LADR: Locality-Aware Dynamic Rescue for Efficient Text-to-Image Generation with Diffusion Large Language Models

Chenglin Wang[1,*], Yucheng Zhou[2,*], Shawn Chen[3], Tao Wang[4], Kai Zhang[1,†]
[1]East China Normal University, [2]University of Macau
[3]Zhejiang University, [4]Nanjing University
52275901013@stu.ecnu.edu.cn, yucheng.zhou@connect.um.edu.mo
kzhang980@gmail.com

Abstract

Discrete Diffusion Language Models have emerged as a compelling paradigm for unified multimodal generation, yet their deployment is hindered by high inference latency arising from iterative decoding. Existing acceleration strategies often require expensive re-training or fail to leverage the 2D spatial redundancy inherent in visual data. To address this, we propose **Locality-Aware Dynamic Rescue (LADR)**, a training-free method that expedites inference by exploiting the spatial Markov property of images. LADR prioritizes the recovery of tokens at the "generation frontier", regions spatially adjacent to observed pixels, thereby maximizing information gain. Specifically, our method integrates morphological neighbor identification to locate candidate tokens, employs a risk-bounded filtering mechanism to prevent error propagation, and utilizes manifold-consistent inverse scheduling to align the diffusion trajectory with the accelerated mask density. Extensive experiments on four text-to-image generation benchmarks demonstrate that our LADR achieves an approximate **4× speedup** over standard baselines. Remarkably, it maintains or even enhances generative fidelity, particularly in spatial reasoning tasks, offering a state-of-the-art trade-off between efficiency and quality.

1 Introduction

The field of generative modeling has witnessed a paradigm shift with the rapid evolution of Discrete Diffusion Language Models (DLMs) (Sahoo et al., 2024; Nie et al., 2025; Xin et al., 2025). Unlike Autoregressive (AR) models (Radford et al., 2018; Touvron et al., 2023; Achiam et al., 2023; Zhou et al., 2026; Song et al., 2026; Zhou et al., 2025b, 2024) that generate sequences strictly left-to-right, or Continuous Diffusion Models (Ho et al., 2020; Rombach et al., 2022; Liu et al., 2023; Yang et al., 2025b) that operate in continuous latent or pixel space, DLMs formulate visual generation (Wang et al., 2025a; Zhou et al., 2025a; Esser et al., 2021) as a bidirectional masked modeling task within a discretized vector-quantized (VQ) latent space. This paradigm not only enables flexible, non-sequential generation orders but also facilitates unified multimodal understanding and generation within a single framework (Li et al., 2025b; You et al., 2025). By treating visual patches as discrete tokens akin to text, DLMs have achieved impressive scalability and fidelity, emerging as a powerful competitor to traditional paradigms.

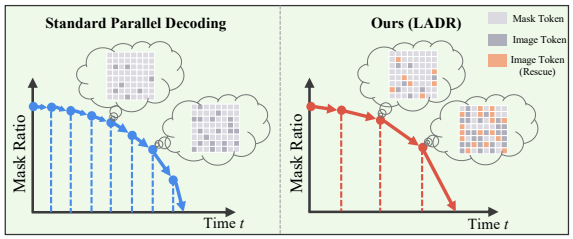

Figure 1: Comparison between Standard Parallel Decoding and our LADR method. While standard parallel decoding follows a fixed schedule, LADR accelerates decoding by exploiting spatial locality to dynamically recover neighbor tokens and keeps generation quality.

However, the iterative nature of DLMs imposes a severe bottleneck on inference efficiency. High-fidelity generation typically requires 50 to 100 forward passes to progressively refine the noisy sequence. Unlike AR models that benefit from KV-caching mechanisms to reuse historical computations, masked diffusion models must re-calculate bidirectional attention interactions at every step. While acceleration techniques exist, they often fall short in practicality: distillation-based methods (Zhu et al., 2025a,b) require computationally expensive re-training and student-teacher alignment, limiting their flexibility. On the other hand, heuristic strategies borrowed from textual Masked Language Models (MLMs) (Li et al., 2025a; Ye

* Equal Contributions.
† Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27170–27183
July 2-7, 2026

et al., 2025) often fail to generalize to the visual domain, as they overlook the fundamental difference between 1D textual dependencies and 2D visual structures.

Our work addresses this inefficiency by exploiting a property intrinsic to images but largely ignored in standard parallel decoding: *Spatial Locality*. As illustrated in Fig. 1, standard decoding schedules (Chang et al., 2022; You et al., 2025) (*e.g.,* Cosine) assume isotropic uncertainty reduction, treating all masked tokens as independent variables. In contrast, we observe that images exhibit a strong spatial Markov property, the uncertainty of a pixel is significantly reduced if its immediate spatial neighbors are known. Based on this insight, we hypothesize that the most efficient decoding path is not random, but topological. By prioritizing the "generation frontier", the unmasked tokens spatially adjacent to observed regions, we can accelerate the transition from noise to structure.

To materialize this insight, we propose **Locality-Aware Dynamic Rescue (LADR)**, a training-free acceleration method tailored for discrete visual generation. LADR dynamically modifies the decoding trajectory through three coupled mechanisms. First, it employs morphological operations to identify the topological neighbors of the current generation frontier. Second, to prevent the "hallucination" risks associated with aggressive acceleration, we introduce a risk-bounded filtering mechanism derived from the confidence gap of the model's posterior. Finally, to address the distribution shift caused by rapid mask reduction, we devise a *Manifold-Consistent Inverse Scheduling* strategy that re-aligns the diffusion timesteps with the actual mask density, ensuring the denoiser operates within its trained support.

We extensively validate LADR on multiple comprehensive benchmarks, including GenEval (Ghosh et al., 2023), UniGenBench (Wang et al., 2025d), DPG-Bench (Hu et al., 2024), and T2I-CompBench (Huang et al., 2023). Experimental results demonstrate that LADR significantly outperforms standard sampling and existing acceleration baselines. Notably, our method achieves an approximate **4× speedup** (reducing inference time from ∼57s to ∼13s) without compromising generative quality. In tasks requiring spatial reasoning (*e.g.,* object positioning and counting), LADR even surpasses the baseline performance, suggesting that enforcing spatial contiguity during decoding acts as a beneficial inductive bias.

In summary, our contributions are as follows:

- We identify *spatial locality* as a critical but underutilized source of information gain in discrete diffusion, theoretically showing that topology-aware decoding minimizes conditional entropy more effectively than random selection.
- We propose **LADR**, a plug-and-play acceleration method that integrates morphological neighbor identification, theoretically grounded risk filtering, and inverse scheduling to safely expedite inference without re-training.
- We achieve state-of-the-art efficiency-quality trade-offs on widely adopted benchmarks, demonstrating that LADR can accelerate large-scale multimodal DLMs by $4\times$ while maintaining robust semantic alignment and visual fidelity.

