$ is defined via a unified predictor $\mathbf{f}_\theta$:

$$p_\theta(z_{t-1} \mid z_t, c) = \mathcal{N}\big(z_{t-1}; \mu_t(c), \sigma_t^2 \mathbf{I}\big), \quad \mu_t(c) = b_t(z_t) + \lambda_t \mathbf{f}_\theta(z_t, t, c), \tag{14}$$

where $b_t(z_t)$ does not depend on c.

For two Gaussians with identical covariance matrices $\Sigma = \sigma_t^2 \mathbf{I}$, the KL divergence simplifies to the scaled squared Euclidean distance between their means. At the same state z_t, the shared term $b_t(z_t)$ cancels, giving

$$\mu_t(\tilde{c}_{\text{im}}) - \mu_t(c_{\text{im}}) = \lambda_t \big[\mathbf{f}_\theta(z_t, t, \tilde{c}_{\text{im}}) - \mathbf{f}_\theta(z_t, t, c_{\text{im}})\big]. \tag{15}$$

The step-wise KL term therefore reduces to a weighted L_2 distance:

$$\begin{aligned} D_{\text{KL}}^t &= \frac{1}{2\sigma_t^2} \big\| \lambda_t \big(\mathbf{f}_\theta(z_t, t, \tilde{c}_{\text{im}}) - \mathbf{f}_\theta(z_t, t, c_{\text{im}})\big) \big\|_2^2 \\ &= \frac{\lambda_t^2}{2\sigma_t^2} \big\| \mathbf{f}_\theta(z_t, t, \tilde{c}_{\text{im}}) - \mathbf{f}_\theta(z_t, t, c_{\text{im}}) \big\|_2^2 = \ell_t. \end{aligned} \tag{16}$$

For deterministic updates with $\sigma_t = 0$, we use the same predictor distance with finite positive weights as a regularizer rather than an exact KL. In practice, we evaluate predictor discrepancies along multi-step candidate rollouts as a surrogate for the trajectory KL.

Case I: Noise-Predicting Frameworks

In the DDPM framework, the reverse transition is derived from the posterior distribution. We map the unified predictor $\mathbf{f}_\theta$ to the noise predictor ϵ_θ through the following derivation:

Step A: Mean Difference via $\hat{z}_0$. The posterior mean $\tilde{\mu}_t$ is a linear combination of the current state z_t and the predicted clean latent $\hat{z}_0$. When calculating the difference between two conditions at the same z_t, the z_t terms cancel:

$$\mu_t(\tilde{c}_{\text{im}}) - \mu_t(c_{\text{im}}) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \left(\hat{z}_0(z_t, \tilde{c}_{\text{im}}) - \hat{z}_0(z_t, c_{\text{im}})\right). \tag{17}$$

Step B: Substitution of Noise Predictor. Using the standard unclipped prediction $\hat{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(z_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta)$, we substitute $\hat{z}_0$ to express the mean difference in terms of the noise predictor ϵ_θ:

$$\mu_t(\tilde{c}_{\text{im}}) - \mu_t(c_{\text{im}}) = -\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t\sqrt{1 - \bar{\alpha}_t}}{(1 - \bar{\alpha}_t)\sqrt{\bar{\alpha}_t}} \left(\epsilon_\theta(z_t, t, \tilde{c}_{\text{im}}) - \epsilon_\theta(z_t, t, c_{\text{im}})\right). \tag{18}$$

Step C: Coefficient Simplification. For transitions with posterior variance $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t > 0$, and using $\bar{\alpha}_t = \alpha_t \bar{\alpha}_{t-1}$ and $\beta_t = 1 - \alpha_t$, the final weight is simplified:

$$\begin{aligned} D_{\text{KL}}^t &= \frac{1}{2\sigma_t^2} \frac{\bar{\alpha}_{t-1}\beta_t^2}{(1 - \bar{\alpha}_t)\bar{\alpha}_t} \left\| \epsilon_\theta(z_t, t, \tilde{c}_{\text{im}}) - \epsilon_\theta(z_t, t, c_{\text{im}}) \right\|_2^2 \\ &= \frac{1 - \alpha_t}{2\alpha_t(1 - \bar{\alpha}_{t-1})} \left\| \epsilon_\theta(z_t, t, \tilde{c}_{\text{im}}) - \epsilon_\theta(z_t, t, c_{\text{im}}) \right\|_2^2 = \ell_t. \end{aligned} \tag{19}$$

Case II: Flow Matching

Flow Matching models learn a probability path via a velocity field $v_\theta(z_t, t, c)$.

