DICT: Data Injection and Contrastive Trajectory Refinement for Conditional Image Generation with Diffusion Models

Chunnan Shang, Xin Zhang, Zhizhong Wang ✉, and Hongwei Wang ✉

Zhejiang University, Hangzhou, China
{chunnan.22, hongweiwang}@intl.zju.edu.cn
{22421264, endywon}@zju.edu.cn
https://github.com/scn-00/DICT

Abstract. Diffusion models have become a dominant paradigm for conditional image generation, yet existing approaches generally follow two directions: task-specific designs that can improve performance but limit generalization, and training-free loss guidance that compresses rich conditions into scalar objectives and applies stepwise guidance, leading to information bottlenecks and error accumulation along the sampling trajectory. Given the urgent need for an effective unified framework across diverse conditional image generation tasks, we propose **D**ata **I**njection and **C**ontrastive **T**rajectory Refinement (**DICT**), a training-free inference method that enhances conditional image generation without introducing task-dependent architectures. DICT introduces Data Injection, where noise-perturbed conditional signals are integrated into early denoising stages; by performing guided denoising on these injected signals, DICT adaptively selects and distills task-salient information from the raw condition, effectively preserving spatial richness and ensuring precise condition-to-generation alignment. Furthermore, DICT applies Contrastive Trajectory Refinement across adjacent denoising states, enabling pairwise comparisons that progressively improve sample quality. These designs keep inference simple while improving cross-task transfer under a unified diffusion formulation. Extensive experiments on conditional image generation tasks (e.g., style transfer, image super-resolution, and image deblurring) show consistent gains in fidelity and perceptual quality over representative task-specific and loss-guided baselines.

Keywords: Conditional Image Generation · Diffusion Models · Training-Free Inference

1 Introduction

Diffusion models (DMs) [8, 22] have become a dominant paradigm for conditional image generation, largely due to their strong priors and flexible sampling dynamics. In many practical conditional image generation settings (e.g., style

✉Corresponding authors.

transfer, image super-resolution, and image deblurring), the informative value of the given conditions is often restricted. For instance, a condition might capture merely partial target features [12,15] or represent a corrupted observation [11,26]. Such suboptimal conditions inherently hinder faithful generation, requiring the model to extract useful signals while suppressing misleading cues.

Existing approaches can be categorized into two major classes of methods: (1) *condition-specific designs*, which tailor architectures or constraints to particular condition types, such as feature injection and fusion for style transfer [4,15,25] or data-consistency and trajectory relaxation for image super-resolution and image deblurring tasks [11, 23, 28]; and (2) *generic inference-time loss guidance*, which formulates condition satisfaction as a scalar loss and utilizes its gradients to guide the reverse diffusion trajectory without additional training [2, 29, 30].

While broadly applicable, loss-guided inference suffers from two key limitations. First, it compresses the condition reference into a single scalar loss, so guidance is driven primarily by the gradient of this objective. This scalarization creates an information bottleneck: the structured, high-dimensional content in the condition reference is not fully retained in the resulting gradient signal, leading to updates that are often too coarse to provide fine-grained control in complex tasks. For instance, in style transfer, a loss function might indicate how far a stylized image is from the target style overall, but it lacks the granular information needed to specify specific elements like texture patterns and brush strokes, which exist only within the high-dimensional data itself and are not reducible to a single scalar. Furthermore, loss guidance using scalar objectives leads to compounding error accumulation. In the iterative denoising process, inaccuracies at any step persist and propagate along the sampling trajectory, increasingly biasing the generation and degrading final sample quality. Alternatively, while latent initialization methods like SDEdit [17] avoid scalarization by adding noise to a reference image and denoising it, they typically employ a one-shot integration strategy. This lacking of continuous condition anchoring often leads to trajectory drift, and their inability to adaptively filter out irrelevant condition information restricts their general efficacy.

To address these issues, we propose **D**ata **I**njection and **C**ontrastive **T**rajectory Refinement (**DICT**), a general inference framework for conditional image generation that goes beyond scalar loss guidance and avoids task-specific redesign. Our motivation is twofold. First, although task-specific architectures or constraints can better exploit a given condition, they often encode condition-dependent assumptions and thus transfer poorly across tasks and condition types. To improve generality, DICT avoids explicit constraints and instead leverages the *condition as data*: we diffuse the condition signal and inject its noise-perturbed counterpart into the early denoising steps. This design allows the model to access richer conditional information than a scalar loss, selectively amplifying informative cues while suppressing redundant or degraded components. Second, even with a better condition representation, stepwise loss optimization remains prone to error accumulation along the reverse process. To stabilize sampling in a model-agnostic manner, compatible with U-Net [22] and Transformer-based backbones [19], we

exploit the temporal structure of the reverse diffusion process and introduce a *Contrastive Trajectory Refinement* mechanism. Concretely, we optimize a contrastive trajectory objective that enforces step-to-step improvement, ensuring that the prediction at each timestep is more aligned with the condition signal than the prediction at the previous timestep, thereby reducing error propagation and promoting consistent refinement along the sampling path.

Overall, our main contributions are as follows:

- We analyze limitations of existing conditional image generation methods: condition-specific designs can improve task performance but generalize poorly across tasks, while generic inference-time loss guidance compresses rich conditions into scalar objectives, leading to information bottlenecks and error accumulation along the sampling process.
- We propose **D**ata **I**njection and **C**ontrastive **T**rajectory Refinement (**DICT**), a training-free inference method that enhances conditional image generation without introducing task-dependent architectures. DICT injects noise-perturbed conditional signals in early denoising steps to strengthen condition alignment, and further refines the reverse process via a contrastive trajectory objective across adjacent denoising states to progressively improve quality.
- Extensive experiments on conditional image generation tasks (e.g., style transfer, image super-resolution, and image deblurring) demonstrate consistent improvements in fidelity and perceptual quality over representative task-specific and loss-guided baselines.

