Discrete Diffusion Bridges for Spatiotemporally Aligned Image Translation and Generation

Xing Xie[1,2], Jiawei Liu[1], Shijun Zhou[1,2], Huijie Fan[1⋆], Zhi Han[1], Yandong Tang[1], and Liangqiong Qu[3⋆]

[1] State Key Laboratory of Robotics and Intelligent Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] School of Computing and Data Science, The University of Hong Kong, Hong Kong
{xiexing, liujiawei, zhoushijun, fanhuijie, hanzhi, ytang}@sia.cn, liangqqu@hku.hk

Abstract. We propose Discrete Diffusion Bridges (DDB), a novel framework designed to resolve the fundamental spatiotemporal misalignment of standard discrete diffusion in image translation and generation. By corrupting data into a pure mask state via a random schedule, the conventional forward process induces a twofold misalignment: spatially, this pure-mask destination entirely discards the rich structural priors of the source image; temporally, the random masking order inherently contradicts the "easy-first, hard-last" decoding mechanism used during inference. To address this, DDB constructs a direct and efficient trajectory between domains. Spatially, we introduce a hybrid absorption mechanism that redefines the absorbing state to a stochastic mixture of mask and source tokens, effectively injecting source prior as spatial anchors into the latent space. Temporally, we design an information-guided noise schedule that quantifies semantic variation to prioritize the corruption of high-information regions at earlier timesteps. This ensures the model learns to resolve difficult semantic changes using robust context from invariant regions. Extensive experiments validate the versatility and robustness of our framework across diverse generative paradigms. DDB effectively balances edit alignment with structural fidelity across both text-guided semantic manipulation and pure structural image translation, while inherently complementing text-to-image generation and guaranteeing robust high-quality decoding under extremely low sampling steps. Code and models are available at https://github.com/HKU-HealthAI/DDB.

Keywords: Discrete Diffusion Model · Image Translation · Image Generation

1 Introduction

Discrete diffusion models (DDMs) [30,39] have recently emerged as a compelling paradigm in generative AI. Unlike the continuous diffusion models that rely on

⋆ Corresponding author.

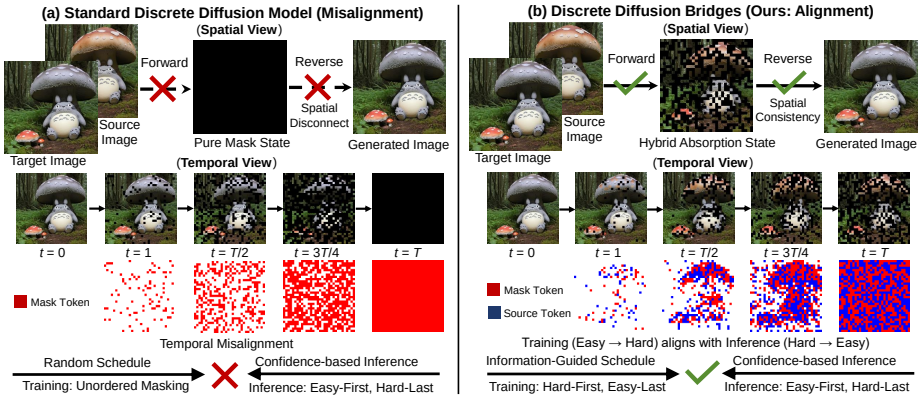

Fig. 1: Comparison between Standard Discrete Diffusion and the proposed Discrete Diffusion Bridges (DDB). (a) Standard models suffer from spatiotemporal misalignment: **spatially** discarding structural priors via a pure mask state, and **temporally** employing a random training schedule that contradicts the "easy-first, hard-last" inference. (b) DDB resolves these discrepancies. **Spatially**, our Hybrid Absorption mechanism injects source tokens into the mask state as spatial anchors, maintaining structural consistency. **Temporally**, our Information-Guided Schedule prioritizes masking complex regions early, aligning the training and inference trajectories.

Gaussian noise corruption, DDMs operate in a compressed latent space using absorbing state formulations [1, 40]. They employ a discrete masking and prediction paradigm: the forward process absorbs the target data into a pure mask state, while the reverse process progressively recovers the original content, as shown in Fig 1(a). This discrete nature allows the model to learn the conditional probability of corrupted tokens, enabling high-fidelity generation through iterative refinement. The success of this paradigm has rapidly extended to complex tasks [61, 63], demonstrating remarkable potential in capturing long-range dependencies and high-frequency details.

However, directly applying the standard discrete diffusion paradigm to image translation and generation tasks reveals a fundamental limitation, which we identify as spatiotemporal misalignment, as shown in Fig 1(a). Spatially, standard models treat image generation as a "void-to-image" reconstruction process. The forward process blindly absorbs the data into a generic, all-mask state (a pure absorbing state), ignoring the explicit structural priors provided by the source image. This forces the model to hallucinate content from scratch rather than transforming the existing semantic layout, leading to a loss of fidelity in identity-preserving tasks. Temporally, a critical discrepancy exists between the training and inference. During training, tokens are masked via a random schedule, treating complex semantic regions (e.g., edited subjects) and simple background regions equally. Conversely, during inference, DDMs typically employ a confidence-based sampling strategy, where easy tokens are determined first and hard tokens last. This contradiction between an unordered training curriculum

and a difficulty-aware inference path prevents the model from effectively learning the transformation logic for complex translation tasks.

To bridge these gaps, we propose Discrete Diffusion Bridges, a novel framework that reformulates image translation and generation as a spatiotemporally consistent trajectory. Our core insight is to redefine the "noise" in discrete diffusion not as an absence of information, but as a guided transition state, as shown in Fig 1(b). Specifically, we introduce two synergistic designs. First, to resolve spatial disconnect, we propose hybrid absorption. Instead of driving forward process toward a pure mask, we construct a hybrid absorption space, where the absorbing state is a probabilistic mixture of mask tokens and source tokens. This mechanism effectively injects source image as a spatial anchor directly into diffusion latent space, ensuring generation is grounded in the source context. Second, to resolve temporal disconnect, we devise an information-guided noise schedule. By measuring the translation difficulty between the source and target, we construct a robust stochastic masking curriculum that prioritizes the corruption of high-information regions at earlier timesteps. This ensures that the training process mirrors the "easy-first, hard-last" dynamics of inference, forcing model to tackle difficult semantic changes only when sufficient context is available.