Step A: Predictor Mapping. For an Euler step, the unified predictor $\mathbf{f}_\theta$ corresponds directly to the velocity field v_θ, and the update step λ_t aligns with the signed time increment Δt:

$$\mathbf{f}_\theta(z_t, t, c) = v_\theta(z_t, t, c), \quad \lambda_t = \Delta t. \tag{20}$$

Step B: Deterministic Updates. In the ODE-based formulation, the transition is deterministic ($\sigma_t = 0$), so the Gaussian KL reduction does not apply. Instead, we use the weighted L_2 distance between velocity fields as a fidelity regularizer:

$$\begin{aligned} \ell_t &= w_t \left\| \mathbf{f}_\theta(z_t, t, \tilde{c}_{\text{im}}) - \mathbf{f}_\theta(z_t, t, c_{\text{im}}) \right\|_2^2 \\ &= w_t \left\| v_\theta(z_t, t, \tilde{c}_{\text{im}}) - v_\theta(z_t, t, c_{\text{im}}) \right\|_2^2, \end{aligned} \tag{21}$$

where w_t is a finite positive weight.

B.2 Per-Step Optimization Objective: Details

B.2.1 Reward term and decoded evaluation

Since optimization is performed at intermediate noisy states z_t, we form a clean estimate $\hat{z}_0^t$ (using the denoiser/Tweedie-style estimator) and evaluate perceptual metrics on the decoded image:

$$\mathcal{L}_{\text{reward}} = -\sum_k w_k \, \mathcal{R}_k\big(\mathcal{D}(\hat{z}_0^t)\big), \tag{22}$$

where $\mathcal{R}_k$ can be black-box or differentiable metrics (e.g., MUSIQ, CLIPIQA), and $w_k \geq 0$.

B.2.2 Trajectory Consistency as a Trust Region

Maximizing perceptual rewards alone may over-steer $\tilde{c}_{\text{im}}$ toward implausible updates. We therefore use consistency as a trust region: reward improvements are only pursued among candidates whose induced predictions stay close to those of the original condition. Following the trajectory-level KL in Sec. 3.2.1, the trajectory consistency of candidate i is

$$\mathcal{L}_{\text{traj}}^{(i)} = \big\| \mathbf{f}_\theta\big(z_t, t, \tilde{c}_{\text{im}}^{(i)}\big) - \mathbf{f}_\theta\big(z_t, t, c_{\text{im}}\big) \big\|_2^2, \tag{23}$$

where $\mathbf{f}_\theta$ is the noise or velocity predictor. The weight w_t in Eq. 5 is omitted, since it is shared by all candidates at a given step and does not affect their ranking.

Predictor-space agreement does not directly measure deviations in the decoded image on which the reward is computed. Since the clean-image estimate is derived from the predicted noise or velocity, it provides a complementary image-space view of the denoising dynamics. We therefore add an image-space regularizer:

$$\mathcal{L}_{\text{img}}^{(i)} = -\,\text{PSNR}\big(\hat{x}_0^{t,(i)}, \hat{x}_0^{\text{ref}}\big), \tag{24}$$

where $\hat{x}_0^{t,(i)}$ is the decoded clean-image estimate of candidate i, and $\hat{x}_0^{\text{ref}}$ is the reference estimate obtained before the current optimization step. Higher PSNR indicates stronger agreement and hence a smaller penalty.

Since $\mathcal{L}_{\text{traj}}$, $\mathcal{L}_{\text{img}}$, and $\mathcal{L}_{\text{cond}}$ differ in scale, we convert each into a percentile rank among the candidates at the current key step and combine them into the joint consistency score:

$$\mathcal{L}_{\text{cons}}^{(i)} = \beta_{\text{traj}} \mathcal{R}\big(\mathcal{L}_{\text{traj}}^{(i)}\big) + \beta_{\text{img}} \mathcal{R}\big(\mathcal{L}_{\text{img}}^{(i)}\big) + \beta_{\text{cond}} \mathcal{R}\big(\mathcal{L}_{\text{cond}}^{(i)}\big), \tag{25}$$

where $\mathcal{R}(\cdot) \in [0, 1]$ denotes percentile-rank normalization in ascending order, so lower values indicate better consistency. β_{traj}, β_{img}, β_{cond} are weights for balancing consistency items. This score serves as the consistency criterion in the two-stage scoring of Sec. 3.2.2.