2 Related Work

Style Transfer. We focus on text-driven style transfer, which aims to generate images that follow the text content while adopting the style of a reference image. Existing methods can be divided into two categories: those that learn feature mappings and those that exploit prior features. Feature-mapping approaches learn to project style/content attributes into dedicated spaces and model their relations, e.g., mapping style features into the diffusion model's CLIP space (StyleShot [6]), learning a style feature space from CLIP image features (StyleCrafter [15]), or explicitly decomposing CLIP image features into content and style representations (CSGO [27], DEADiff [21]); related work also maps style attributes via lightweight tuning with human feedback (StyleDrop [24]). Prior-feature exploitation approaches instead leverage pretrained representations to extract or inject style cues, such as subtracting text content embeddings from image embeddings (InstanStyle [25]), decoupling content via ControlNet [32] to obtain cleaner style features (StyleStudio [12]), or injecting and aligning attention features (queries/keys/values) from the style image (StyleAlign [7]).

Super-resolution and Deblurring. We consider super-resolution and deblurring, recovering a clean image from a low-resolution or blurred observation. Existing approaches mainly follow two directions: methods based on strict constraints and methods based on flexible sampling. Constraint-based methods

enforce explicit data-consistency to match the degraded observation, e.g., multi-stage consistency (SITCOM [1]), correction/projection constraints (PSLD [23], [28]), or lightweight gradient-based consistency enforcement (DCDP [14]). Flexible sampling methods relax or reinterpret the reverse diffusion process, such as viewing posterior sampling as Bayesian filtering (FPS-SMC [5]), formulating sampling as discretized optimal control (DOC [13]), or steering via vector-field objectives and scheduled constraint relaxation (FlowChef [18], FlowDPS [11]). Beyond these general strategies, many task-specific designs have also been explored for super-resolution, e.g., SeeSR [26] and InvSR [31].

Compared to task-specific methods, DICT directly conditions on the given image and adaptively extracts useful cues by injecting noise-perturbed conditional signals into early denoising steps and denoising them along the reverse process, avoiding biases from learned condition-dependent feature mappings and the partial information captured by fixed pretrained features. Moreover, DICT does not rely on strict task-specific consistency constraints or modified sampling procedures; instead, it leverages the conditional input to apply a Contrastive Trajectory Refinement objective across adjacent denoising states, enabling pairwise step-to-step comparisons that progressively improve condition alignment and sample quality in a unified and more general manner.

Loss-gradient Guidance. Training-free loss-guided methods have emerged as a versatile paradigm to incorporate complex constraints into pretrained diffusion models without the need for task-specific fine-tuning. These methods typically leverage the gradients of external loss functions to shift the sampling trajectory towards a target region defined by the condition. For instance, FreeDoM [30] achieves conditional generation by utilizing off-the-shelf networks to construct energy functions, employing an iterative time-travel strategy to enforce semantic consistency. TFG [29] unifies them into a comprehensive design space, proposing an efficient search strategy to dynamically optimize guidance parameters for various tasks. However, a major bottleneck of these methods is the increased computational overhead caused by repeated gradient computations. To address this, Adaptive Guidance [2] accelerates inference by adaptively omitting the unconditional score evaluation once it converges with the conditional estimate.

Unlike previous methods, DICT goes beyond scalar loss gradients by injecting a noise-perturbed condition in early denoising steps to exploit rich conditional signals. It further introduces a Contrastive Trajectory Refinement that enforces step-to-step improvement, mitigating error accumulation, and improving quality.

3 Method

3.1 Revisit Latent Diffusion Models

The latent diffusion models [22] operate in latent space, using a forward process that gradually adds noise to reach a standard Gaussian and a reverse process that steadily denoises to generate samples.

Thus, the clean latent vector z_0 can be first obtained:

$$z_0 = E(x), \tag{1}$$

where E is an encoder of [22]. x denotes the clean data.

The forward diffusion is a Markov chain, where the latent vector z_t at each state $t \in (1, T)$ is obtained by adding Gaussian noise to the clean latent vector z_0:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_t)$, $\beta_t \in (0, 1)$ adopts a fixed variance schedule. $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the noise.

The reverse diffusion starts from a Gaussian sample $z_T \sim \mathcal{N}(0, \mathbf{I})$ and progressively denoises. Based on PLMS [16], each preceding state z_{t-1} is obtained:

$$\begin{aligned} z_{0|t} &= \frac{z_t - \sqrt{1 - \bar{\alpha}_t}\,\epsilon_\theta(t)}{\sqrt{\bar{\alpha}_t}}, \\ z_{t-1} &= \sqrt{\alpha_{t-1}}\, z_{0|t} + \sqrt{1 - \alpha_{t-1} - \sigma_t^2}\epsilon_\theta(t) + \sigma_t z, \end{aligned} \tag{3}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$ and $\epsilon_\theta(t) = \sum_{i=0}^{k-1} w_i \epsilon_\theta(z_{t-i}, t-i)$. $k = 4$ is for the historical steps. w_i is the weighting coefficient. ϵ_θ is predicted noise, implemented by a backbone such as U-Net [22] or DiT [19]. For this paper, we use the U-Net as our foundational model. $\sigma_t^2 = 0$ for deterministic sampling.

Finally, the target image x can be obtained via decoding:

$$x = D(z_0), \tag{4}$$

where D represents the decoder of [22].

3.2 Task Setup for Conditional Image Generation

Conventionally, super-resolution and deblurring are treated as fundamentally distinct from text-driven style transfer due to their differing natures—explicit physical degradations versus abstract semantic conditions. To bridge this conceptual gap, we mathematically frame them as generalized posterior sampling, providing a unified formulation under conditional image generation.