Extensive experiments demonstrate that DDB establishes a highly efficient and unified bridge across diverse generative paradigms. Specifically, in Image-to-Image (I2I) translation, DDB achieves a state-of-the-art balance between edit alignment and source context preservation. Whether executing precise semantic manipulation (e.g., image editing, style transfer) or performing structural mapping (e.g., all-in-one restoration, modality translation), DDB consistently delivers high-fidelity outputs. Furthermore, DDB demonstrates highly competitive performance in Text-to-Image (T2I) generation, proving that our spatiotemporal alignment mechanism naturally complements universal generative priors. Overall, these results confirm the broad applicability and robustness of our framework. Driven by the spatiotemporal trajectory, DDB effectively balances generation diversity and structural fidelity, while maintaining excellent quality at extremely low sampling steps. Our contributions are summarized as follows:

(i) We identify the spatiotemporal misalignment in applying discrete diffusion to image translation and generation, and propose Discrete Diffusion Bridges, a unified framework that harmonizes state space and generative trajectory.

(ii) We introduce Hybrid Absorption, a novel transition mechanism that replaces the standard pure-mask absorbing state with a source-aware hybrid state, robustly preserving spatial structure.

(iii) We design an Information-Guided Noise Schedule, which decouples the stochastic nature of the noise state from a deterministic, difficulty-aware order, aligning training objectives with inference process.

(iv) Extensive experiments validate the versatility of DDB across diverse generative paradigms. It effectively balances edit alignment with structural fidelity in I2I translation tasks, inherently complements T2I generation, and maintains excellent quality under limited sampling steps.

2 Related Work

2.1 Discrete Diffusion Modeling

Discrete diffusion models [1,39] have emerged as a powerful generative paradigm, offering a highly parallelizable alternative to autoregressive models. This approach traces back to masked language modeling in natural language processing, like BERT [9], which reconstructs obscured tokens using bidirectional context. MaskGIT [6] and MUSE [5] adapted this to the visual domain via iterative token refinement for parallel decoding. Modern masked diffusion models [3,35,69] formulate this as a discrete-state Markov chain, corrupting data to an absorbing state and utilizing bidirectional attention for parallel recovery. Recently, this paradigm has scaled to billion-parameter language models [28,42,62], achieving performance comparable to advanced AR models at the billion-parameter scale. It has also driven the development of discrete multimodal large language models [2,27,63,65,70]. Notably, LaViDa-O [22] introduces an Elastic-MoT architecture and stratified sampling to efficiently balance multimodal tasks. Muddit [41] adopts a pure diffusion transformer for high-quality, unified generation.

2.2 Image Translation and Generation

Continuous diffusion models [15,31,36] have achieved remarkable success across various image translation [7,17,32,49,51] and generation tasks [4,10,11,16,58]. Early models like Palette [37] and SR3 [38] condition on input images to map pure noise to clean targets. To enhance cross-domain mapping, subsequent works such as DDBM [68], RDDM [25], DDIB [43] and BBDM [21] establish direct trajectories between source and target domains utilizing residuals and diffusion bridges. Inspired by large language models, recent research explores the autoregressive paradigm [12,34] for visual tasks. Several studies [24,26,44,46,59] attempt to unify text and image modeling within a single LLM framework, boosting performance via tokenizer optimization [44], early fusion modeling [46], and flexible resolution adaptation [24]. However, their strict left-to-right decoding intrinsically suffers from slow inference speeds and struggles to maintain global spatial coherence. To overcome these AR limitations, discrete diffusion models have emerged as a promising solution for unified generation. Frameworks like UniDisc [45] and Muddit [41] refine tokens iteratively, enabling rapid, parallel decoding. Lumina-DiMOO [60] leverages structural tokens to natively support arbitrary-resolution generation.

3 Method

3.1 Motivation: Misalignment in Standard DDM

We first revisit the standard discrete diffusion models (DDMs) to formalize both their forward and reverse processes, thereby illuminating their inherent limitations in image translation and generation. Let $\mathcal{K} = \{1, \ldots, K\}$ be a discrete

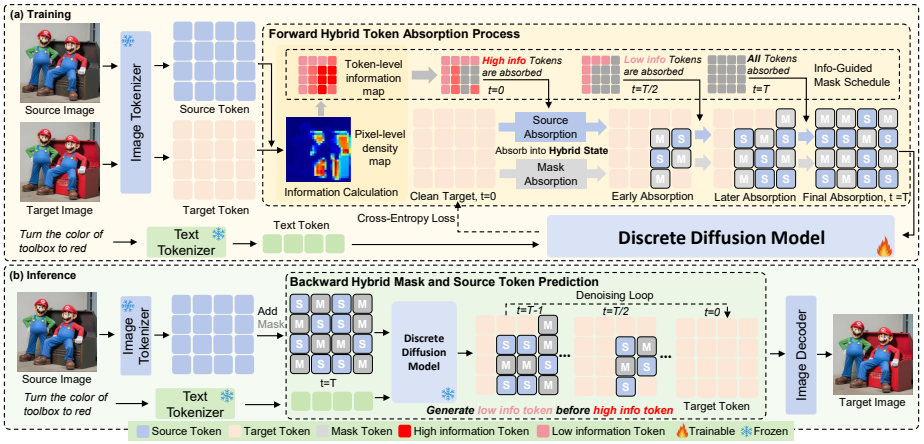

Fig. 2: Overview of the Discrete Diffusion Bridges (DDB) framework. (a) Training: The forward process employs Hybrid Absorption to corrupt target tokens into a source-mask mixture, establishing explicit spatial anchors. It employs an Information-Guided Schedule: a pixel-level density map is computed and downsampled into a token-level information map, prioritizing the early absorption of high-information tokens. (b) Inference: Starting from the hybrid state, confidence-based decoding resolves low-information regions before complex edits, achieving spatiotemporal alignment with training process.

codebook vocabulary of size K. A target image can be tokenized into a discrete sequence $\mathbf{z}_0 \in \mathcal{K}^N$, where N is the sequence length.

Forward Process. Standard DDMs define the forward process $q(\mathbf{z}_t|\mathbf{z}_0)$ as a gradual corruption that progressively replaces input tokens with a special mask token [M]. Once a token is masked, it remains in that state throughout the remainder of the process, making [M] an absorbing state. At timestep $t \in [0, T]$, the forward transition for the i-th token is defined as:

$$q(z_t^{(i)}|z_0^{(i)}) = \text{Cat}(z_t^{(i)}; (1-\gamma(t))z_0^{(i)} + \gamma(t)m). \tag{1}$$

$\text{Cat}(\cdot)$ denotes a categorical distribution. Where $z_0^{(i)}$ and m are the one-hot vectors corresponding to original target token and [M] token, respectively. $\gamma(t) \in [0, 1]$ is a monotonically increasing noise schedule such that $\gamma(0) \approx 0$ and $\gamma(T) = 1$. Crucially, this transition probability $\gamma(t)$ is applied uniformly across all spatial locations, meaning the masking process is entirely unordered and random.