B.2.3 Condition Fidelity as Likelihood Regularization

To specify the relaxation term $\mathcal{D}(\tilde{c}_{\text{im}}, c_{\text{im}})$ in Eq. 6, we choose an empirical L_2 fidelity penalty:

$$\mathcal{D}(\tilde{c}_{\text{im}}, c_{\text{im}}) = \frac{1}{n} \big\| \tilde{c}_{\text{im}} - c_{\text{im}} \big\|_2^2, \tag{26}$$

In the context of image restoration, this penalty discourages excessive deviations from the original observation while allowing moderate condition refinement. Thus, the refined condition $\tilde{c}_{\text{im}}$ remains close to the original condition while allowing precise, localized adjustments to eliminate artifacts.

C More Experimental Results

C.1 More Results on Image Super-resolution

The main manuscript presents our image super-resolution comparison on DRealSR. We provide complementary results on the RealSR dataset [61] in Tab. 8. The results follow a trend similar to that observed on DRealSR: our method performs competitively across the evaluated metrics among the best overall, further demonstrating its effectiveness on real-world super-resolution data.

Table 8: Real-world image super-resolution comparisons using DDIM and DDPM on RealSR [61], and photorealistic face restoration comparisons on CelebA-Test [62].

Method	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	ARNIQA↑	CLIPIQA↑	MANIQA↑	NIQE↓	LIQE↑
LDM (DDIM)	26.20	0.7321	0.3155	56.02	0.6254	0.5735	0.3345	6.631	2.608
BoN [17](CVPR'25)	26.20	0.7324†	0.3164	56.34†	0.6241	**0.5977**†	0.3406†	6.528†	**2.673**†
FK [20](ICML'25)	26.17	0.7308	0.3143†	55.57	0.6223	0.5698	0.3357†	6.517†	2.594
DSearch [18](arxiv'25)	26.25†	**0.7361**†	0.3095†	55.09	0.6286†	0.5587	0.3330	6.604†	2.586
KDS [21](NeurIPS'25)	26.21†	0.7324†	0.3152†	56.08†	0.6227	0.5712	0.3343	**6.410**†	2.636†
CATTO (Ours)	**26.27**†	0.7344†	**0.3078**†	**56.41**†	**0.6302**†	0.5950†	**0.3407**†	6.448†	2.652†
SeeSR (DDPM)	25.41	0.7270	0.2935	68.84	0.7147	0.6513	0.5238	5.442	3.992
BoN [17](CVPR'25)	25.35	0.7317†	0.2994	69.83†	0.7171†	0.6982†	0.5464†	5.476	4.122†
FK [20](ICML'25)	25.37	0.7313†	0.2972	69.21†	0.7154†	0.6754†	0.5310†	5.433†	4.041†
DSearch [18](arxiv'25)	25.43†	**0.7320**†	0.3030	69.24†	0.7150†	0.6768†	0.5507†	5.813	4.061†
KDS [21](NeurIPS'25)	25.22	0.7270	**0.2795**†	67.96	0.7095	0.6353	0.4901	**5.253**†	3.863
CATTO (Ours)	**25.46**†	0.7283†	0.3005	**69.97**†	**0.7172**†	**0.7237**†	**0.5552**†	5.431†	**4.158**†
DiffBIR (DDPM)	24.47	0.6353	0.3815	68.49	0.6755	0.7091	0.5540	5.799	3.981
BoN [17](CVPR'25)	24.52†	0.6377†	0.3817	69.17†	0.6840†	0.7397†	0.5715†	5.796†	4.142†
FK [20](ICML'25)	24.51†	0.6430†	0.3695†	69.02†	0.6836†	0.7336†	0.5635†	**5.691**†	4.088†
DSearch [18](arxiv'25)	24.49†	0.6426†	0.3698†	69.76†	0.6924†	0.7433†	0.5698†	5.705†	4.185†
KDS [21](NeurIPS'25)	24.54†	0.6345	0.3683†	68.40	0.6776†	0.7147†	0.5709†	5.993	4.032†
CATTO (Ours)	**24.56**†	**0.6448**†	**0.3667**†	**69.82**†	**0.6945**†	**0.7579**†	**0.5828**†	6.015	**4.235**†
PMRF [16] (Rectified-Flow)	24.47	**0.6935**	0.1657	62.23	0.7054	0.5574	0.5236	6.909	4.547
BoN [17](CVPR'25)	24.45	0.6889	0.1645†	62.35†	0.7071†	0.5648†	**0.5269**†	6.739†	**4.594**†
FK [20](ICML'25)	24.52†	0.6930	0.1661	62.20	0.7019	0.5651†	0.5214	6.839†	4.535
DSearch [18](arxiv'25)	24.52†	0.6933	0.1652†	62.32†	0.7018	0.5678†	0.5240†	7.028	4.536
KDS [21](NeurIPS'25)	**25.21**†	0.6771	0.2005	47.71	0.6271	0.5391	0.4964	8.235	3.301
CATTO (Ours)	24.49†	0.6902	**0.1638**†	**62.75**†	**0.7079**†	**0.5780**†	**0.5269**†	**6.648**†	4.545

† denotes improvement over the corresponding backbone baseline.