Formally, let y denote the unknown clean image (or latent) to be generated, and let x denote the available condition, which can be a semantic/style specification (e.g., text and a style reference image) or a degraded observation (e.g., low-resolution/blurred image). Both tasks can be cast as sampling from the conditional distribution:

$$y \sim p(y|x), \quad p(y|x) \propto p(x|y)p(y), \tag{5}$$

where $p(y)$ is the diffusion prior learned from data, and $p(x|y)$ measures the compatibility between the generated image y and the condition x. Equivalently, we can target the Maximum a Posteriori (MAP) estimation:

$$y^* = arg\min_y -logp(x|y) - logp(y). \tag{6}$$

Style Transfer. For text-driven style transfer, the condition $x = (t, s)$ combines a text prompt t (content) and a style image s (style). Under our unified posterior view:

$$p(y|t, s) \propto p(t, s|y)p(y). \tag{7}$$

Assuming t and s are conditionally independent given y, we can factorize the conditional term as $p(t, s|y) = p(t|y)p(s|y)$. This leads to:

$$p(y|t, s) \propto p(s|y)p(y|t), \tag{8}$$

where $p(y|t) \propto p(t|y)p(y)$ represents the text-conditioned prior. In practice, $p(y|t)$ is implicitly learned by a pretrained diffusion model (e.g., via classifier-free guidance), while $p(s|y)$ is explicitly defined using an energy-based surrogate:

$$-\log p(s|y) \approx \mathcal{L}_{style}(y, s). \tag{9}$$

Therefore, MAP inference under these conditions can be formulated as:

$$y^* = \arg\min_y \mathcal{L}_{style}(y, s) - \log p(y|t), \tag{10}$$

where the gradient of $-\log p(y|t)$ is provided by the conditioned score function of the diffusion model.

Super-resolution and Deblurring. For super-resolution or deblurring, the condition x results from a known degradation operator $\mathcal{A}(\cdot)$:

$$x = \mathcal{A}(y) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}), \tag{11}$$

which yields a standard data-fidelity term $-logp(x|y) \propto ||x - \mathcal{A}(y)||_2^2$. Therefore, MAP inference under degraded conditioning can be written as:

$$y^* = arg\min_y \frac{1}{2\sigma^2}||x - \mathcal{A}(y)||_2^2 - logp(y). \tag{12}$$

This unified view motivates inference procedures that exploit conditional signals while remaining robust to condition information.

3.3 Overall Framework of DICT

We introduce **D**ata **I**njection and **C**ontrastive **T**rajectory Refinement, hereafter **DICT**, for conditional image generation (Fig. 1), a general inference method beyond loss guidance. It injects noise-perturbed conditions in early denoising and enforces step-to-step improvement with a contrastive trajectory objective.

Data Injection. In conditional image generation, the fidelity of the output is dictated by how effectively the condition x is integrated into the sampling trajectory. Existing training-free methods predominantly rely on a scalar-guidance paradigm, where rich, high-dimensional conditions x are compressed into a single scalar loss objective. This drastic reduction in dimensionality creates a severe information bottleneck: a scalar value is fundamentally insufficient to encapsulate

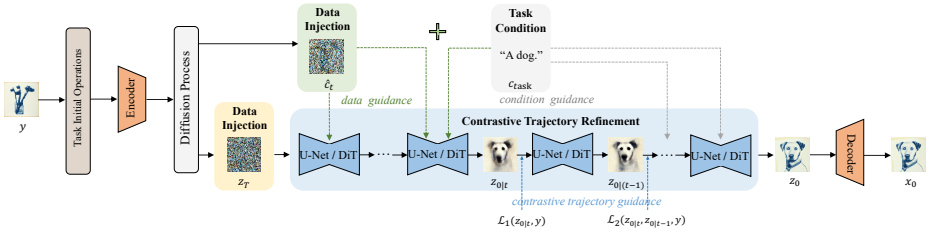


Fig. 1: Overview of DICT. Data Injection is integrated into the early reverse diffusion stages to retain conditional richness, leveraging guided denoising to adaptively extract task-salient information from the perturbed condition. Simultaneously, Contrastive Trajectory Refinement spans the entire trajectory to ensure coherence and refine the generative path for superior quality. Task-specific conditions, where applicable, remain persistent throughout the process to ensure consistent guidance.

the spatial complexity and fine-grained textures of the condition. Consequently, the resulting gradients often fail to recover subtle cues, leading to a significant loss of fidelity as the model hallucinates details to fill the information void.

To circumvent this information collapse, we propose a direct Data Injection mechanism. Instead of relying solely on a simplified scalar proxy, we incorporate noise-perturbed versions of the condition x directly into the diffusion inference loop during the early denoising stages (Fig. 1, Fig. 2(a)). By denoising these injected signals alongside the latent variables throughout the reverse process, the model can adaptively distill task-relevant information from the raw condition. This enables the generative trajectory to be grounded in the target data manifold, ensuring that the final output preserves high-frequency details and structural nuances that are typically lost in traditional loss-guided frameworks.

Specifically, given a condition x, we choose an operation $\mathcal{M}(\cdot)$ to transform x into a higher-quality initial condition that amplifies useful signals for guidance:

$$\hat{x} = \mathcal{M}(x), \tag{13}$$

where $\mathcal{M}$ is a task-dependent but lightweight preprocessing step that better aligns the condition with the guidance objective. It is worth noting that although $\mathcal{M}$ is inherently task-specific (analogous to task-dependent loss functions), it introduces no trainable parameters and does not modify the internal structure of the base model. Consequently, our "architecture-independent" claim remains valid. Details of $\mathcal{M}$ are provided in Sec. B of the Supplementary Materials.