2 Related Work

Discrete Diffusion Language Models (DLMs) cast image generation as iterative masked token recovery in a discretized VQ space, enabling parallel decoding and substantially fewer sampling steps than continuous diffusion (Sahoo et al., 2024; Nie et al., 2025; Song et al., 2025; Arriola et al., 2025; Ho et al., 2020; Rombach et al., 2022; Chen et al., 2025; Yang et al., 2025c,a). Initiated by MaskGIT (Chang et al., 2022), this framework has been extended by Paella (Rampas et al., 2022) and Muse (Chang et al., 2023) to improve robustness and semantic control, and more recently generalized to unified multimodal generation by modeling visual and textual tokens as a single sequence (You et al., 2025; Swerdlow et al., 2025; Xin et al., 2025; Li et al., 2025b). Despite these advances, masked discrete diffusion remains latency-bound due to its reliance on iterative refinement with bidirectional attention and dynamically changing masks, which precludes computation reuse and contrasts sharply with KV-cached autoregressive decoding (Li et al., 2024; Bai et al., 2023; Guo et al., 2025; Cai et al., 2024). Distillation-based acceleration methods compress multi-step diffusion trajectories (Hinton, 2014; Song et al., 2023; Deschenaux and Gulcehre, 2025), but adapting consistency-style objectives to discrete VQ spaces is non-trivial and typically requires expensive retraining or relaxation techniques (Zhu et al., 2025a,b). In parallel, training-free acceleration heuristics have shown promise in text diffusion and sequence models (Wu et al., 2025a; Hu et al., 2025; Wu et al., 2025b; Wang et al., 2025c; Li et al., 2025a; Israel et al., 2025),

yet their direct transfer to image generation remains limited, as they fail to explicitly exploit the strong 2D spatial locality inherent in visual tokens. The extended version can be found in Appendix B.

3 Methodology

As illustrated in Fig. 2, we proposed **Locality-Aware Dynamic Rescue (LADR)**, a method designed to accelerate Discrete Diffusion Language Models (DLMs) while preserving generation quality. Our approach is grounded in the observation that standard parallel decoding typically treats tokens as independent variables given the global context. However, image representations derived from convolutional encoders inherently exhibit strong *spatial locality*. In this section, we first formalize the generation process and then provide the theoretical motivation grounded in information theory and risk estimation to drive our three algorithmic components: morphological neighbor identification, risk-bounded filtering, and manifold-consistent inverse scheduling.

3.1 Preliminaries: Discrete Diffusion and Variational Bounds

Let $\mathbf{z}_0 = [z_{0,1}, \dots, z_{0,N}] \in \mathcal{V}^N$ represent a discrete image sequence flattened from a $H \times W$ feature map, where each token belongs to a codebook $\mathcal{V}$. The discrete diffusion process is a forward Markov chain $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ that progressively corrupts $\mathbf{z}_0$ by replacing tokens with a special [MASK] token. The marginal distribution at time $t \in [0, 1]$ is given by:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \prod_{i=1}^{N} q(z_{t,i}|z_{0,i}),$$
$$\text{where } q(z_{t,i} = \texttt{[MASK]}|z_{0,i}) = \gamma(t), \quad (1)$$

where $\gamma(t)$ is a monotonic masking schedule (*e.g.,* cosine) representing the probability of a token being masked. The reverse process $p_\theta(\mathbf{z}_0|\mathbf{z}_t)$ approximates the true posterior $q(\mathbf{z}_0|\mathbf{z}_t)$. The training objective is to minimize the negative Evidence Lower Bound (ELBO), which simplifies to the negative log-likelihood over masked regions $\mathcal{M}_t$:

$$\mathcal{L} \approx \mathbb{E}_{t,\mathbf{z}_0}\left[-\sum_{i\in\mathcal{M}_t} \log p_\theta(z_{0,i}|\mathbf{z}_{t,\mathcal{O}_t})\right], \quad (2)$$

where $\mathcal{O}_t$ denotes the set of observed indices. During inference, iterative decoding approximates the joint distribution via conditional independence assumption: $p_\theta(\mathbf{z}_0|\mathbf{z}_t) \approx \prod_{i\in\mathcal{M}_t} p_\theta(z_{0,i}|\mathbf{z}_{t,\mathcal{O}_t})$. Standard acceleration methods strictly follow $\gamma(t)$, discarding potentially correct predictions in early stages.

3.2 Theoretical Motivation

Instead of relying on heuristic acceleration, we formulate LADR by analyzing the entropy reduction and risk bounds within the discrete latent space.

3.2.1 Entropy Reduction via Local Information Gain

Standard decoding assumes isotropic uncertainty reduction. However, since discrete image tokens $\mathbf{z}$ are typically obtained via CNN-based encoders (*e.g.,* VQGAN), the dependency between tokens decays with spatial distance due to bounded *Effective Receptive Fields (ERFs)*. We quantify the uncertainty of a masked token z_i using Conditional Entropy $H(z_i|\mathbf{z}_\mathcal{O})$. The reduction in uncertainty gained by observing an auxiliary set $\mathcal{S}$ is quantified by the Conditional Mutual Information:

$$I(z_i; \mathbf{z}_\mathcal{S}|\mathbf{z}_\mathcal{O}) = H(z_i|\mathbf{z}_\mathcal{O}) - H(z_i|\mathbf{z}_\mathcal{O}, \mathbf{z}_\mathcal{S}). \quad (3)$$

Definition 1 (Generation Frontier)**.** Given a binary mask $\mathbf{M}$, the generation frontier $\mathcal{F}$ is defined as the set of masked tokens spatially adjacent to currently observed tokens: $\mathcal{F} = \{i \mid m_i = 1 \wedge \exists j \in \mathcal{N}(i), m_j = 0\}$, where $\mathcal{N}(i)$ is the local spatial neighborhood.

Proposition 1 (Locality-Driven Information Lower Bound)**.** *Given the spatial inductive bias of the encoder, the mutual information between a token z_i and its immediate neighborhood $\mathcal{N}(i)$ dominates that of distant context $\mathcal{S}_{dist}$. Formally:*

$$I(z_i; \mathbf{z}_{\mathcal{N}(i)}|\mathbf{z}_\mathcal{O}) \gg I(z_i; \mathbf{z}_{\mathcal{S}_{dist}}|\mathbf{z}_\mathcal{O}). \quad (4)$$

Remark. This proposition provides the theoretical justification for our **Morphological Neighbor Identification** strategy: by prioritizing the generation frontier $\mathcal{F}$, LADR maximizes the expected information gain per decoding step, guiding sampling along the local structure of the latent manifold. This locality assumption is also empirically illustrated in Fig. 3, where perturbing a small number of VQ tokens induces only spatially confined changes in the decoded image, while the global structure remains largely intact.