Table 9: Quantitative comparison of image super-resolution performance under different noise levels ($\sigma = 5$, 25, 50 and 75) on DRealSR. We evaluate our method based on SeeSR across various metrics.

Noise Level	Method	PSNR↑	SSIM↑	DISTS↓	LPIPS↓	ARNIQA↑	MANIQA↑	CLIPIQA↑	NIQE↓	LIQE↑
$\sigma = 5$	SeeSR	27.68	0.7567	0.2400	0.3357	0.6878	0.5088	0.6864	6.386	4.143
	Ours	**27.72**	**0.7616**	**0.2387**	**0.3295**	**0.6908**	**0.5237**	**0.7029**	**6.187**	**4.201**
$\sigma = 25$	SeeSR	24.59	0.6768	**0.2758**	0.4244	0.6978	**0.5502**	0.7225	6.170	**4.284**
	Ours	**24.79**	**0.6870**	0.2760	**0.4212**	**0.7001**	0.5488	**0.7332**	**6.140**	4.270
$\sigma = 50$	SeeSR	21.12	0.5853	0.3122	0.5197	0.6902	0.5592	0.7472	6.222	**4.372**
	Ours	**21.46**	**0.5908**	**0.3080**	**0.5149**	**0.6963**	**0.5643**	**0.7569**	**6.131**	4.336
$\sigma = 75$	SeeSR	18.74	0.4699	0.3473	0.6169	**0.6831**	0.5555	0.7428	6.236	4.212
	Ours	**18.91**	**0.4838**	**0.3442**	**0.6122**	0.6734	**0.5597**	**0.7491**	**6.203**	**4.232**

C.2 More Results on Image Restoration

In Tab. 8, we provide the results of PMRF [16] on the task of photo-realistic face restoration on CelebA-Test [62]. We can see that applying CATTO to PMRF also achieves better performance on most metrics than using other methods, such as FK and BoN. Note that, we utilize its y-conditional variant for evaluation.

C.3 Robustness Evaluation

To evaluate the robustness of our proposed method against complex and severe degradation scenarios, we conduct a comprehensive comparison by injecting varying levels of noise ($\sigma \in \{5, 25, 50, 75\}$) into the low-resolution inputs. As reported in Tab. 9, our method outperforms the strong baseline, SeeSR [10], across all evaluated noise levels.

Notably, as the degradation becomes more severe (from $\sigma = 5$ to $\sigma = 75$), the baseline experiences significant performance degradation. In contrast, our approach maintains a stable performance margin. Even under the most extreme noise condition ($\sigma = 75$), our method still shows improvements in both structural fidelity and perceptual quality. These sustained gains across almost all no-reference metrics (CLIPIQA, NIQE, *etc.*) demonstrate that our condition-aware refinement strategy is robust to input perturbations.

Table 10: Real-world image super-resolution comparison using different rewards with SeeSR on DRealSR [60].