We then use the processed condition $\hat{x}$ to initialize the reverse diffusion with z_T at time step T, and to compute the condition latent $\hat{c}_t$ used at each denoising step t:

$$\begin{aligned} \hat{z} &= E(\hat{x}), \\ z_T &= \sqrt{\bar{\alpha}_T}\hat{z} + \sqrt{1-\bar{\alpha}_T}\epsilon, \\ \hat{c}_t &= \sqrt{\bar{\alpha}_t}\hat{z} + \sqrt{1-\bar{\alpha}_t}\epsilon_{ti}, i = 1, \ldots, N_{iter1}, \end{aligned} \tag{14}$$

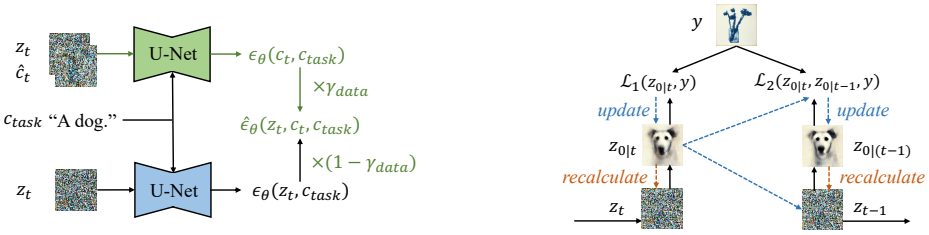
where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. For $i = 1$, we set $\epsilon_{ti} = \epsilon$; otherwise, $\epsilon_{ti} = \epsilon_\theta(t)$ (Eq. 15). N_{iter1} denotes the number of iterations. The condition latent $\hat{c}_t$ is a noise-perturbed version of the condition x that preserves its high-dimensional spatial richness. Rather than injecting the condition data directly, we add noise to it to facilitate selective information exploitation. This strategy is grounded in the observation that task-relevant features—such as stylistic textures in style transfer or semantic structures in image restoration—typically manifest as the dominant signals within the image space. Stochastic noise effectively disrupts redundant and irrelevant information, whereas the dominant conditional structures remain resilient and perceptible to the diffusion prior, thanks to their high-dimensional robustness. Consequently, compared to conventional guidance methods that compress rich conditions into a single scalar loss—thereby creating an irreversible information bottleneck—our noise injection strategy effectively maintains the integrity of the conditional cues. By grounding the sampling process in the noise-perturbed image manifold rather than a simplified scalar proxy, DICT enables the model to adaptively distill task-salient signals while suppressing harmful interference, leading to superior generation quality and fidelity.

Subsequently, we inject lightweight variants of the condition data into the early reverse-diffusion steps to improve output fidelity and sampling efficiency:

$$\begin{aligned} c_t &= \alpha_{data} \times z_t + (1 - \alpha_{data}) \times \hat{c}_t, \\ \hat{\epsilon}_\theta(z_t, c_t, c_{task}) &= \gamma_{data} \times \epsilon_\theta(c_t, c_{task}) \\ &\quad + (1 - \gamma_{data}) \times \epsilon_\theta(z_t, c_{task}), \\ \epsilon_\theta(t) &= \hat{\epsilon}_\theta(z_t, \hat{c}_t, c_{task}), \end{aligned} \tag{15}$$

where z_t and $\hat{c}_t$ are integrated via weighted summation, enabling their information to interact and be adaptively refined during denoising. α_{data} and γ_{data} are weighting factors, with values reported in Sec. B of the Supplementary Materials. c_{task} denotes the task-level condition: the text prompt for style transfer, and empty for super-resolution and deblurring.

Discussion. SDEdit [17] synthesizes images by noising a user-provided image to a fixed timestep t_0 and then denoising it to reach the data manifold. Our DICT differs in three key aspects: (i) *Role of the Condition.* SDEdit treats the condition primarily as a latent initialization (a starting point). It relies on the model's unconditional prior to "hallucinate" missing details based on the noisy layout of the input. In contrast, DICT treats the condition as an active reference source. Instead of just starting from latent initialization, DICT explicitly distills specific high-dimensional cues (e.g., stylistic textures for style transfer or semantic skeletons for restoration) from the condition throughout the denoising guidance process. This ensures the output is grounded in the condition's actual features rather than just its coarse layout. (ii) *Integration Frequency.* SDEdit is a one-shot strategy that only conditions the generative process at the initial timestep t_0. This often leads to a "drift" where the model loses track of the condition in later stages. In contrast, DICT performs multi-step injection over a subset of the denoising trajectory. This continuous anchoring provides a more direct and stable path toward the target, effectively reducing suboptimal solutions



(a) Data Injection (b) Contrastive Trajectory Refinement

Fig. 2: The detailed schematics illustrate the Data Injection and Contrastive Trajectory Refinement within the DICT framework.

common in fixed-initialization methods. (iii) *Adaptive Selectivity.* SDEdit lacks a mechanism to distinguish between relevant and irrelevant information—it attempts to reconstruct the entire noisy input. DICT, however, leverages the noise prediction, i.e., $\epsilon_{ti} = \epsilon_\theta(t)$ in Eq. 15 as a natural filter. By injecting noise and denoising under guidance, DICT adaptively ignores harmful or irrelevant content (e.g., the objects in a style image) and only attends to the task-salient signals.

We then compute the predicted clean latent $z_{0|t}$:

$$z_{0|t} = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(t)}{\sqrt{\bar{\alpha}_t}}, \tag{16}$$

where $z_{0|t}$ is expected to be consistent with the condition x. We therefore optimize $z_{0|t}$ to better capture condition-relevant target information:

$$\begin{aligned} \mathcal{L}_1(z_{0|t}, x) =& f_{loss}(D(z_{0|t}), x), \\ z_{0|t} =& z_{0|t} - \eta_1 \times \nabla_{z_{0|t}} \mathcal{L}_1(z_{0|t}, x), \end{aligned} \tag{17}$$

where f_{loss} is the task loss function. η_1 controls the step size of gradient updates to $z_{0|t}$. More details are in Sec. B and Sec. C of the Supplementary Materials.