Reverse Process. During inference, generation proceeds via an iterative refinement from the pure mask state $\mathbf{z}_T = [\text{M}]^N$ to $\mathbf{z}_0$. Tokens unmasked in previous steps are carried over unchanged. At each reverse step $t \to t-1$, the network predicts the categorical distribution $p_\theta(\mathbf{z}_0|\mathbf{z}_t, \mathbf{c})$ given a condition $\mathbf{c}$. For each currently masked token position i (where $z_t^{(i)} = [\text{M}]$), the most probable candidate $\hat{z}_0^{(i)}$ and its corresponding confidence score $s_t^{(i)}$ are computed:

$$\hat{z}_0^{(i)} = \arg\max_{k \in \mathcal{K}} p_\theta(z_0^{(i)} = k|\mathbf{z}_t, \mathbf{c}), \quad s_t^{(i)} = \max_{k \in \mathcal{K}} p_\theta(z_0^{(i)} = k|\mathbf{z}_t, \mathbf{c}). \tag{2}$$

The model then ranks the newly predicted candidates by their confidence scores and retains a specific proportion of the most confident tokens as governed by the noise schedule. The reverse transition is thus formulated as:

$$z_{t-1}^{(i)} = \begin{cases} z_t^{(i)} & \text{if } z_t^{(i)} \neq [\mathrm{M}] \\ \hat{z}_0^{(i)} & \text{if } z_t^{(i)} = [\mathrm{M}] \text{ and } s_t^{(i)} \geq \tau_t \\ [\mathrm{M}] & \text{otherwise} \end{cases} \tag{3}$$

where τ_t is a dynamic confidence threshold implicitly determined by the schedule $\gamma(t-1)$ to regulate the decoding pace.

Spatiotemporal Misalignment. When applied to conditional image translation and generation, the discrepancy between Eq. 1 and Eq. 3 exposes a fundamental spatiotemporal misalignment. **Spatially**, the forward process strictly drives the data into a pure mask state ($\mathbf{z}_T = [\mathrm{M}]^N$). This completely severs the spatial correspondence between the target and the explicit source priors, forcing the model to hallucinate invariant structures from a semantic void rather than transforming existing source image features. **Temporally**, the generation trajectory in Eq. 3 employs a confidence-driven progressive decoding mechanism. It naturally tends to resolve deterministic features before generating complex and ambiguous semantic details. Conversely, the forward process in Eq. 1 applies uniform random masking, forcing the model to learn the corruption of simple and complex regions without any difficulty-aware curriculum. This stark contradiction between random training trajectory and deterministic, confidence-based inference dynamics makes learning the transition mapping highly inefficient.

3.2 Overview of DDB

To overcome these misalignments, we propose **Discrete Diffusion Bridges** (see Fig. 2), a novel framework that reconceptualizes the generative trajectory. DDB constructs a direct, efficient bridge connecting the target domain to the source domain. The framework comprises two core phases: (1) A Forward Process (Sec. 3.3) that builds the bridge. It introduces a Hybrid Absorption mechanism to establish spatial anchors by injecting source priors into the absorbing state, and an Information-Guided Schedule to map difficulty-aware temporal path. (2) A Reverse Process (Sec. 3.4) that efficiently traverses this bridge. Guided by the spatial anchors, it resolves simple invariant regions before complex semantic edits, naturally matching the training trajectory.

3.3 Forward Process: Constructing the Bridge

Spatial Bridge Anchors: Hybrid Absorption. Standard discrete diffusion replaces all corrupted tokens with [M], which mathematically corresponds to driving the generative state towards a pure mask vector m. This mechanism entirely severs the spatial connection between the source and target domains. To construct a direct bridge, we propose a Hybrid Absorption mechanism that injects source priors into the latent space to establish robust spatial anchors.

(1) Hybrid State Space. We define the absorbing state as a source-conditioned hybrid distribution. For a given source image token $y^{(i)}$ represented as a one-hot vector, we define the hybrid absorbing state $h^{(i)}$ as a stochastic mixture of the source prior and the mask:

$$h^{(i)} = \lambda y^{(i)} + (1 - \lambda)m \tag{4}$$

where $\lambda \in [0, 1]$ is the source injection ratio. When $\lambda > 0$, the retained source tokens successfully serve as explicit "spatial anchors" to bridge the two domains. When $\lambda = 0$, the framework can naturally degrade to a standard discrete diffusion model without other modification.

(2) Hybrid Forward Transition. The forward transition $q(\mathbf{z}_t|\mathbf{z}_0, \mathbf{y})$ is consequently conditioned on both the target and the source. Let $\mathbf{m}_t \in \{0, 1\}^N$ be the binary mask sequence at timestep t (which will be determined by our information-guided schedule), where $m_t^{(i)} = 0$ indicates that the i-th token is selected for corruption. The transition for our discrete bridge is formulated as a categorical distribution:

$$q(z_t^{(i)}|z_0^{(i)}, y^{(i)}, m_t^{(i)}) = \mathrm{Cat}\left(z_t^{(i)}; m_t^{(i)} z_0^{(i)} + (1 - m_t^{(i)})h^{(i)}\right) \tag{5}$$

Here, $\mathrm{Cat}(\cdot)$ denotes a categorical distribution. This formulation ensures endpoint $\mathbf{z}_T$ retains rich structural priors.

Temporal Bridge Trajectory: Information-Guided Mask Schedule. To explicitly align the training curriculum with inference dynamics, we introduce an Information-Guided Schedule. By quantifying the translation difficulty (information density) of each region, we construct a temporal trajectory that prioritizes corrupting high-information tokens at earlier timesteps.

(1) Information Metric Formulation. We define a token-level information map $\mathcal{I} \in \mathbb{R}^N$ to guide mask schedule, derived from a pixel-level density map $\mathcal{M} \in \mathbb{R}^{H_{px} \times W_{px}}$. For translation tasks with source images (e.g., image editing), $\mathcal{M}$ is the absolute pixel-wise difference between the target $\mathbf{X}_0$ and source $\mathbf{X}_{src}$ across C channels, computed as $\mathcal{M} = \frac{1}{C}\sum_{c=1}^{C} |\mathbf{X}_{0,c} - \mathbf{X}_{src,c}|$. For generation tasks without source images (e.g., text-to-image), we evaluate structural complexity using the local pixel variance of the grayscale target $\mathbf{X}_{0,g}$ within a $k \times k$ window, formulated as $\mathcal{M} = \mathrm{AvgPool}_{k\times k}(\mathbf{X}_{0,g}^2) - (\mathrm{AvgPool}_{k\times k}(\mathbf{X}_{0,g}))^2$. The pixel map $\mathcal{M}$ is then downsampled via average pooling to match the tokenizer's spatial stride, yielding the flattened token-level map $\mathcal{I}$.