Reward	Method	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	ARNIQA↑	CLIPIQA↑	MANIQA↑	NIQE↓	LIQE↑
NA	SeeSR [10]	28.07	0.7684	0.3174	65.08	0.6885	0.6903	0.5128	6.381	4.127
CLIPIQA	BoN [17]	28.11	0.7675	0.3221	65.75	0.6964	0.7197	0.5216	6.624	4.162
	FK [20]	28.19	0.7684	**0.3203**	65.04	0.6939	0.7098	0.5092	6.720	4.102
	DSearch [18]	28.08	0.7674	0.3246	65.46	0.6998	0.7216	0.5273	6.482	4.171
	CATTO (Ours)	**28.21**	**0.7698**	0.3291	**66.09**	**0.7008**	**0.7258**	**0.5279**	**6.461**	**4.180**
NIQE	BoN [17]	28.31	0.7719	0.3146	64.82	0.6908	0.6760	0.5008	5.859	4.052
	FK [20]	**28.39**	**0.7748**	**0.3101**	63.58	0.6900	0.6500	0.4894	5.950	3.986
	DSearch [18]	28.07	0.7669	0.3170	65.14	0.6942	0.6743	0.4999	**5.763**	4.062
	CATTO (Ours)	28.14	0.7673	0.3261	**66.12**	**0.6964**	**0.6920**	**0.5183**	5.944	**4.129**
ARNIQA	BoN [17]	28.18	0.7714	0.3181	65.48	0.7136	0.6875	0.5119	6.577	4.159
	FK [20]	**28.35**	**0.7760**	**0.3149**	64.45	0.7015	0.6717	0.5014	6.727	4.019
	DSearch [18]	27.94	0.7661	0.3240	66.01	**0.7190**	0.6971	0.5199	6.536	4.185
	CATTO (Ours)	28.16	0.7702	0.3229	**66.42**	0.7157	**0.7054**	**0.5267**	**6.399**	**4.189**
CLIPIQA + NIQE + ARNIQA	BoN [17]	28.12	0.7679	0.3189	65.59	0.7059	0.7036	0.5131	**6.238**	4.139
	FK [20]	27.96	0.7688	0.3208	65.54	0.7078	0.7055	0.5130	6.259	4.153
	DSearch [18]	**28.15**	**0.7702**	**0.3184**	65.22	0.6995	0.7024	0.5145	6.331	4.141
	CATTO (Ours)	28.13	0.7674	0.3294	**66.44**	**0.7087**	**0.7140**	**0.5275**	6.330	**4.180**

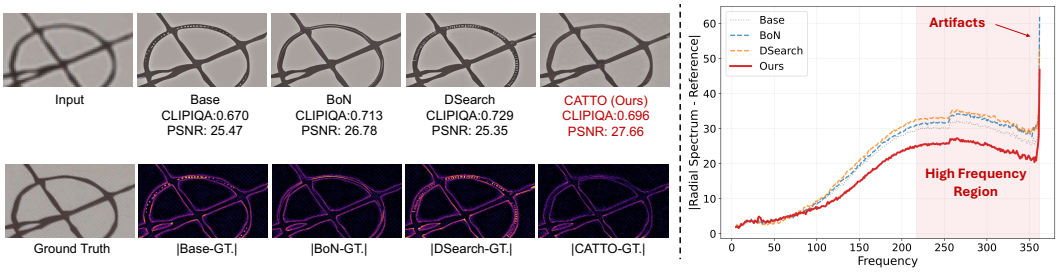

Figure 5: Illustration of image reconstruction before and after optimization. Left: reconstructed patches and error maps with respect to the reference (Ref.). Right: log-magnitude spectrum difference with respect to Ref. Our method produces cleaner reconstructions with fewer high-frequency artifacts.

C.4 More Ablation on IQA Reward

In addition to CLIPIQA used in the main experiments, Tab. 10 compares CATTO with BoN [17], FK [20], and DSearch [18] using NIQE, ARNIQA, and the combination of all three metrics as rewards. All methods use the same reward within each setting. Across the four settings, CATTO achieves competitive no-reference scores among the compared methods. Improvements also extend to other perceptual metrics in several settings, although the full-reference metrics vary and optimizing a metric does not always yield its best score. These results suggest that CATTO's perceptual gains are not confined to the reward used for selection.

C.5 Advantages of Condition-Level Optimization

Optimizing or searching over the latent variables z_t often diverts the sampling trajectory from the natural image distribution. This yields "adversarial" artifacts that overfit the reward model at the expense of true visual fidelity (see BoN [17] and DSearch [18] in Fig. 5). In contrast, refining the visual condition c_{im} routes updates through the pre-trained image encoder, inherently enforcing structural awareness and ensuring the refinements remain visually and physically valid. Furthermore, unlike text embeddings that offer only global semantic control, the visual condition provides fine-grained, spatially-aware guidance.

C.6 Additional Image Restoration Visual Results

To demonstrate the effectiveness of our method, we provide additional qualitative examples in Figs. 6, 7, 8, 9, 10, 11 and 12, including image super-resolution, face restoration, inpainting, and low-light enhancement, comparing the baseline model with different inference-time methods.