After that, we obtain the z_{t-1} with rich information:

$$z_{t-1} = \sqrt{\alpha_{t-1}}\, z_{0|t} + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\epsilon_\theta(t) + \sigma_t z, \tag{18}$$

where $\sigma_t = 0$. While Eq. 17 encourages $z_{0|t}$ to match x, it does not guarantee that predictions at later steps are more accurate than those at earlier steps, since errors can accumulate monotonically along the trajectory without an explicit correction mechanism.

Data Injection as a Non-gradient Score Proxy. To bypass scalar gradient bottlenecks, our Data Injection (Eq. 15) uses latent spatial displacement as a score surrogate. DICT models $\hat{c}_t$ as an empirical Dirac distribution to implicitly maximize high-dimensional likelihood and preserve complex information. Details are in Sec. A of the Supplementary Materials.

Contrastive Trajectory Refinement. In conditional image generation tasks, relying on a scalar-level loss $\mathcal{L}_1$ has fundamental limitations. Such a loss

Table 1: Hyperparameter settings for different tasks in DICT.

Hyperparameters	Style Transfer	Image Super-resolution	Image Deblurring
Base Model	SD v1.4	LDM	LDM
Sampling Steps	50	200	200
Latent Weight (α_{data})	0.88	0.88	0.88
Noise Weight (γ_{data})	0.2	0.1	0.1
Iterative Updates (N_{iter1}, N_{iter2})	3, 3	3, 3	3, 4
Initial Update Rate (η_1, η_2)	$2^\diamond$, $2^\diamond$	[0.08, 0.015, 0.006]†, 0.02	[0.13, 0.015, 0.006]*, 0.02
Data Injection Steps (T_1)	1-18	1-8	1-50
Contrastive Trajectory Margin (α_{margin})	1.0	1.0	1.0

$^\diamond$ This task employs a varying parameter schedule [29] for η_1 and η_2.
† This task employs a three-stage schedule for η_1: steps 1–130, 131–160, and 161–200.
* This task employs a three-stage schedule for η_1: steps 1–130, 131–150, and 151–200.

provides only a static target at a single timestep and does not reflect the temporal structure of the reverse-diffusion trajectory. In practice, this leads to locally optimal updates at each discrete timestep, rather than a coherent refinement trajectory, i.e., $z_{0|t-1}$ is not guaranteed to be a principled improvement over $z_{0|t}$. Without an explicit mechanism to correct earlier deviations, small inaccuracies introduced at early steps can propagate and compound, biasing the sampling trajectory and degrading the final result. To address these issues, we leverage the reverse-diffusion trajectory structure (Fig. 1 and Fig. 2(b)) and impose a Contrastive Trajectory Refinement principle: each intermediate prediction should be closer to the condition target than the previous one. This trajectory-aware design supports incremental correction and favors more reliable denoising paths, thereby mitigating error accumulation via a contrastive trajectory objective:

$$\begin{aligned} \mathcal{L}_2 &= \max\big(\mathcal{L}_1(z_{0|t-1}, y) - \mathcal{L}_1(z_{0|t}, y) + \alpha_{margin}, 0\big), \\ z_{0|t-1} &= z_{0|t-1} - \eta_2 \times \nabla_{z_{0|t-1}} \mathcal{L}_2, \end{aligned} \tag{19}$$

where α_{margin} enforces a margin, requiring the negative sample $z_{0|t}$ to remain at least a fixed distance away from the positive sample $z_{0|t-1}$. η_2 controls the step size of the gradient-based update applied to $z_{0|t-1}$. All parameter settings are provided in Sec. C of the Supplementary Materials. With this strengthened trajectory prior, $z_{0|t-1}$ is encouraged to be closer to the target than $z_{0|t}$.

Ultimately, we obtain the noisy latent z_{t-1} which integrates more accurate and higher-quality condition information:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} z_{0|t-1} + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon. \tag{20}$$

Iterating this reverse process yields the final output:

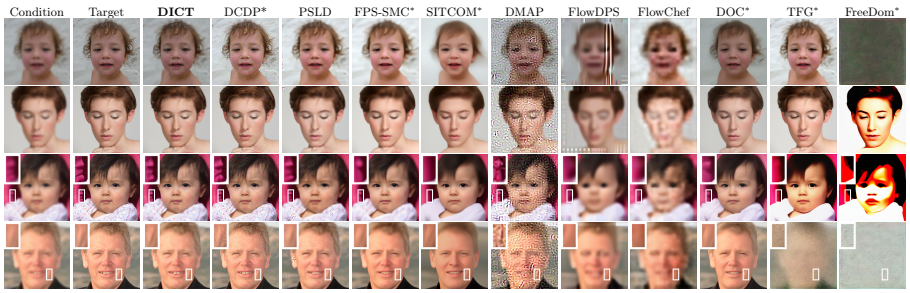
$$x_0 = D(z_0), \tag{21}$$

where $x_0 = y$ denotes the final output, i.e., the target sample generated under the provided condition.

Contrastive Trajectory Refinement as a Discrete Lyapunov Stabilizer. Solver truncation introduces Local Truncation Errors (LTE, $\mathcal{O}(\Delta t^{k+1})$)

(a) Qualitative comparisons of text-to-image style transfer task.

(b) Qualitative comparisons of image super-resolution task.



(c) Qualitative comparisons of image deblurring task.