(2) Robust Stochastic Masking. Strictly sorting tokens by $\mathcal{I}$ yields a deterministic trajectory but restricts exposure to diverse semantic contexts, risking overfitting. To preserve contextual robustness, we propose a stochastic masking strategy. We normalize $\mathcal{I}$ into a probability distribution $P_{info}^{(i)} = \mathcal{I}^{(i)} / \sum_{j=1}^{N} \mathcal{I}^{(j)}$. The final sampling probability $P_{final}^{(i)}$ for the i-th token is formulated as a weighted combination of $P_{info}^{(i)}$ and a uniform distribution $P_{uniform}^{(i)} = 1/N$:

$$P_{final}^{(i)} = (1 - \rho) \cdot P_{info}^{(i)} + \rho \cdot P_{uniform}^{(i)}, \tag{6}$$

where $\rho \in [0, 1]$ regulates the stochasticity. At timestep t, we sample exactly $n_t = \lfloor N \cdot \gamma(t) \rfloor$ mask indices based on P_{final}. This multinomial sampling statistically prioritizes high-information regions, maintaining the "hard-first" temporal bridge trajectory while intrinsically injecting contextual diversity via ρ.

3.4 Reverse Process: Inference along the Bridge

During inference, the reverse generation precisely traces back along the constructed bridge via an iterative decoding and anchor-resampling strategy. The initial state $\mathbf{z}_T$ is instantiated by independently sampling each token $z_T^{(i)}$ from the hybrid absorbing distribution $h^{(i)}$.

Crucially, tokens that have been confidently resolved in previous steps are carried over unchanged. At each reverse step $t \to t-1$, the network predicts the categorical distribution $p_\theta(\mathbf{z}_0|\mathbf{z}_t, \mathbf{y}, \mathbf{c})$ given the source anchors $\mathbf{y}$ and condition $\mathbf{c}$. For each currently unresolved token position i, we determine the most probable candidate $\hat{z}_0^{(i)}$ and its confidence score $s_t^{(i)}$:

$$\hat{z}_0^{(i)} = \arg\max_{k \in \mathcal{K}} p_\theta(z_0^{(i)} = k|\mathbf{z}_t, \mathbf{y}, \mathbf{c}), \quad s_t^{(i)} = \max_{k \in \mathcal{K}} p_\theta(z_0^{(i)} = k|\mathbf{z}_t, \mathbf{y}, \mathbf{c}) \tag{7}$$

To align with the training trajectory, the model ranks these newly predicted candidates by their confidence scores. We retain a specific proportion of the most confident tokens as governed by the discrete schedule. Instead of reverting the uncertain regions to a static state or a pure semantic void, we dynamically resample them from the hybrid distribution to perfectly match the stochastic source injection used during training. The transition is formulated as:

$$z_{t-1}^{(i)} = \begin{cases} z_t^{(i)} & \text{if token } i \text{ is already unmasked} \\ \hat{z}_0^{(i)} & \text{if token } i \text{ is newly unmasked } (s_t^{(i)} \geq \tau_{t-1}) \\ \tilde{h}^{(i)} & \text{otherwise} \quad (\text{where } \tilde{h}^{(i)} \sim h^{(i)}) \end{cases} \tag{8}$$

where τ_{t-1} is the dynamic confidence threshold implicitly determined by the schedule $\gamma(t-1)$, and $\tilde{h}^{(i)}$ is a newly drawn sample from the hybrid absorbing distribution $h^{(i)} = \lambda y^{(i)} + (1-\lambda)m$. This ensures undecoded regions receive continuous guidance from dynamic spatial anchors.

Training Objective. Let $\mathbf{c}$ be the conditioning signal (e.g., text prompt) and $\mathbf{y}$ be the source image tokens. The model is trained to predict the original target tokens $\mathbf{z}_0$ from the hybrid corrupted state $\mathbf{z}_t$. We optimize the model using a standard cross-entropy loss applied specifically to the absorbed regions:

$$\mathcal{L}_{DDB} = \mathbb{E}_{t, \mathbf{z}_0, \mathbf{y}, \mathbf{c}} \left[\sum_{i=1}^{N} -(1 - m_t^{(i)}) \log p_\theta(z_0^{(i)}|\mathbf{z}_t, \mathbf{y}, \mathbf{c}) \right] \tag{9}$$

4 Experiments Analysis and Results

In the experiments, we comprehensively evaluate the effectiveness of DDB across diverse translation and generation paradigms. First, we validate the superiority of DDB framework in Image-to-Image (I2I) translation across two distinct paradigms: (1) instruction-based translation (image editing, style transfer and sub-driven generation), which validates model's capability for precise **semantic manipulation** while preserving invariant contexts; and (2) pure image translation (all-in-one restoration, modality translation and super resolution), which verifies its capability for robust **structural mapping** without textual cues. Next, we show that DDB's advantages extend to unanchored Text-to-Image (T2I) generation, confirming that our spatiotemporal alignment mechanism inherently complements universal **generative priors**. We then conduct comprehensive component analyses to verify the efficacy of our core designs, including hybrid absorption mechanism and information-guided mask schedule. Finally, we investigate DDB's inference efficiency and robustness under few-step settings.

4.1 Experimental Setup

Datasets. For TI2I translation tasks, we use OmniEdit [54] dataset for image editing and style transfer tasks, and Graph-200K [23] dataset for subject-driven generation task. For pure I2I translation tasks, we use CDD-11 [13] dataset for All-in-One restoration task, SynthRAD [47] dataset for CT-to-MRI and MRI-to-CT modality translation task, and IXI [18] dataset for super resolution task. For T2I generation, we use MIMIC-CXR [19] for medical report-to-image generation.

Evaluation Metrics. For TI2I tasks, we evaluate background preservation via DINO [29] and edit semantic alignment via CLIP-T [33], and overall editing quality using the LLM-based Edit Score [56]. For pure I2I tasks, we assess pixel-level fidelity using PSNR and SSIM [52], and perceptual similarity using LPIPS [66]. For the T2I task, we use FID [14] and MS-SSIM [53] to evaluate image fidelity, and CLIP-Score [67] to evaluate text-image alignment.