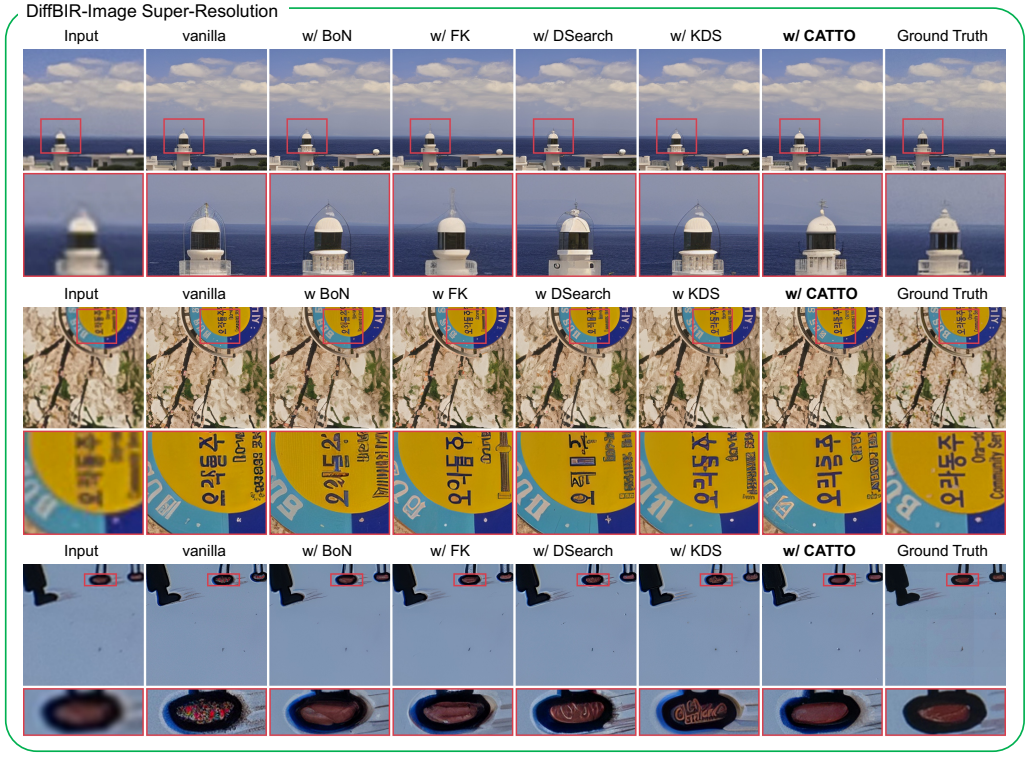

Figure 6: More Comparison Results on Image Super-Resolution with DiffBIR [14].

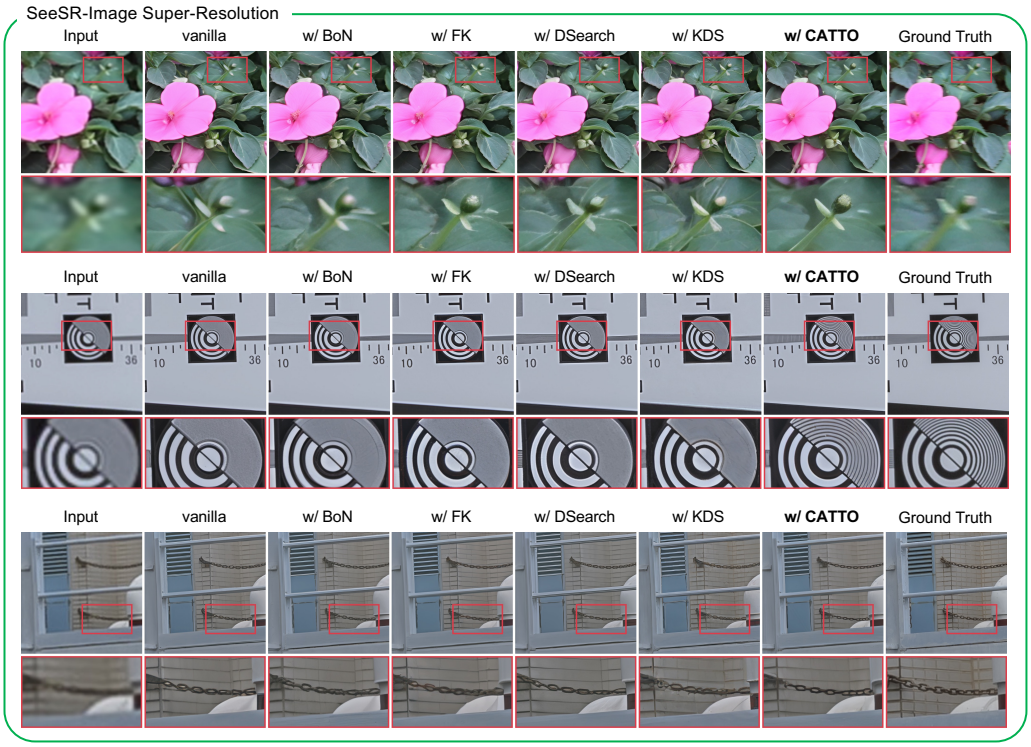

Figure 7: More Comparison Results on Image Super-Resolution with SeeSR [10].