Fig. 3: Qualitative comparisons. An asterisk $(*)$ denotes models in the pixel space; those unmarked operate in the latent space.

causing tangential drift. Our $\mathcal{L}_2$ acts as a discrete Lyapunov constraint enforcing monotonic energy decay ($\Delta\mathcal{L}_1 \leq -\alpha_{margin}$). Functioning as a soft Manifold Projection Operator, it generates a normal force to mitigate orthogonal LTE, promoting coherence. Details are in Sec. A of the Supplementary Materials.

4 Experiments

4.1 Experiment Settings

We study conditional image generation under three tasks: style transfer, super-resolution, and deblurring. For text-driven style transfer, we evaluate on 40,000

Table 2: Quantitative comparisons of the conditional image generation. The best results are in **bold**. The second best results are underlined.

(a) Quantitative comparisons of text-to-image style transfer task.

	DICT	StyleShot	StyleStudio	StyleCrafter	DEADiff	InstantStyle	StyleAlign	CSGO	StyleDrop	TFG	FreeDom
Text Score ↑	0.2952	0.2740	0.2852	0.2582	0.2907	0.1920	0.2553	0.2688	0.2780	**0.3092**	0.2933
Style Loss ↓	**0.6313**	0.6747	0.9996	0.7444	1.7697	1.2691	0.8252	0.6793	1.3341	2.1120	2.7492
CLIP Loss ↓	**4.2334**	6.0415	7.7899	7.0266	9.3485	8.4820	6.9670	6.5944	8.5434	7.7799	11.1131

(b) Quantitative comparisons of image super-resolution task.

	DICT	InvSR	PSLD	FPS-SMC*	SITCOM*	DMAP	FlowDPS	FlowChef	DOC*	TFG*	FreeDom*
PSNR Score ↑	**28.8600**	21.5851	24.9210	27.6305	25.3733	26.3395	25.0459	23.1086	26.7617	26.3441	10.7963
SSIM Score ↑	0.8233	0.6713	0.6397	**0.8283**	0.7805	0.7840	0.7438	0.7458	0.8167	0.7979	0.2546
LPIPS Loss ↓	**0.1573**	0.2374	0.2560	0.2029	0.2460	0.2499	0.3114	0.2868	0.1717	0.2106	0.7190

(c) Quantitative comparisons of image deblurring task.

	DICT	DCDP*	PSLD	FPS-SMC*	SITCOM*	DMAP	FlowDPS	FlowChef	DOC*	TFG*	FreeDom*
PSNR Score ↑	27.5794	**27.9110**	25.8065	25.7486	23.0995	18.8024	21.1828	20.5104	25.2189	22.6209	12.3003
SSIM Score ↑	**0.7736**	0.7384	0.7566	0.7665	0.7082	0.3096	0.5987	0.6067	0.7362	0.7149	0.3105
LPIPS Loss ↓	**0.2236**	0.2325	0.2675	0.2540	0.3100	0.5541	0.4887	0.4934	0.2448	0.2869	0.6764

stylized images at 512×512 resolution, constructed by pairing 200 text prompts with 200 style images randomly sampled from WikiArt [20]. For super-resolution and deblurring, we evaluate on 1,000 randomly selected FFHQ images [10]. In super-resolution, we downsample each image by a factor of $4\times$ and add Gaussian noise with standard deviation 0.01, and generate outputs at 256×256 resolution. In deblurring, we apply a Gaussian blur kernel of size 61 with standard deviation 3.0, add Gaussian noise with standard deviation 0.01, and generate outputs at 256×256 resolution. Hyperparameter settings for different tasks of DICT are in Tab. 1. Detailed settings are in Sec. B of the Supplementary Materials.

4.2 Comparison

For style transfer, we primarily compare against latent-space methods (including StyleShot [6], StyleStudio [12], StyleCrafter [15], DEADiff [21], InstantStyle [25], StyleAlign [7], CSGO [27], StyleDrop [24], TFG [29], and FreeDom [30]), as they align better with generation conditioned on text and images. For super-resolution and deblurring, prior work spans two mainstream paradigms: pixel-space approaches (including FPS-SMC [5], SITCOM [1], DOC [13], TFG [29], FreeDom [30], and DCDP [14]), which often excel in pixel-level fidelity metrics, and latent-space approaches (including InvSR [31], PSLD [23], DMAP [28], FlowDPS [11], and FlowChef [18]), which are typically more efficient and flexible. Although our method is implemented in the latent space, we include both pixel-space and latent-space baselines for a comprehensive evaluation.

Qualitative Comparisons. We present a qualitative comparison.

In Fig. 3(a), prior methods often exhibit weak stylization (e.g., DEADiff, StyleDrop, TFG, and FreeDom), inaccurate prompt adherence (e.g., StyleStudio, StyleCrafter, and CSGO), or content leakage from the reference style image (e.g., StyleShot, InstantStyle, and StyleAlign). In contrast, our DICT better preserves

text semantics (e.g., snow in the $2nd$ row and a white towel in the $3rd$ row) while achieving stronger stylization (e.g., the dog in the $1st$ row).

In Fig. 3(b), competing approaches frequently produce artifacts (e.g., PSLD and DMAP), reduced photorealism (e.g., InvSR, FPS-SMC, and SITCOM), or overly smooth results with missing details (e.g., FlowDPS, FlowChef, and DOC), and in some cases deviate from the data distribution (e.g., TFG and FreeDom). DICT avoids noticeable artifacts and yields clearer, more realistic, and visually richer outputs (e.g., the $1st$ and $2nd$ rows), while recovering fine-grained details such as the mole in the $3rd$ row and the temple hair in the $4th$ row.

In Fig. 3(c), prior methods often exhibit distribution bias (e.g., DCDP, PSLD, FPS-SMC, SITCOM, FlowChef, and DOC), artifacts (e.g., DCDP, DMAP, and FlowDPS), and blurriness (e.g., FlowChef), and some results deviate from the data distribution (e.g., TFG and FreeDom). By comparison, DICT produces outputs that better match the target distribution (e.g., the background color in the $1st$ row) and restores fine details (e.g., the folds of the hat in the $3rd$ row).