Implementation Details. We adopt Lumina-DiMOO [60] as the backbone of our DDB framework, which is an advanced discrete diffusion model with 8B parameters. This backbone uses VQ-VAE [48] for latent encoding. Additional training details and all other experimental settings are provided in Appendix.

4.2 Performance on I2I Translation Tasks

To comprehensively evaluate the capability of DDB on I2I translation tasks, we design experiments from two primary perspectives: instruction-based translation (TI2I) for evaluating complex semantic manipulation, and pure image translation (pure I2I) for assessing robust structural mapping.

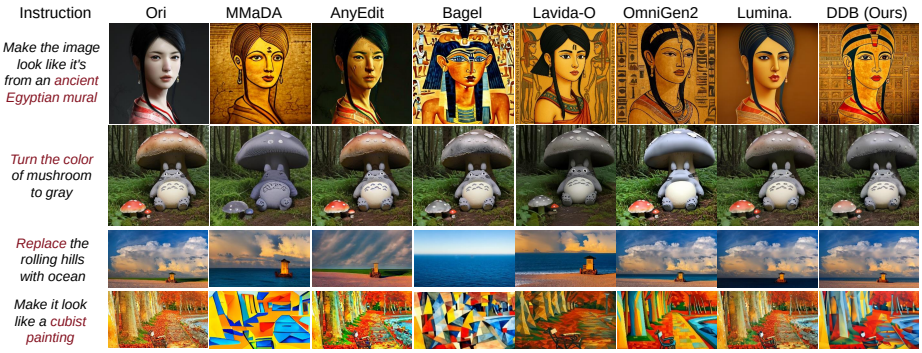

Fig. 3: Visual comparison of different methods and DDB for image editing and style transfer tasks on the OmniEdit dataset.

Table 1: Performance comparisons of Image Editing and Style Transfer tasks on the OmniEdit [54] benchmark, and Sub-driven generation tasks on the Graph-200K [23] benchmark. Best results are highlighted in **red** and the second-best results are blue.

Method	Image Editing			Style Transfer		Sub-Driven Generation		
	Edit Score ↑	CLIP-T↑	DINO ↑	Edit Score ↑	CLIP-T ↑	CLIP-I↑	CLIP-T ↑	DINO ↑
MMaDA [61]	-0.233	0.231	0.643	-0.188	0.244	-	-	-
OmniGen [57]	0.129	0.242	0.636	0.587	0.246	0.828	0.343	0.710
AnyEdit [64]	0.110	0.239	0.762	0.319	0.228	0.736	0.319	0.542
Bagel [8]	0.694	0.233	**0.836**	0.128	0.242	**0.858**	0.348	0.745
OmniGen2 [55]	**0.752**	**0.240**	0.784	0.628	0.240	0.843	0.344	0.739
Lavida-O [22]	0.619	**0.242**	0.681	**1.233**	0.246	0.856	**0.359**	**0.774**
Lumina. [60]	0.734	0.234	0.766	0.958	**0.246**	0.843	0.344	0.767
DDB (Ours)	**0.779**	0.236	**0.790**	**1.077**	**0.250**	**0.867**	**0.349**	**0.789**

Semantic Manipulation: DDB for TI2I Translation. To assess DDB's capability in text-guided semantic manipulation, we evaluate three distinct subtasks: modifying specific regions (Image Editing), globally transforming visual appearance (Style Transfer), and contextualizing background-free target (Subject Driven Generation). As reported in Table 1, standard discrete diffusion models struggle to balance edit extent (CLIP-T) with source preservation (DINO). In contrast, DDB achieves the highest overall Edit Score by excelling in both metrics. This quantitative superiority reflects two core capabilities of our model: the information-guided trajectory enables precise spatial localization for regional edits, while the hybrid absorption mechanism ensures strict identity retention during aggressive global transformations. Visually (Fig. 3), DDB strictly preserves unedited background regions and seamlessly blends edited subjects into the scene with high fidelity, mitigating the unintended background alterations commonly observed in baseline methods.

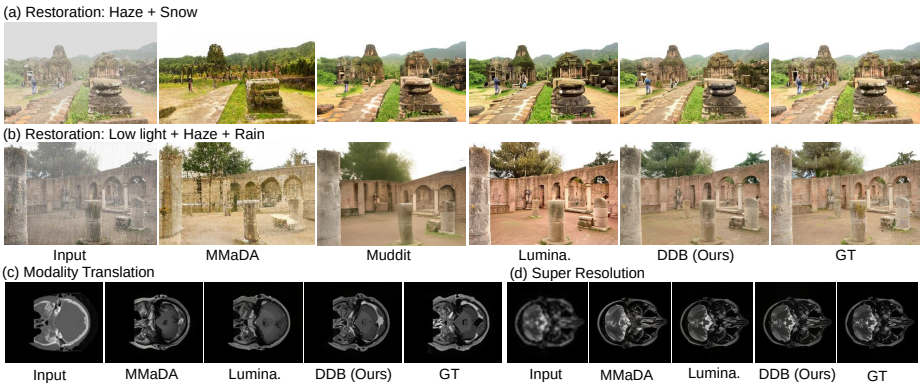

Fig. 4: Visual comparison of different methods for all-in-one restoration, modality translation and super resolution tasks on the CDD-11, SynthRAD and IXI datasets.

Table 2: Evaluate All-in-One Restoration, Modality translation, and Super Resolution tasks on CDD-11 [13], SynthRAD [47], and IXI [18] datasets, respectively.

Method	All-in-One Restoration			Modality Translation			Super Resolution		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
MmaDA [61]	17.92	0.595	0.187	22.24	0.742	0.110	22.89	0.704	0.087
Muddit [41]	20.29	0.608	0.140	21.50	0.585	0.164	23.75	0.648	0.109
Lumina. [60]	21.96	0.751	0.092	23.05	0.732	0.095	25.15	0.719	0.066
DDB (Ours)	**22.61**	**0.759**	**0.081**	**23.36**	**0.750**	**0.083**	**25.46**	**0.733**	**0.056**

Structural Mapping: DDB for Pure I2I Translation. To verify DDB's capacity to capture complex spatial correspondences relying solely on visual priors, we evaluate the model on pure I2I tasks without explicit textual guidance. We deliberately design experiments across three distinct tasks to validate different dimensions of the model's structural mapping capabilities: (1) *All-in-One Restoration* tests the model's multi-task versatility in handling diverse and unpredictable degradations within a single unified framework; (2) *Modality Translation* evaluates its capability to bridge severe cross-domain gaps while maintaining strict anatomical consistency; and (3) *Super Resolution* assesses its capability to recover high-frequency, fine-grained details from severe structural information loss. As shown in Table 2, DDB consistently outperforms baselines, achieving significantly higher PSNR and SSIM, alongside lower LPIPS scores across all three settings. These results confirm that the explicit spatial anchors injected via our hybrid absorption process successfully empower the model to mitigate severe domain shifts and execute highly accurate structural mapping. As shown in Fig. 4, DDB successfully restores complex textures from degraded images, and generates highly realistic target modalities with sharp anatomical boundaries in CT-to-MRI translation, proving the robustness and precision of our spatiotemporally aligned diffusion trajectory. More results of the I2I task are in Appendix.