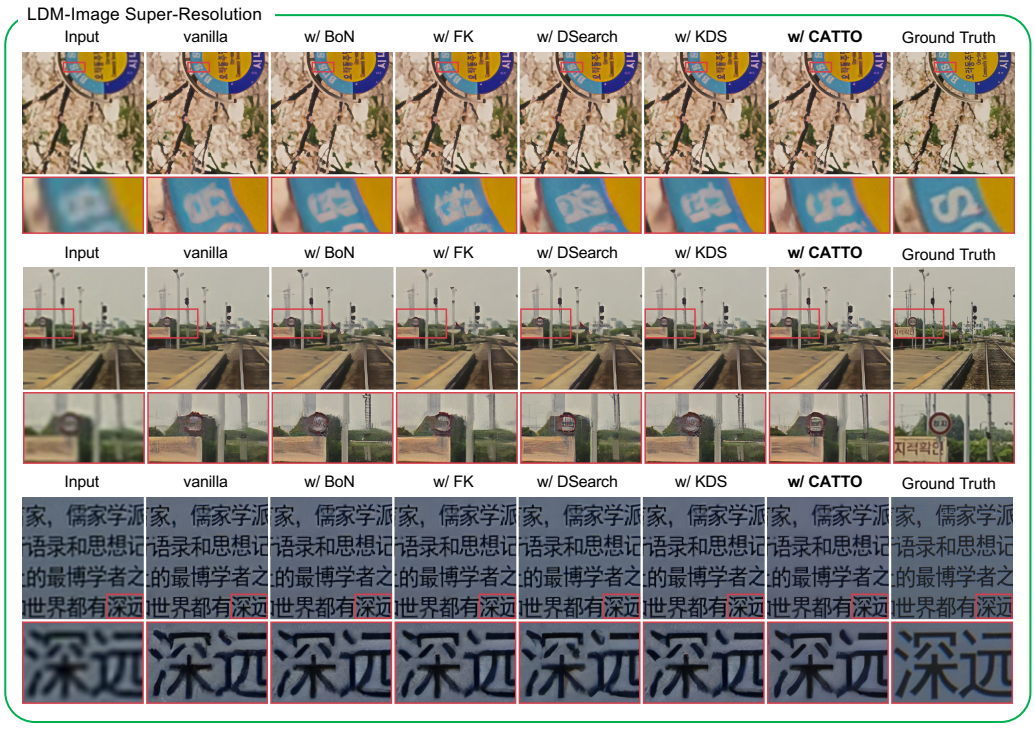

Figure 8: More Comparison Results on Image Super-Resolution with LDM [6].

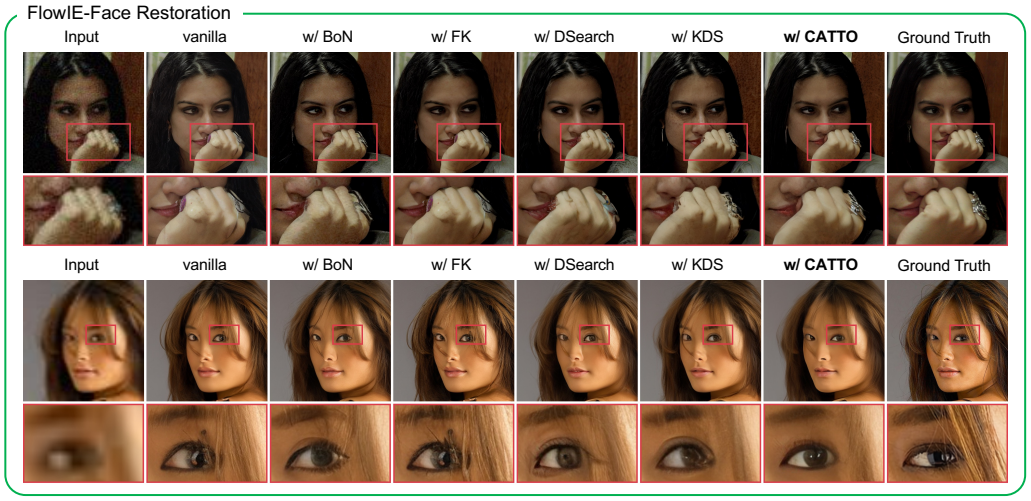

Figure 9: More Comparison Results on Face Restoration with FlowIE [57].