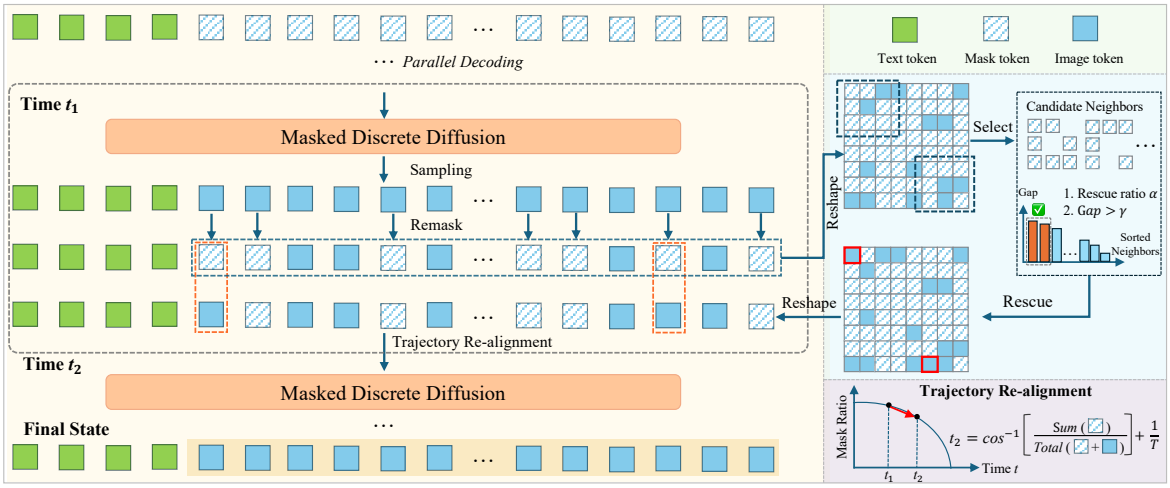

Figure 2: Overview of the LADR method. At each timestep, the flattened discrete tokens were reshaped into a 2D grid to identify candidate neighbors adjacent to resolved regions. These candidates are evaluated using the *Confidence Margin* (confidence top1-top2 gap) and are dynamically "rescued" (unmasked) based on an adaptive rescue ratio α and threshold γ. To synchronize the generation timeline with this accelerated accumulation of tokens, the *Trajectory Re-alignment* module utilizes an inverse cosine function to re-calculate the effective timestep t_2, allowing the scheduler to skip redundant iterations while maintaining consistency.

3.2.2 Safety Guarantee via Margin Bounds

Accelerating generation involves "rescuing" tokens before their scheduled timestamp. To control the quality, we must bound the probability of misclassification. We employ the *Confidence Gap*, $\Delta_i = p_{(1)} - p_{(2)}$, where $p_{(1)}$ and $p_{(2)}$ are the top-1 and top-2 probabilities.

Theorem 1 (Margin-based Error Bound)**.** *Consider a classification task over K classes. If the predicted distribution satisfies a confidence margin $\Delta \geq \tau$, the probability of error $P(\mathcal{E})$ is strictly upper bounded. Specifically, in the worst-case distribution scenario:*

$$P(\mathcal{E}) \leq 1 - \left(\frac{1+\tau}{2}\right). \tag{5}$$

Proof. See Appendix A.1.

Remark. This theorem provides a controllable mechanism. By enforcing a dynamic threshold $\tau(t)$, we can theoretically bound the error rate of our acceleration module, ensuring that the rescued tokens satisfy a minimum reliability standard.

3.2.3 Manifold Consistency via Inverse Scheduling

A critical challenge in acceleration is the **Training-Inference Mismatch**. Aggressive rescue reduces the mask ratio ρ_{act} faster than the scheduler $\gamma(t)$ expects, potentially pushing the state out-of-distribution (OOD).

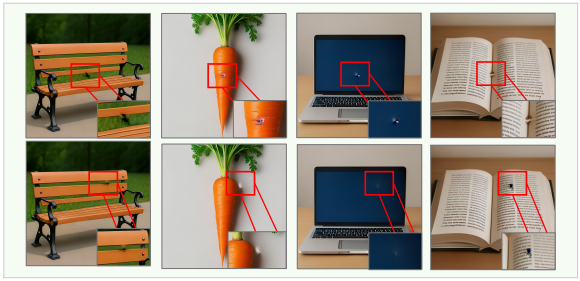

Figure 3: Visualization of localized semantic changes caused by perturbing a small set of VQ tokens. The impact remains spatially confined, supporting the locality assumption that nearby tokens dominate information gain.

Proposition 2 (Manifold Consistency Condition)**.** *To ensure the input state remains within the support of the trained diffusion manifold, the conditioning timestep t must be re-aligned such that the expected mask density matches the actual observation density:*

$$t_{new} = \gamma^{-1}(\rho_{act}) \quad s.t. \quad \mathbb{E}[\rho(t_{new})] \approx \rho_{act}. \tag{6}$$

Remark. This necessitates our **Inverse Scheduling** technique, which acts as a temporal projection operator to correct the trajectory after aggressive rescue operations. γ^{-1} is the inverse function of γ.

3.3 The LADR Method

Guided by the theoretical motivations above, LADR dynamically updates the mask tokens $\mathbf{M}_t$

at each timestep t through three coupled steps.

3.3.1 Morphological Neighbor Identification

Leveraging Proposition 1, we aim to identify the frontier $\mathcal{F}$. We map the 1D mask sequence to the 2D spatial grid $\Phi : \{0,1\}^N \to \{0,1\}^{H\times W}$. Using a spatial kernel $\mathbf{K}$ (*e.g.,* 3×3), the candidate neighbors $\mathcal{C}_t$ are identified via morphological dilation:

$$\mathbf{M}_{\text{grid}} = \Phi(\mathbf{M}_t), \tag{7}$$

$$\mathbf{M}_{\text{frontier}} = \mathbf{M}_{\text{grid}} \wedge (\neg \mathbf{M}_{\text{grid}} \oplus \mathbf{K}), \tag{8}$$

$$\mathcal{C}_t = \{i \mid \Phi^{-1}(\mathbf{M}_{\text{frontier}})[i] = 1\}. \tag{9}$$

This explicitly selects masked tokens that share spatial connectivity with observed regions.