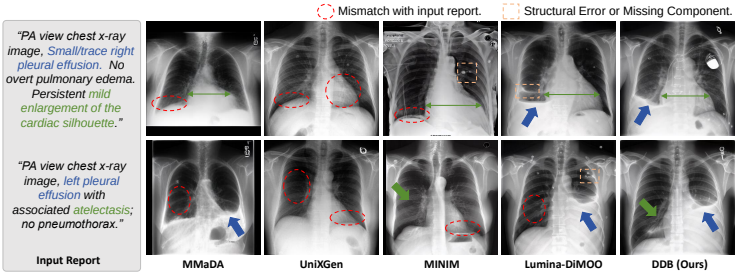

Fig. 5: Visual comparison of different methods and DDB for report-to-image generation task on the MIMIC-CXR dataset.

4.3 Performance on T2I Generation Tasks

To investigate the generation capability of our model without source images as explicit spatial source anchors, we perform a report-to-image generation (T2I) task. This task requires generating coherent X-ray images strictly from complex medical text descriptions. As shown in Table 3, DDB achieves superior FID, MS-SSIM, and CLIP-Score compared to other methods. Visual comparisons in Fig. 5 further show that our generated X-rays exhibit more realistic anatomical features and fewer structural artifacts. This confirms that introducing spatiotemporal alignment does not compromise, but rather enhances, the model's fundamental generative priors.

Table 3: Evaluate Report to Image generation task on MIMIC-CXR [19] datasets.

Method	Report-to-Image		
	FID ↓	MS-SSIM ↑	CLIP-Score ↑
MmaDA [61]	15.25	0.381	0.446
UniXGen [20]	30.75	0.361	0.413
MINIM [50]	15.62	0.317	0.442
Lum.-DiMOO [60]	13.86	0.398	0.441
DDB (Ours)	**9.81**	**0.413**	**0.451**

4.4 Component Analysis of DDB Framework

To better understand the core design principles of DDB and build an effective model, we analyze how different components affect the effectiveness of our discrete diffusion process. We focus on the following key questions:

- Source injection ratio: How does the hybrid injection ratio balance structural preservation against editing alignment during forward process? (Table 4)
- Source injection mechanism: Which spatial assignment strategy (stochastic mixing or deterministic allocation) yields the most robust hybrid absorbing state? (Table 5)
- Information metrics: How do varying definitions of information impact scheduling effectiveness across different tasks? (Table 6)

Source injection ratio. We analyze the impact of the source token injection ratio during the forward process. As shown in Table 4, an injection ratio of

Table 4: Impact of source token injection ratios on the OmniEdit benchmark, comparing models trained with a fixed ratio against trained with a random ratio ($r \in [0, 1]$).

Ratio	Train on specified ratio			Train on random ratio		
	Edit Score ↑	CLIP-T↑	DINO ↑	Edit Score ↑	CLIP-T↑	DINO ↑
0.0	0.743	0.235	0.755	0.760	0.235	0.765
0.2	0.745	0.233	0.761	0.770	0.234	0.774
0.5	0.757	0.235	0.778	**0.779**	**0.236**	**0.790**
0.8	0.756	0.234	0.774	0.768	0.233	0.785
1.0	0.750	0.233	0.781	0.764	0.232	0.788

Table 5: The impact of varying source injection mechanisms. We compare deterministic allocation (M1: masking high-information regions; M2: the inverse) against stochastic proportional mixing (M3).

Metric	Image Editing			Modality Translation			Super Resolution		
	Edit Score ↑	CLIP-T↑	DINO ↑	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o inj.	0.743	0.235	0.755	23.15	0.741	0.089	25.19	0.723	0.059
M1	0.695	0.231	0.783	21.96	0.714	0.102	23.59	0.688	0.073
M2	0.747	0.234	0.781	23.11	0.746	0.088	25.10	0.725	0.061
M3	**0.779**	**0.236**	**0.790**	**23.36**	**0.750**	**0.083**	**25.46**	**0.733**	**0.056**

0.5 combined with random-ratio training yields superior performance. We draw three key conclusions. First, a larger injection ratio yields better source and background preservation (higher DINO), whereas a smaller ratio favors stronger edit semantic alignment (higher CLIP-T). Second, an injection ratio of 0.5, where noise and source tokens are evenly mixed, achieves the optimal balance between CLIP-T and DINO, resulting in the highest overall Edit Score. This validates the efficacy of hybrid noising. Furthermore, training with a random ratio significantly outperforms any fixed ratio configuration. This occurs because random ratio training exposes the model to diverse contextual combinations, thereby enhancing its robustness and generalization capabilities during inference. This leads to a key insight: **Takeaway 1.** *An optimal balance of source and mask tokens (ratio around 0.5) coupled with random-ratio training maximizes both structural preservation and editing flexibility.*

Source injection mechanism. We evaluate three injection mechanisms for corrupted regions to determine the best strategy for mixing source and mask tokens: (1) Deterministically injecting mask tokens to high-information regions and source tokens to low-information regions; (2) The inverse of (1); (3) Randomly assigning source or mask tokens based on a fixed ratio. Empirical results in Table 5 demonstrate that the stochastic proportional mix (M3) achieves the best performance. This happens because deterministic allocation often causes the model to learn fixed mappings, excessively limiting the flexibility of generation. In contrast, stochastic injection forces the network to robustly learn transition mappings under diverse and unpredictable contextual guidance. This leads to a

Table 6: The impact of using different information metrics on image editing, style transfer, and report-to-image generation tasks.

Metric	Image Editing			Style Transfer		Report-to-Image		
	Edit Score ↑	CLIP-T↑	DINO ↑	Edit Score ↑	CLIP-T↑	FID ↓	MS-SSIM ↑	CLIP-Score ↑
w/o infor.	0.755	0.235	0.778	0.982	0.245	14.88	0.397	0.441
Gradient	0.686	0.236	0.764	0.979	0.250	11.50	0.403	0.450
Frequency	0.758	0.235	0.780	1.069	0.247	11.59	0.410	0.449
Variance	0.723	0.236	0.772	0.983	0.249	**9.81**	**0.413**	**0.451**
Image Diff	**0.779**	**0.236**	**0.790**	**1.077**	**0.250**	-	-	-

Table 7: Quantitative evaluation under few-step inference settings.Our DDB maintains robust performance even at extremely low sampling steps.