Overall, DICT synergistically integrates Data Injection to preserve spatial information richness and Contrastive Trajectory Refinement to ensure temporal trajectory coherence. This framework enables high-fidelity conditional image generation that consistently outperforms task-specific and loss-guided baselines.

Quantitative Comparisons. We also conduct quantitative comparisons.

For text-driven style transfer, we evaluate using Text Score [24] (prompt-image CLIP similarity), Style Loss [9] (VGG feature statistics discrepancy), and CLIP Loss [29] (CLIP Gram loss). As shown in Tab. 2(a), our method achieves the second-highest Text Score while obtaining the lowest Style Loss and CLIP Loss, indicating strong prompt fidelity and the most consistent stylization. Although TFG achieves a marginal lead in Text Score, it incurs substantially higher Style and CLIP losses, suggesting weaker style consistency.

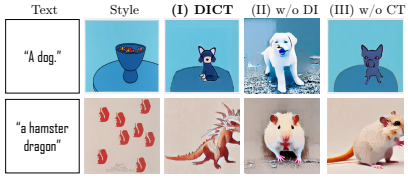
For super-resolution and deblurring, we evaluate PSNR [1], SSIM [18], and LPIPS [28]. PSNR and SSIM measure pixel-level fidelity and structural similarity, respectively, while LPIPS quantifies perceptual distance. In Tab. 2(b), DICT attains the highest PSNR and the lowest LPIPS; its SSIM is slightly below FPS-SMC, which exhibits a much higher LPIPS. In Tab. 2(c), DICT achieves the highest SSIM and the lowest LPIPS, with PSNR marginally below DCDP.

Overall, DICT integrates Data Injection and Contrastive Trajectory Refinement to achieve high-fidelity and precise conditional image generation.

4.3 Ablation Study

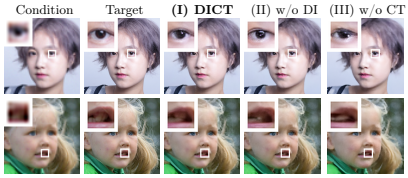
To provide a more comprehensive validation of our framework, we present ablation studies on Data Injection and Contrastive Trajectory Refinement. Additional ablation results are provided in Sec. D of the Supplementary Materials.

Data Injection. We conduct an ablation study on Data Injection to evaluate its effectiveness. As illustrated in columns (I) and (II) of the left figure in Fig. 4(a), removing data conditioning causes the stylized results in (II) to exhibit a stronger inherent model bias (e.g., the overly white dog in the first row). Quantitatively, for style transfer, while omitting Data Injection leads to a marginal



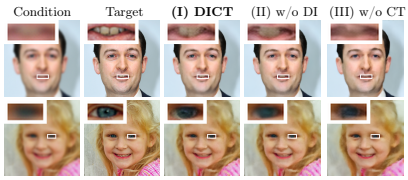
Metric	**(I) DICT**	(II) w/o DI	(III) w/o CT
Text Score ↑	0.2952	0.2943	**0.3008**
Style Loss ↓	**0.6313**	0.8098	0.9201
CLIP Loss ↓	**4.2334**	4.7909	5.2108

(a) Style transfer ablation study.



Metric	**(I) DICT**	(II) w/o DI	(III) w/o CT
PSNR ↑	**28.8600**	28.8155	28.7759
SSIM ↑	**0.8233**	0.8224	0.8148
LPIPS ↓	**0.1573**	0.1574	0.1818

(b) Super-resolution ablation study.



Metric	**(I) DICT**	(II) w/o DI	(III) w/o CT
PSNR ↑	**27.5794**	27.5188	26.8616
SSIM ↑	**0.7736**	0.7711	0.7496
LPIPS ↓	**0.2236**	0.2241	0.2590

(c) Deblurring ablation study.

Fig. 4: Qualitative and quantitative ablation results. "w/o DI" and "w/o CT" denote without Data Injection and Contrastive Trajectory Refinement, respectively.

improvement in Text Score, it results in a substantial surge in Style Loss and CLIP Loss. This suggests that while the model may follow the text prompt slightly more easily without the Data Injection, it fails to capture the essential style, leading to a significant drop in overall fidelity. For super-resolution and deblurring, the absence of Data Injection leads to a consistent degradation across all metrics compared to the full DICT. This is also reflected in column (II) of the qualitative comparisons: specifically, the super-resolution results in the first row of Fig. 4(b) suffer from a loss of fine details (e.g., double eyelids), while the deblurring outputs in the second row of Fig. 4(c) fail to recover clear structures, such as eye regions. These observations further corroborate that Data Injection is indispensable for maintaining high-fidelity conditional alignment.

Contrastive Trajectory Refinement. We evaluate the impact of Contrastive Trajectory Refinement by comparing the full DICT against a variant without it (column (III)). Visually, omitting Contrastive Trajectory Refinement leads to noticeable stylistic discrepancies and structural artifacts. For style transfer, results in column (III) exhibit pronounced color shifts and inconsistent stylization, such as the unnatural hue of the dog in the 1*st* row and the distorted textures of the "hamster dragon" in the 2*nd* row in Fig. 4(a). In low-level vision tasks, the absence of Contrastive Trajectory Refinement compromises local structural fidelity. For super-resolution, outputs in the second row of Fig. 4(b)

Fig. 5: Qualitative DiT-based stylization results.

suffer from inaccurate local details, particularly around the teeth region; for deblurring, the recovery of fine textures is significantly constrained, as seen in the blurred eye regions in the second row of Fig. 4(c). These qualitative shortcomings are reflected in the quantitative results in Fig. 4(b) and Fig. 4(c), where removing Contrastive Trajectory Refinement consistently leads to inferior metrics, thereby validating its essential role for higher-quality generation.