3.3.2 Phase-Aware Dynamic Filtering

For every candidate $i \in \mathcal{C}_t$, we compute the confidence gap Δ_i. Guided by Theorem 1, we employ a dynamic policy $\Pi(t) = (\alpha_t, \tau_t)$ that adapts to the entropy of the generation phase defined by the effective timestep t_{eff}:

- **Exploration Phase** ($t_{eff} < 0.2$)**:** Global entropy is high. We apply a strict threshold $\tau = 0.05$ and limit the rescue ratio $\alpha = 0.1$ to prevent error propagation.
- **Structure Phase** ($0.2 \le t_{eff} < 0.7$)**:** As semantics emerge, we relax constraints ($\tau = 0.05, \alpha = 0.3$).
- **Refinement Phase** ($t_{eff} \ge 0.7$)**:** We aggressively rescue neighbors ($\tau = \emptyset, \alpha = 1.0$) to fill texture details.

The set of rescued tokens $\mathcal{R}_t$ is:

$$\mathcal{R}_t = \text{TopK}_{\Delta}\left(\{i \in \mathcal{C}_t \mid \Delta_i > \tau_t\}, \lfloor |\mathcal{C}_t| \cdot \alpha_t \rfloor\right). \tag{10}$$

3.3.3 Trajectory Re-alignment

After unmasking $\mathcal{R}_t$, the sequence sparsity decreases to ρ_{new}. Crucially, continuing with the original schedule t would violate the manifold consistency (Proposition 2). We thus re-calculate the next sampling step t_{next} using the inverse schedule:

$$t_{next} = \text{clamp}\left(\gamma^{-1}(\rho_{new}) + \frac{1}{T}, 0, 1\right). \tag{11}$$

This adjustment ensures that the noise level estimates remain accurate, effectively "skipping" redundant diffusion steps. The complete procedure is summarized in Algorithm 1.

Algorithm 1 Locality-Aware Dynamic Rescue (LADR)

Require: Pre-trained DLM p_θ, Scheduler $\gamma(\cdot)$, Steps T
Ensure: Discrete tokens $\mathbf{z}$

Initialize: $\mathbf{z} \leftarrow [\texttt{MASK}]^N$, $\mathbf{M} \leftarrow \mathbf{1}^N$
$N_{total} \leftarrow N$
for $step = 0$ **to** $T-1$ **do**
  if $\sum \mathbf{M} = 0$ **then**
    break
  end if
  {*Step 1: Inverse Scheduling (Prop. 1)*}
  $\rho_{curr} \leftarrow (\sum \mathbf{M})/N_{total}$
  $t_{eff} \leftarrow \gamma^{-1}(\rho_{curr})$ // Align time with mask density
  $t_{next} \leftarrow \text{clamp}(t_{eff} + 1/T, 0, 1)$
  $n_{mask} \leftarrow \lfloor N_{total} \cdot \gamma(t_{next}) \rfloor$ // Target mask count
  {*Step 2: Parallel Prediction*}
  $\mathbf{L} \leftarrow p_\theta(\mathbf{z}, t_{next})$
  $\mathbf{P} \leftarrow \text{Softmax}(\mathbf{L})$
  $\mathbf{z}_{pred} \leftarrow \text{Sampling}(\mathbf{P})$
  $\mathbf{\Delta} \leftarrow \text{Top1}(\mathbf{P}) - \text{Top2}(\mathbf{P})$
  {*Step 3: Standard Selection (Global)*}
  $\mathcal{I}_{rank} \leftarrow \text{Argsort}(\text{Top1}(\mathbf{P}) \cdot \mathbf{M}, \text{descending})$
  $\mathbf{M}_{std} \leftarrow \mathbf{1}^N$
  $\mathbf{M}_{std}[\mathcal{I}_{rank}[n_{mask} : N]] \leftarrow 0$ // Unmask most confident
  {*Step 4: Neighbor Rescue (Lemma 1 & Thm 1)*}
  $\mathbf{M}_{grid} \leftarrow \text{Reshape}(\mathbf{M}_{std}, H, W)$
  $\mathbf{M}_{front} \leftarrow \mathbf{M}_{std} \wedge \text{Flatten}(\neg \mathbf{M}_{grid} \oplus \mathbf{K}_{3\times 3})$
  $\mathcal{C}_{neigh} \leftarrow \{i \mid \mathbf{M}_{front}[i] = 1\}$
  if $|\mathcal{C}_{neigh}| > 0$ **then**
    Get α, τ based on t_{eff} (Sec 3.3.2)
    $\mathcal{C}_{valid} \leftarrow \{i \in \mathcal{C}_{neigh} \mid \mathbf{\Delta}[i] > \tau\}$
    $k_{res} \leftarrow \min(\lfloor |\mathcal{C}_{neigh}| \cdot \alpha \rfloor, |\mathcal{C}_{valid}|)$
    $\mathcal{S}_{res} \leftarrow \text{Argsort}(\mathbf{\Delta}[\mathcal{C}_{valid}], \text{descending})[0 : k_{res}]$
    $\mathbf{M}_{std}[\mathcal{S}_{res}] \leftarrow 0$
  end if
  {*Step 5: State Update*}
  $\mathbf{M} \leftarrow \mathbf{M}_{std}$
  $\mathbf{z} \leftarrow \mathbf{z}_{pred} \odot (\mathbf{1} - \mathbf{M}) + [\texttt{MASK}] \odot \mathbf{M}$
end for
return $\mathbf{z}$

4 Experiments

4.1 Experimental setup

Benchmarks and Baselines. To strictly evaluate the effectiveness of our method on both inference efficiency and visual fidelity, we conducted evaluations across four publicly popular text-to-image generation benchmarks: **GenEval** (Ghosh et al., 2023), **UniGen-Bench** (Wang et al., 2025d), **DPG-Bench** (Hu et al., 2024), and **T2I-CompBench** (Huang et al., 2023). These benchmarks provide a comprehensive assessment spanning from basic object semantics to complex compositional generation. Furthermore, to ensure a fair and focused evaluation, we compared our method against two representative training-free acceleration method: (1) **ML-Cache** (Xin et al., 2025), the native caching optimization strategy embedded in the Lumina-DiMOO backbone; and (2) **Prophet** (Li et al., 2025a), a heuristic-based ac-

celerated decoding method originally designed for text generation, which we adapted to the visual domain to investigate the cross-modal applicability of textual heuristics.