Figure 10: More Comparison Results on Face Restoration with PMRF [16].

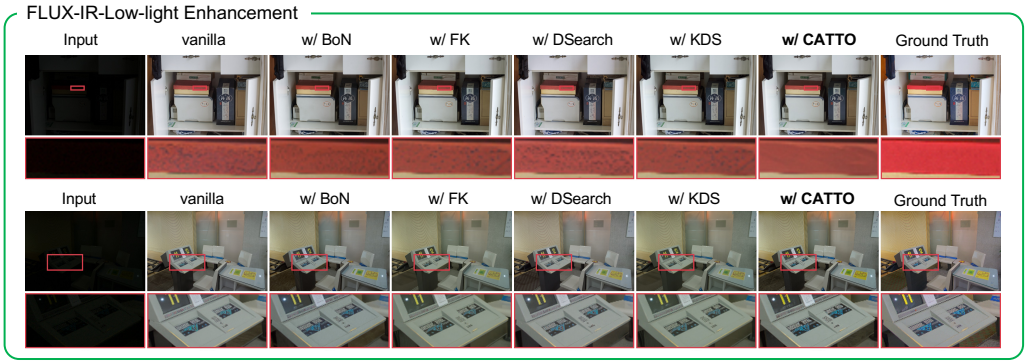

Figure 11: More Comparison Results on Low-light Enhancement with FLUX-IR [58].

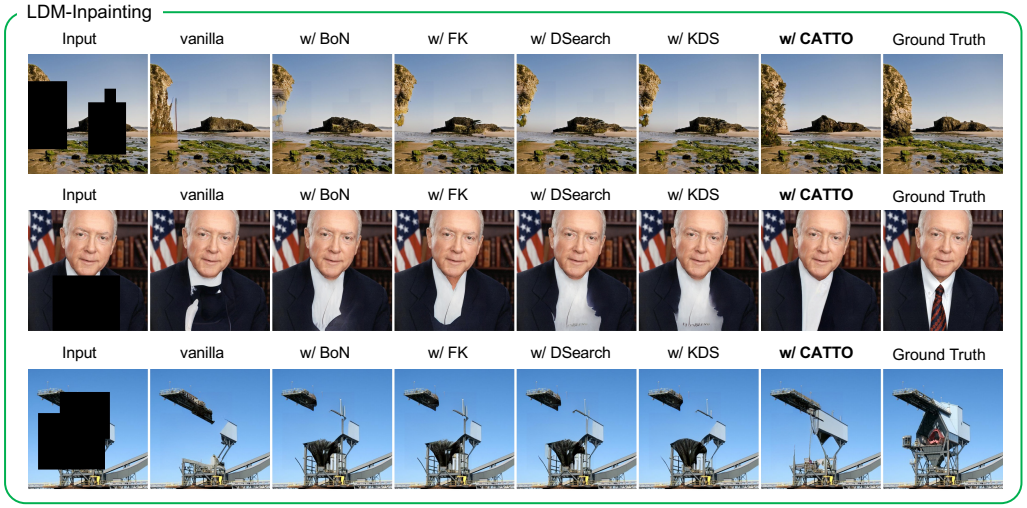

Figure 12: More Comparison Results on Image Inpainting with LDM [6].

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the paper's contributions and scopes.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See the main manuscript.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: See the "Experiments" section in the paper.

   Guidelines:

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

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

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

Answer: [No]

Justification: The details provided in the paper are sufficient to reproduce the proposed method. The training and testing datasets are derived from publicly available datasets. And the code will be released in the future.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: See the "Experiments" section in the paper.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: The experimental setup of this paper follows the methodology from previous work, which did not provide error bars. Besides, the experiments require significant memory and time resources, making statistical experiments too costly to provide.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: See the "Experiments" section in the paper.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: On the positive side, CATTO provides a training-free and efficient framework to improve image restoration quality, making high-fidelity enhancement more accessible for applications like remote sensing and legacy media preservation. By refining image conditions instead of just trajectories, it offers a more structured way to balance visual appeal with structural accuracy. However, the use of generative priors and perceptual rewards may still introduce "hallucinated" details that look realistic but lack factual authenticity. This poses risks in high-stakes fields such as medical imaging or legal forensics, where misinterpreted artifacts could lead to erroneous conclusions.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: See the "Experiments" section in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLM is used only for writing, editing, or formatting purposes.

    Guidelines:

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