5 Additional Experiments

To demonstrate its backbone-agnostic generalization, we integrate DICT into a Diffusion Transformer (DiT), specifically PixArt [3]. On the demanding text-driven style transfer task, DICT consistently yields high-fidelity results (Fig. 5), confirming its architectural independence.

6 Conclusion

In this paper, we present a unified perspective on conditional image generation. We reveal that traditional loss-guided methods compress high-dimensional conditional cues into a single scalar objective, causing severe information bottlenecks and cumulative sampling errors. To overcome these challenges, we introduce DICT, a versatile inference framework that synergistically combines Data Injection with Contrastive Trajectory Refinement. By directly injecting noise-perturbed conditional signals and denoising them under guidance, DICT preserves the spatial richness lost in scalar-based methods and adaptively distills task-relevant information. Furthermore, its contrastive refinement mechanism ensures coherence between adjacent trajectory steps, progressively improving generation quality. Extensive evaluations across diverse tasks, including style transfer, super-resolution, and deblurring, demonstrate that DICT significantly enhances both fidelity and perceptual quality. Ultimately, DICT offers a robust, training-free, and unified solution for high-quality conditional image generation without requiring task-specific architectures.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant Nos. 2024YFF0907802 and 2024YFF0907803, and the Research Fund for International Scientists of the National Natural Science Foundation of China under Grant No. 72350710798.

References

1. Alkhouri, I., Liang, S., Huang, C.H., Dai, J., Qu, Q., Ravishankar, S., Wang, R.: Sitcom: Step-wise triple-consistent diffusion sampling for inverse problems. arXiv preprint arXiv:2410.04479 (2025)
2. Castillo, A., Kohler, J., Pérez, J.C., Pérez, J.P., Pumarola, A., Ghanem, B., Arbeláez, P., Thabet, A.: Adaptive guidance: Training-free acceleration of conditional diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1962–1970 (2025)
3. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: International conference on learning representations. vol. 2024, pp. 57611–57640 (2024)
4. Deng, Y., He, X., Tang, F., Dong, W.: Z*: Zero-shot style transfer via attention reweighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6934–6944 (2024)
5. Dou, Z., Song, Y.: Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In: The Twelfth International Conference on Learning Representations (2024)
6. Gao, J., Sun, Y., Liu, Y., Tang, Y., Zeng, Y., Qi, D., Chen, K., Zhao, C.: Styleshot: A snapshot on any style. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
7. Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4775–4785 (2024)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
10. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
11. Kim, J., Kim, B.S., Ye, J.C.: Flowdps: Flow-driven posterior sampling for inverse problems. arXiv preprint arXiv:2503.08136 (2025)
12. Lei, M., Song, X., Zhu, B., Wang, H., Zhang, C.: Stylestudio: Text-driven style transfer with selective control of style elements. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23443–23452 (2025)
13. Li, H., Pereira, M.: Solving inverse problems via diffusion optimal control. Advances in Neural Information Processing Systems **37**, 73549–73571 (2024)
14. Li, X., Kwon, S.M., Liang, S., Alkhouri, I.R., Ravishankar, S., Qu, Q.: Decoupled data consistency with diffusion purification for image restoration. arXiv preprint arXiv:2403.06054 (2024)
15. Liu, G., Xia, M., Zhang, Y., Chen, H., Xing, J., Wang, Y., Wang, X., Shan, Y., Yang, Y.: Stylecrafter: Taming artistic video diffusion with reference-augmented adapter learning. ACM Transactions on Graphics (TOG) **43**(6), 1–10 (2024)
16. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778 (2022)
17. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)

18. Patel, M., Wen, S., Metaxas, D.N., Yang, Y.: Steering rectified flow models in the vector field for controlled image generation. arXiv preprint arXiv:2412.00100 (2024)
19. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
20. Phillips, F., Mackintosh, B.: Wiki art gallery, inc.: A case for critical thinking. Issues in Accounting Education **26**(3), 593–608 (2011)
21. Qi, T., Fang, S., Wu, Y., Xie, H., Liu, J., Chen, L., He, Q., Zhang, Y.: Deadiff: An efficient stylization diffusion model with disentangled representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8693–8702 (2024)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
23. Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A., Shakkottai, S.: Solving linear inverse problems provably via posterior sampling with latent diffusion models. Advances in Neural Information Processing Systems **36**, 49960–49990 (2023)
24. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)
25. Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., Chen, A.: Instantstyle: Free lunch towards style-preserving in text-to-image generation. arXiv preprint arXiv:2404.02733 (2024)
26. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25456–25467 (2024)
27. Xing, P., Wang, H., Sun, Y., Wang, Q., Bai, X., Ai, H., Huang, R., Li, Z.: Csgo: Content-style composition in text-to-image generation. arXiv preprint arXiv:2408.16766 (2024)
28. Xu, T., Cai, X., Zhang, X., Ge, X., He, D., Sun, M., Liu, J., Zhang, Y.Q., Li, J., Wang, Y.: Rethinking diffusion posterior sampling: From conditional score estimator to maximizing a posterior. arXiv preprint arXiv:2501.18913 (2025)
29. Ye, H., Lin, H., Han, J., Xu, M., Liu, S., Liang, Y., Ma, J., Zou, J.Y., Ermon, S.: Tfg: Unified training-free guidance for diffusion models. Advances in Neural Information Processing Systems **37**, 22370–22417 (2024)
30. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23174–23184 (2023)
31. Yue, Z., Liao, K., Loy, C.C.: Arbitrary-steps image super-resolution via diffusion inversion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23153–23163 (2025)
32. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)