Implementation Details. For a fair and controlable comparison, we adopt the unified multimodal model Lumina-DiMOO (Xin et al., 2025) as the foundational DLM backbone for all experiments since comparable open-source models are limited. The generated image resolution is 1024×1024. To ensure a consistent evaluation of inference latency, all models and baselines were executed locally on a single NVIDIA A100 (80GB) GPU. We adhered to the standard inference configurations of the backbone model, reporting the performance following the evaluation scripts of each benchmark. Notably, for the UniGen-Bench, we used the version of their released scripts about the open-source vision-language model Qwen2.5-VL-72B (Team, 2025) to evaluate. Follow Prophet (Li et al., 2025a), we divided the parallel decoding process into three phases, and set phase-aware thresholds to regulate the rescued neighbors in the decoding process. Much like it established a proof-of-concept for heuristic-based acceleration in text decoding, our work aims to pioneer a similar trajectory for visual decoding. Consequently, we did not perform an exhaustive grid search to obtain these parameters for each specific dataset.

4.2 Main Results

We empirically investigated the effectiveness of our proposed accelerated method LADR by answering two critical research questions: (1) *Does the method deliver substantial speedups compared to existing caching and heuristic strategies?* (2) *Can it maintain or even enhance generative fidelity in some scenarios?*

Decoding Efficiency Analysis. The primary motivation of our approach is to alleviate the computational bottleneck of parallel iterative decoding. As presented in Tables 1 through 4, our method demonstrates a dramatic reduction in inference latency across all benchmarks. On average, our accelerated model completes inference in approximately $\sim$13-14 seconds, representing a $\mathbf{4}\times$ speedup over the No-Cache ($\sim$57s) and a $\mathbf{2}\times$ speedup over the optimized ML-Cache and Prophet ($\sim$32s). Notably, our method outperforms the text-optimized Prophet algorithm, confirming that our locality-aware rescue strategy is inherently more suitable for the 2D visual domain than heuristics transplanted from 1D text generation.

Generative Quality Analysis. Beyond efficiency, our results indicate that the significant reduction in sampling steps does not come at the cost of visual fidelity, while it yields a highly competitive performance profile across diverse benchmarks. Conversely, we note a performance drop in fine-grained high-frequency tasks, such as the text rendering score in UniGen-Bench, suggesting that the model remains sensitive to the reduction of iterative refinement steps in some scenarios. To further investigate this, we conducted an additional experiment on UniGen-Bench's text category exploring a delayed rescue strategy. Specifically, we completely disabled the rescue operation during the Exploration Phase ($t_{eff} < 0.2$) and kept LADR's other settings unchanged. As illustrated in Table 5, the text rendering score recovers to 26.10, which is comparable to the base model's 27.87. Meanwhile, the inference time is 22.51s, still maintaining a highly efficient 2.5x speedup. This demonstrates that LADR could achieve a highly adaptable efficiency-quality trade-off. Overall, the experiment results demonstrate that our method achieves a superior efficiency-fidelity trade-off, delivering efficient decoding speeds while preserving robust generative capabilities in image generation.

Case Visualization. To intuitively assess the impact of acceleration on perceptual quality, We visualized some cases generated by our proposed LADR method alongside the base model and two training-free baselines. As shown in Figure 4, our approach achieves a significant speedup ($4\times$ faster than the backbone) while preserving intricate visual details (*e.g.,* reflections in the "wooden boats") and correct semantic composition (*e.g.,* "red cup and pink handbag"), demonstrating that our spatial-aware acceleration could maintain the generative quality of the underlying model.

4.3 Ablation Studies and Analysis

Impact of Spatial Selection Strategy. To strictly validate our hypothesis that spatial locality is the critical factor for acceleration, we conducted an ablation study on the token selection criteria. Specifically, we first determined the counts k of rescued neighbor tokens via LADR at each timestep, and then enforced this exact budget on two strategies:

- **Non-Neighbor Prioritization.** This strategy

Method	Avg. t (s)↓	Two Obj.	Colors	Attribute	Single Obj.	Position	Counting	Overall ↑
Lumina-DiMOO	57.01	**93.94**	**91.49**	73.00	97.50	79.00	85.00	86.66
+ ML-Cache	31.95	93.75	89.63	**75.50**	**100.00**	84.50	**85.94**	**87.83**
+ Prophet	32.15	91.16	86.44	70.75	96.56	74.50	84.06	83.91
+ **LADR(Ours)**	**13.22**	91.41	91.22	74.75	99.06	**85.50**	81.88	87.30

Table 1: Performance comparison on the **GenEval** (Ghosh et al., 2023).

Method	Avg. t (s)↓	Style	Know.	Attr.	Action	Rel.	Cmp.	Gram.	Logic.	Lay.	Text
Lumina-DiMOO	57.21	91.52	**89.87**	79.29	71.48	78.55	73.45	**69.79**	43.58	**85.63**	**27.87**
+ ML-Cache	31.96	91.40	88.77	79.17	73.67	79.06	74.61	69.65	**45.87**	83.96	26.15
+ Prophet	32.47	87.40	84.34	75.64	68.25	75.63	65.85	64.97	39.91	83.02	25.57
+ LADR(**Ours**)	**13.94**	**94.80**	88.61	**81.73**	**75.95**	**81.98**	**77.71**	66.31	45.41	85.26	16.38

Table 2: Performance comparison on **UniGen-Bench** (Wang et al., 2025d).

Method	Avg. t (s) ↓	Color	Shape	Texture	Spatial	Non-spatial	Complex
Lumina-DiMOO	56.92	81.07	57.02	69.30	46.70	31.70	34.98
+ ML-Cache	31.78	81.52	57.75	70.28	**46.79**	**31.83**	35.23
+ Prophet	32.43	80.92	55.48	70.34	42.37	31.46	34.92
+ LADR(**Ours**)	**13.41**	**82.25**	**58.94**	**72.34**	46.35	31.60	**36.19**

Table 3: Performance Comparison on **T2I-CompBench** (Huang et al., 2023).

Method	Avg. t (s) ↓	Global	Entity	Attribute	Relation	Other	Overall ↑
Lumina-DiMOO	58.16	77.20	90.36	87.93	93.04	82.80	83.61
+ ML-Cache	32.01	81.46	90.37	88.16	93.27	**84.40**	84.02
+ Prophet	34.63	81.76	89.56	87.33	92.76	83.20	82.91
+ LADR(**Ours**)	**14.52**	**84.19**	**91.47**	**89.12**	**94.20**	81.20	**85.42**

Table 4: Performance evaluation on **DPG-Bench** (Hu et al., 2024).

Method	Avg.t (s) ↓	Text
Lumina-DiMOO	57.21	27.87
LADR	13.94	16.38
LADR(delayed rescue)	22.51	26.10

Table 5: Efficiency and performance comparison of different strategies on the text category of UniGen-Bench.

prioritizes isolated tokens with the top-k confidence gaps, only reverting to neighbors if the non-neighbor set is exhausted.

- **Random Selection.** This strategy randomly sampled from the remask token, no better neighbor or non-neighbor tokens.