Step	Lumina-DiMOO			DDB(Ours)		
	Edit Score ↑	CLIP-Acc ↑	DINO ↑	Edit Score ↑	CLIP-Acc ↑	DINO ↑
5 steps	0.541	0.233	0.743	0.606	0.235	0.757
15 steps	0.709	0.235	0.756	0.735	0.235	0.762
30 steps	0.716	0.234	0.758	0.763	0.235	0.769
45 steps	0.721	0.235	0.757	0.776	0.235	0.770
64 steps	0.734	0.234	0.766	**0.779**	**0.236**	**0.790**

key insight: **Takeaway 2.** *Stochastic proportional injection is superior because it ensures contextual robustness and effectively decouples the stochastic spatial state from the deterministic temporal schedule.*

Information metrics. We study the influence of different information metrics used to guide our masking schedule. As shown in Table 6, for translation tasks utilizing explicit source images (e.g., Image Editing, Style Transfer), Image Difference achieves highest performance. Conversely, for unanchored generation tasks lacking source images (e.g., Report-to-Image), Target variance emerges as the optimal metric. This indicates that structural translation relies on relative cross-domain changes, whereas pure generation depends on image complexity to quantify generation difficulty. This leads to a key insight: **Takeaway 3.** *The optimal information metric is task-dependent: latent difference is crucial for anchored image-to-image translation, while target variance is optimal for unanchored text-to-image generation.*

4.5 Performance on Few-Step Inference Setting

We analyze the model's performance retention across varying sampling steps to evaluate its inference efficiency and robustness. Standard discrete diffusion models typically suffer severe performance degradation when inference steps are heavily reduced, primarily due to compounded prediction errors and the lack of explicit guidance. As shown in Table 7, while the baseline model exhibits a significant performance drop at low step counts (5 steps), DDB maintains robust

generation quality and editing accuracy. This demonstrates that our spatiotemporally aligned trajectory effectively simplifies the learning objective, enabling highly efficient and accurate decoding even under accelerated inference settings.

5 Conclusion

We propose Discrete Diffusion Bridges (DDB), a novel framework that resolves the fundamental spatiotemporal misalignment inherent in standard discrete diffusion models. Our work reconceptualizes the discrete generative trajectory from a random, void-to-image process into a direct, efficient bridge between data domains. By introducing hybrid absorption to establish robust spatial source anchors and an information-guided schedule to dictate an optimal, difficulty-aware temporal path, DDB perfectly synchronizes training objectives with inference dynamics. Extensive experiments demonstrate DDB's versatility and superiority across a wide range of tasks. Our work opens new perspectives on optimal trajectory design in discrete latent spaces and provides novel insights into modeling unified multimodal translation and generation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U23A20343), the National Natural Science Foundation of China under Grant (T2594604030, T2594604035, 62306253), the Early Career Fund (27207025), the National Natural Science Foundation of China under Grant (U24A20282), the Guangdong Natural Science Fund-General Program (2024A1515010233), the China Postdoctoral Science Foundation under Grant Number 2025M781669, and the Fundamental Research Project of SIA (2025JC1K05).

References

1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in neural information processing systems **34**, 17981–17993 (2021) 2, 4
2. Bai, J., Lei, Y., Wu, H., Zhu, Y., Li, S., Xin, Y., Li, X., Tao, M., Grover, A., Yang, M.H.: From masks to worlds: A hitchhiker's guide to world models. arXiv preprint arXiv:2510.20668 (2025) 4
3. Bai, J., Ye, T., Chow, W., Song, E., Chen, Q.G., Li, X., Dong, Z., Zhu, L., Yan, S.: Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In: The Thirteenth International Conference on Learning Representations (2024) 4
4. Cao, P., Zhou, F., Song, Q., Yang, L.: Controllable generation with text-to-image diffusion models: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 4
5. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023) 4

6. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11315–11325 (2022) 4
7. Chen, I., Chen, W.T., Liu, Y.W., Chiang, Y.C., Kuo, S.Y., Yang, M.H., et al.: Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17969–17979 (2025) 4
8. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 10
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019) 4
10. Dong, J., Lyu, Q., Liu, B., Wang, X., Liang, W., Zhang, D., Tu, J., Li, H., Zhao, H., Ding, H., Zhang, Y., Han, Z., Sebe, N., Khan, F.S., Khan, S., Shah, M., Torr, P., Yang, M.H., Tao, D.: Learning to model the world: A survey of world models in artificial intelligence. TechRxiv (2026) 4
11. Dong, J., Wang, X., Liang, W., Han, Z., Cao, M., Zhang, D., Zhao, H., Han, Z., Khan, S., Khan, F.S.: Bring your dreams to life: Continual text-to-video customization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 3623–3631 (2026) 4
12. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 4
13. Guo, Y., Gao, Y., Lu, Y., Zhu, H., Liu, R.W., He, S.: Onerestore: A universal restoration framework for composite degradation. In: European conference on computer vision. pp. 255–272. Springer (2024) 9, 11
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 9
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 4
16. Huang, G., Mao, J., Huang, F., Liu, F., Luo, X., Liang, Y., Lu, J., Wang, X., Liu, P., Fu, R., Huang, R., Huang, S.L.: Exposure bias can alleviate itself via directional and frequency rectification in flow matching (2026) 4
17. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 4
18. Information eXtraction from Images (IXI) Project: The IXI Dataset. https://brain-development.org/ixi-dataset/, accessed: 2026-03-01 9, 11
19. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019) 9, 12
20. Lee, H., Lee, D.Y., Kim, W., Kim, J.H., Kim, T., Kim, J., Sunwoo, L., Choi, E.: Vision-language generative model for view-specific chest x-ray generation. arXiv preprint arXiv:2302.12172 (2023) 12

21. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdm: Image-to-image translation with brownian bridge diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 1952–1961 (2023) 4
22. Li, S., Gu, J., Liu, K., Lin, Z., Wei, Z., Grover, A., Kuen, J.: Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. arXiv preprint arXiv:2509.19244 (2025) 4, 10
23. Li, Z.Y., Du, R., Yan, J., Zhuo, L., Li, Z., Gao, P., Ma, Z., Cheng, M.M.: Visualcloze: A universal image generation framework via visual in-context learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18969–18979 (2025) 9, 10
24. Liu, D., Zhao, S., Zhuo, L., Lin, W., Qiao, Y., Li, H., Gao, P.: Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. arXiv preprint arXiv:2408.02657 (2024) 4
25. Liu, J., Wang, Q., Fan, H., Wang, Y., Tang, Y., Qu, L.: Residual denoising diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2773–2783 (2024) 4
26. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26439–26455 (2024) 4
27. Luo, Y., Hu, X., Fan, K., Sun, H., Chen, Z., Xia, B., Zhang, T., Chang, Y., Wang, X.: Reinforcement learning meets masked generative models: Mask-grpo for text-to-image generation. arXiv preprint arXiv:2510.13418 (2025) 4
28. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.R., Li, C.: Large language diffusion models. arXiv preprint arXiv:2502.09992 (2025) 4
29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023) 9
30. Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., Li, C.: Your absorbing discrete diffusion secretly models the conditional distributions of clean data. arXiv preprint arXiv:2406.03736 (2024) 1
31. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 4
32. Qu, L., Tian, J., Han, Z., Tang, Y.: Pixel-wise orthogonal decomposition for color illumination invariant and shadow-free image. Optics express **23**(3), 2220–2239 (2015) 4
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 9
34. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021) 4
35. Rojas, K., Zhu, Y., Zhu, S., Ye, F.X.F., Tao, M.: Diffuse everything: Multimodal diffusion models on arbitrary state spaces. arXiv preprint arXiv:2506.07903 (2025) 4

36. 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) 4
37. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022) 4
38. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4713–4726 (2022) 4
39. Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., Kuleshov, V.: Simple and effective masked diffusion language models. Advances in Neural Information Processing Systems **37**, 130136–130184 (2024) 1, 4
40. Shi, J., Han, K., Wang, Z., Doucet, A., Titsias, M.: Simplified and generalized masked diffusion for discrete data. Advances in neural information processing systems **37**, 103131–103167 (2024) 2
41. Shi, Q., Bai, J., Zhao, Z., Chai, W., Yu, K., Wu, J., Song, S., Tong, Y., Li, X., Li, X., et al.: Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. arXiv preprint arXiv:2505.23606 (2025) 4, 11
42. Song, Y., Zhang, Z., Luo, C., Gao, P., Xia, F., Luo, H., Li, Z., Yang, Y., Yu, H., Qu, X., et al.: Seed diffusion: A large-scale diffusion language model with high-speed inference. arXiv preprint arXiv:2508.02193 (2025) 4
43. Su, X., Song, J., Meng, C., Ermon, S.: Dual diffusion implicit bridges for image-to-image translation. arXiv preprint arXiv:2203.08382 (2022) 4
44. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024) 4
45. Swerdlow, A., Prabhudesai, M., Gandhi, S., Pathak, D., Fragkiadaki, K.: Unified multimodal discrete diffusion. arXiv preprint arXiv:2503.20853 (2025) 4
46. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024) 4
47. Thummerer, A., Van der Bijl, E., Galapon Jr, A., Verhoeff, J.J., Langendijk, J.A., Both, S., van den Berg, C.N.A., Maspero, M.: Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. Medical physics **50**(7), 4664–4674 (2023) 9, 11
48. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 9
49. Wang, D., Lu, Y., Tian, J.: Polarization state tracing for reflection removal and color-consistent reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5680–5689 (June 2026) 4
50. Wang, J., Wang, K., Yu, Y., Lu, Y., Xiao, W., Sun, Z., Liu, F., Zou, Z., Gao, Y., Yang, L., et al.: Self-improving generative foundation model for synthetic medical image generation and clinical applications. Nature Medicine **31**(2), 609–617 (2025) 12
51. Wang, X., Chen, X., Ren, W., Han, Z., Fan, H., Tang, Y., Liu, L.: Compensation atmospheric scattering model and two-branch network for single image dehazing. IEEE Transactions on Emerging Topics in Computational Intelligence **8**(4), 2880–2896 (2024) 4
52. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 9

53. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The thrity-seventh asilomar conference on signals, systems & computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003) 9
54. Wei, C., Xiong, Z., Ren, W., Du, X., Zhang, G., Chen, W.: Omniedit: Building image editing generalist models through specialist supervision. In: The Thirteenth International Conference on Learning Representations (2024) 9, 10
55. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025) 10
56. Wu, K., Jiang, S., Ku, M., Nie, P., Liu, M., Chen, W.: Editreward: A human-aligned reward model for instruction-guided image editing. arXiv preprint arXiv:2509.26346 (2025) 9
57. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13294–13304 (2025) 10
58. Xie, X., Liu, J., Fan, H., Han, Z., Tang, Y., Qu, L.: Dvg-diffusion: Dual-view guided diffusion model for ct reconstruction from x-rays. IEEE Transactions on Image Processing (2026) 4
59. Xie, X., Liu, J., Lin, Z., Fan, H., Han, Z., Tang, Y., Qu, L.: Unleashing the potential of large language models for text-to-image generation through autoregressive representation alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11105–11113 (2026) 4
60. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al.: Lumina-dimoo: An omni diffusion large language model for multimodal generation and understanding. arXiv preprint arXiv:2510.06308 (2025) 4, 9, 10, 11, 12
61. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025) 2, 10, 11, 12
62. Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., Kong, L.: Dream 7b: Diffusion large language models. arXiv preprint arXiv:2508.15487 (2025) 4
63. You, Z., Nie, S., Zhang, X., Hu, J., Zhou, J., Lu, Z., Wen, J.R., Li, C.: Llada-v: Large language diffusion models with visual instruction tuning. arXiv preprint arXiv:2505.16933 (2025) 2, 4
64. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. arXiv preprint arXiv:2411.15738 (2024) 10
65. Yu, R., Ma, X., Wang, X.: Dimple: Discrete diffusion multimodal large language model with parallel decoding. arXiv preprint arXiv:2505.16990 (2025) 4
66. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 9
67. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023) 9
68. Zhou, L., Lou, A., Khanna, S., Ermon, S.: Denoising diffusion bridge models. In: International Conference on Learning Representations. vol. 2024, pp. 8160–8171 (2024) 4

69. Zhu, Y., Wang, X., Lathuilière, S., Kalogeiton, V.: Di [m] o: Distilling masked diffusion models into one-step generator. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18606–18618 (2025) 4
70. Zhu, Y., Wang, X., Lathuilière, S., Kalogeiton, V.: Soft-di [m] o: Improving one-step discrete image generation with soft embeddings. arXiv preprint arXiv:2509.22925 (2025) 4