Figure 5 presents the quantitative comparison on the GenEval benchmark. We observe that the Non-Neighbor strategy yields the lowest performance (Overall score: 84.05), significantly lagging behind our method, which achieves an overall score of 87.30. This confirms that forcing the model to resolve isolated tokens early, even those with high confidence gaps, leads to error propagation, as these predictions may lack sufficient spatial grounding. Interestingly, the random strategy outperforms the Non-Neighbor variant, but it is still lagging behind our proposed accelerated method. This indicates that our locality-aware dynamic rescue strategy provides the optimal balance, ensuring that the accelerated decoding trajectory respects the structural dependencies of the image.

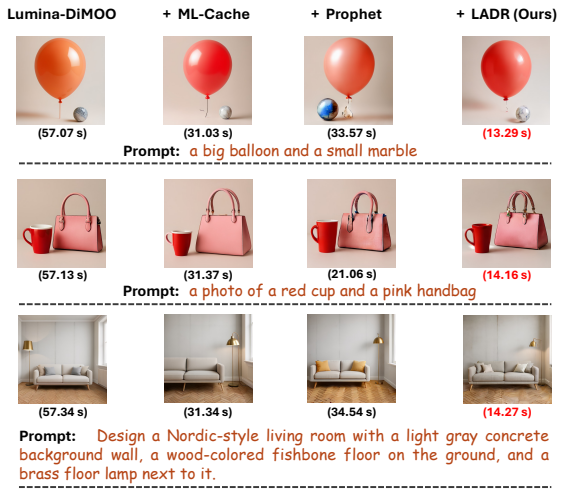

Figure 4: Qualitative comparison of different methods in terms of generation fidelity and inference time, where the corresponding text prompt is provided below each row, with the inference latency displayed in parentheses under the image.

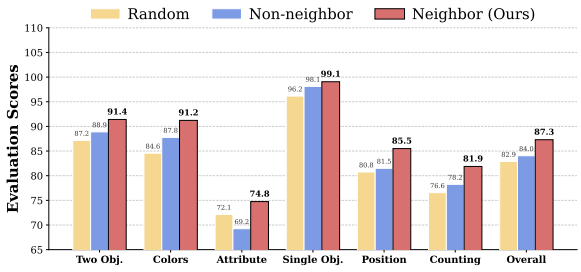

Figure 5: Ablation study of spatial selection strategies on the GenEval benchmark."Random" means *Random Selection*, "Non-neighbor" represents *Non-Neighbor Prioritization*, and "Neighbor (Ours)" is our strategy that rescues neighbor tokens.

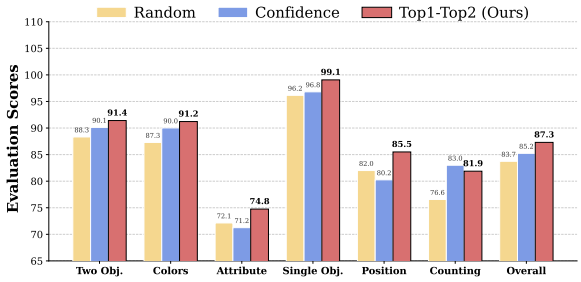

Figure 6: Ablation study about different ranking metrics of neighbor tokens on the GenEval benchmark."Random" means *Random Neighbor Selection*, "Confidence" denotes *Standard Confidence*, and "Top1-Top2 (Ours)" means confidence margin we employed.

Effectiveness of Confidence Margin. Besides the necessity of spatial locality, we further investigate the optimality of the ranking metric used to filter these spatial candidates. To verify whether the *Confidence Margin* (Top1-Top2 gap) provides a superior signal compared to standard confidence scores, we evaluate distinct prioritization criteria for selecting the rescued tokens $\mathcal{R}_t$ in eq (10):

- **Standard Confidence (Top-1 Probability).** This variant ranks neighbors solely by the probability of the most likely token.
- **Random Neighbor Selection.** This baseline selects tokens stochastically from the neighborhood $\mathcal{C}_t$ in eq (9), ignoring predictive certainty entirely.

Figure 6 reports the performance on the GenEval benchmark. The results demonstrate that our confidence margin strategy achieves the highest overall accuracy (87.30), surpassing the standard confidence baseline (85.24). While the standard confidence approach performs strongly in object-centric metrics like *Counting* (83.00), it underperforms in structural categories such as *Position* (80.25 vs. 85.50) and *Attribute* (71.25 vs. 74.75). The Random neighbor selection yields the lowest overall performance (83.74). These findings suggest that the Top1-Top2 gap is a more robust discriminator to recover tokens. It effectively penalizes ambiguously high predictions, where the model is confident in the top choice but equally confident in a competing alternative, thereby preventing the premature fixation of semantically unstable tokens.

5 Conclusion and Future Work

In this paper, we propose an accelerated parallel decoding strategy called LADR, which is a training-free method designed to unlock the inference efficiency of DLMs. By challenging the standard schedule generation difficulty, LADR exploits the intrinsic spatial locality of visual data. It dynamically rescues high-confidence tokens within resolved neighborhoods using a lightweight confidence margin, employing an inverse scheduling mechanism to adaptively re-align the generation timeline. Extensive evaluations across four publicly popular text-to-image generation benchmarks demonstrate that our method achieves a superior efficiency-fidelity trade-off. It delivers a significant $4\times$ speedup over non-cached baselines and $2\times$ speedup over heuristic-based methods, without model re-training or architectural modifications. Our findings underscore that while text-optimized heuristics provide a foundation, optimal accelera-

tion in the visual domain requires strategies that explicitly respect the 2D spatial structure of the modality, paving the way for plug-and-play DLMs.

Future Work. While LADR demonstrates strong efficiency–quality trade-offs for text-to-image diffusion, several directions remain open for future exploration. First, extending locality-aware rescue to temporally structured modalities such as video generation may require jointly modeling spatial and temporal frontiers, where locality spans both space and time. Second, we anticipate that integrating LADR with emerging architectural optimizations (e.g., sparse attention or lightweight distillation) may yield complementary gains, pushing DLMs closer to real-time multimodal generation.

Limitations

While LADR demonstrates the potential of exploiting image spatial locality for acceleration of parallel decoding, our current method still has some limitations. The implementation relies on empirically determined hyperparameters, such as the confidence threshold τ and rescue ratios α. These values were selected to validate the core hypothesis that spatial neighbors facilitate faster convergence rather than to locate the global optimum. Furthermore, as a training-free acceleration method, LADR's performance is influenced by the foundational backbone. To prevent unfaithful generation, existing post-hoc alignment frameworks can be integrated to ensure prompt-image faithfulness.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62276099).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Marianne Arriola, Subham Sekhar Sahoo, Aaron Gokaslan, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Justin T Chiu, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. In *Proceedings of International Conference on Learning Representations*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Baobao Chang, Junjie Hu, and Xiao Wen. 2024. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, and 1 others. 2023. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of International Conference on Machine Learning*, pages 4055–4075.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325.

Guanjie Chen, Xinyu Zhao, Yucheng Zhou, Xiaoye Qu, Tianlong Chen, and Yu Cheng. 2025. Towards stabilized and efficient diffusion transformers through long-skip-connections with spectral constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17708–17718.

Justin Deschenaux and Caglar Gulcehre. 2025. Beyond autoregression: Fast LLMs via self-distillation through time. In *Proceedings of International Conference on Learning Representations*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Proceedings of Advances in Neural Information Processing Systems*, pages 52132–52152.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Satoshi Hayakawa, Yuhta Takida, Masaaki Imaizumi, Hiromi Wakaki, and Yuki Mitsufuji. 2025. Distillation of discrete diffusion through dimensional correlations. In *Proceedings of International Conference on Machine Learning*.

G Hinton. 2014. Distilling the knowledge in a neural network. In *Deep Learning and Representation Learning Workshop in Conjunction with NIPS*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 6840–6851.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Zhanqiu Hu, Jian Meng, Yash Akhauri, Mohamed S Abdelfattah, Jae-sun Seo, Zhiru Zhang, and Udit Gupta. 2025. Accelerating diffusion language model inference via efficient kv caching and guided diffusion. *arXiv preprint arXiv:2505.21467*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Proceedings of Advances in Neural Information Processing Systems*, pages 78723–78747.

Daniel Mingyi Israel, Guy Van den Broeck, and Aditya Grover. 2025. Accelerating diffusion LLMs via adaptive parallel decoding. In *Proceedings of Annual Conference on Neural Information Processing Systems*.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. 2024. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *Proceedings of International Conference on Learning Representations*.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Yi Liang, Soroush Vosoughi, and Shiwei Liu. 2025a. Diffusion language models know the answer before decoding. *arXiv preprint arXiv:2508.19982*.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. 2025b. Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. *arXiv preprint arXiv:2509.19244*.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. 2025c. Lavida: A large diffusion language model for multimodal understanding. *arXiv preprint arXiv:2505.16839*.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024. Snapkv: Llm knows what you are looking for before generation. In *Proceedings of Advances in Neural Information Processing Systems*, pages 22947–22970.

Xingchao Liu, Chengyue Gong, and 1 others. 2023. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Dominic Rampas, Pablo Pernias, and Marc Aubreville. 2022. A novel sampling scheme for text-and image-conditional image synthesis in quantized latent spaces. *arXiv preprint arXiv:2211.07292*.

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

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 130136–130184.

Han Song, Yucheng Zhou, Jianbing Shen, and Yu Cheng. 2026. From broad exploration to stable synthesis: Entropy-guided optimization for autoregressive image generation. In *The Fourteenth International Conference on Learning Representations*.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models. In *Proceedings of International Conference on Machine Learning*, pages 32211–32252.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, and 1 others. 2025. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. 2025. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*.

Qwen Team. 2025. Qwen2.5-vl.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chenglin Wang, Yucheng Zhou, Qianning Wang, Zhe Wang, and Kai Zhang. 2025a. Complexbench-edit: Benchmarking complex instruction-driven image editing via compositional dependencies. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13391–13397.

Chenglin Wang, Yucheng Zhou, Zhe Wang, Zijie Zhai, Jianbing Shen, and Kai Zhang. 2025b. Alternate geometric and semantic denoising diffusion for protein inverse folding. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 350–366. Springer.

Xu Wang, Chenkai Xu, Yijie Jin, Jiachun Jin, Hao Zhang, and Zhijie Deng. 2025c. Diffusion llms can do faster-than-ar inference via discrete diffusion forcing. *arXiv preprint arXiv:2508.09192*.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. 2025d. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Shizhe Diao, Yonggan Fu, Zhijian Liu, Pavlo Molchanov, Ping Luo, Song Han, and Enze Xie. 2025a. Fast-dllm v2: Efficient block-diffusion llm. *arXiv preprint arXiv:2509.26328*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2025b. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, and 1 others. 2025. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*.

Hongji Yang, Wencheng Han, Yucheng Zhou, and Jianbing Shen. 2025a. Dc-controlnet: Decoupling inter-and intra-element conditions in image generation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19065–19074.

Hongji Yang, Yucheng Zhou, Wencheng Han, and Jianbing Shen. 2025b. Self-rewarding large vision-language models for optimizing prompts in text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7332–7349.

Hongji Yang, Yucheng Zhou, Wencheng Han, Runzhou Tao, Zhongying Qiu, Jianfei Yang, and Jianbing Shen. 2025c. Hicogen: Hierarchical compositional text-to-image generation in diffusion models via reinforcement learning. *arXiv preprint arXiv:2511.19965*.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. 2024. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2025. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.

Yucheng Zhou, Hao Li, and Jianbing Shen. 2026. Condition errors refinement in autoregressive image generation with diffusion loss. In *The Fourteenth International Conference on Learning Representations*.

Yucheng Zhou, Jiahao Yuan, and Qianning Wang. 2025a. Draw all your imagine: A holistic benchmark and agent framework for complex instruction-based image generation. *arXiv preprint arXiv:2505.24787*.

Yucheng Zhou, Jihai Zhang, Guanjie Chen, Jianbing Shen, and Yu Cheng. 2024. Less is more: Vision representation compression for efficient video generation with large language models. *OpenReview*.

Yucheng Zhou, Huan Zheng, Dubing Chen, Hongji Yang, Wencheng Han, and Jianbing Shen. 2025b. From medical llms to versatile medical agents: A comprehensive survey.

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. 2025a. Di [m] o: Distilling masked diffusion models into one-step generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18606–18618.

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. 2025b. Soft-di [m] o: Improving one-step discrete image generation with soft embeddings. *arXiv preprint arXiv:2509.22925*.

A Theoretical Proofs and Derivations

In this section, we provide the detailed mathematical derivations for the propositions and theorems presented in the main methodology.

A.1 Proof of Theorem 1 (Margin-based Error Bound)

Problem Statement: Let $\mathbf{p} = [p_1, p_2, \ldots, p_K]$ be the probability distribution over K classes, sorted such that $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(K)}$. The predicted class is $\hat{y} = \text{argmax}_k p_k$. The probability of error is $P(\mathcal{E}) = 1 - p_{(1)}$. Given the margin constraint $p_{(1)} - p_{(2)} \geq \tau$, we seek the upper bound of $P(\mathcal{E})$.

Proof. We aim to find the upper bound for the probability of error $P(\mathcal{E}) = 1 - p_{(1)}$. This is strictly equivalent to finding the minimum possible value for the confidence $p_{(1)}$ subject to the given margin constraints.

1. **Define Constraints:** The probability distribution must sum to 1, and the margin condition must hold:

$$p_{(1)} + p_{(2)} + \sum_{k=3}^{K} p_{(k)} = 1, \tag{12}$$

$$p_{(1)} - p_{(2)} \geq \tau \implies p_{(2)} \leq p_{(1)} - \tau. \tag{13}$$

2. **Worst-Case Analysis for $p_{(1)}$:** To minimize $p_{(1)}$, we must maximize the probability mass assigned to the remaining classes, particularly the closest competitor $p_{(2)}$. The most extreme case occurs when the entire probability mass is concentrated in the top two classes, meaning $\sum_{k=3}^{K} p_{(k)} \approx 0$.

Under this worst-case assumption, we have:

$$p_{(1)} + p_{(2)} \approx 1. \tag{14}$$

3. **Deriving the Lower Bound of Confidence:** Substituting the margin constraint from Eq. (13) into Eq. (14), we obtain:

$$\begin{aligned} 1 &= p_{(1)} + p_{(2)} \\ &\leq p_{(1)} + (p_{(1)} - \tau) \\ &= 2p_{(1)} - \tau. \end{aligned} \tag{15}$$

Rearranging the inequality gives the lower bound for $p_{(1)}$:

$$\begin{aligned} 2p_{(1)} &\geq 1 + \tau \\ p_{(1)} &\geq \frac{1+\tau}{2}. \end{aligned} \tag{16}$$

4. **Calculating the Error Bound:** The probability of error is the complement of the top-1 probability. Using Eq. (16), we bounded the error as:

$$\begin{aligned} P(\mathcal{E}) &= 1 - p_{(1)} \\ &\leq 1 - \frac{1+\tau}{2} \\ &= \frac{1-\tau}{2}. \end{aligned} \tag{17}$$

This concludes the proof. The error is strictly bounded by a linear function of the threshold τ. □

A.2 Justification of Proposition 1 (Locality-Induced Information Gain)

Proposition Statement: The mutual information $I(z_i; \mathbf{z}_{\mathcal{N}(i)})$ dominates $I(z_i; \mathbf{z}_{\mathcal{S}_{dist}})$.

Proof. We derive this property from the definition of Conditional Mutual Information and the Markov property of Convolutional Neural Networks (CNNs).

1. **Definition of Information Gain:** Let Ω be the set of all tokens. The information gain for a target token z_i given a subset $\mathcal{S}$ is:

$$\begin{aligned} IG(\mathcal{S}) &= I(z_i; \mathcal{S} | \Omega \setminus (\{z_i\} \cup \mathcal{S})) \\ &= H(z_i | \Omega \setminus (\{z_i\} \cup \mathcal{S})) - H(z_i | \Omega \setminus \{z_i\}). \end{aligned} \tag{18}$$

2. **Spatial Correlation Decay:** For latent codes z derived from a VQ-GAN encoder E, the covariance between features at spatial locations u and v generally follows a decay function dependent on Euclidean distance $d(u, v)$:

$$\text{Cov}(z_u, z_v) \propto \exp\left(-\frac{d(u,v)^2}{2\sigma^2}\right), \tag{19}$$

where σ corresponds to the Effective Receptive Field (ERF).

3. **Entropy and Correlation:** For Gaussian-like distributions, the conditional entropy is related to the correlation coefficient ρ:

$$H(z_i | z_j) \approx \frac{1}{2} \log(1 - \rho_{ij}^2) + \text{const}. \tag{20}$$

Higher correlation ρ_{ij} leads to lower conditional entropy $H(z_i|z_j)$.

4. **Comparing Neighbors vs. Distant Tokens:** Let $j \in \mathcal{N}(i)$ be a spatial neighbor and $k \in \mathcal{S}_{dist}$

be a distant token.

$$d(i,j) \ll d(i,k) \tag{21}$$
$$\Rightarrow \rho_{ij} > \rho_{ik} \tag{22}$$
$$\Rightarrow H(z_i|z_j) < H(z_i|z_k). \tag{23}$$

5. **Conclusion:** Since observing neighbors reduces the conditional entropy more than observing distant tokens:

$$I(z_i; \mathbf{z}_{\mathcal{N}(i)}) > I(z_i; \mathbf{z}_{\mathcal{S}_{dist}}). \tag{24}$$

Thus, the optimal strategy for variance reduction is to prioritize $\mathcal{N}(i)$. □

A.3 Derivation of Inverse Scheduling (Proposition 2)

Objective: Find the effective timestep t_{new} such that the model's training distribution matches the current observation density.

Proof. 1. **Forward Process Definition:** The masking probability at time t is given by the schedule function $\gamma(t)$:

$$q(z_{t,i} = \texttt{[MASK]}) = \gamma(t). \tag{25}$$

2. **Expected Mask Ratio:** For a sequence of length N, the number of masked tokens M_t follows a Binomial distribution. The expected mask ratio is:

$$\mathbb{E}\left[\frac{|\mathcal{M}_t|}{N}\right] = \gamma(t). \tag{26}$$

3. **Perturbation via Rescue:** The LADR algorithm unmasks a set of tokens $\mathcal{R}$, changing the actual mask ratio to ρ_{act}:

$$\rho_{act} = \frac{|\mathcal{M}_{prev}| - |\mathcal{R}|}{N}. \tag{27}$$

4. **Manifold Alignment:** To ensure the input to the denoiser $p_\theta(\mathbf{z}_0|\mathbf{z}_{t_{new}})$ is In-Distribution (ID), we require the expected mask ratio at t_{new} to equal the actual current ratio:

$$\gamma(t_{new}) = \rho_{act}. \tag{28}$$

5. **Solving for Timestep:** Assuming $\gamma(t)$ is monotonic and invertible (e.g., cosine schedule), we apply the inverse function:

$$t_{new} = \gamma^{-1}(\rho_{act}). \tag{29}$$

This creates the mapping required for the Manifold Consistent Inverse Scheduling. □

B Extended Related Work

In this section, we provide a detailed elaboration on the development of Masked Discrete Diffusion for image generation and the current landscape of acceleration strategies.

B.1 Masked Discrete Diffusion for Image Generation

Discrete Diffusion Language Models (DLMs) (Sahoo et al., 2024; Nie et al., 2025; Song et al., 2025; Arriola et al., 2025) have reformulated the generation process as masked modeling within a discretized vector-quantized (VQ) space. Pioneered by MaskGIT (Chang et al., 2022), this paradigm utilizes a bidirectional Transformer coupled with a mask-scheduling strategy to enable image synthesis via iterative parallel decoding. Compared to standard continuous diffusion models (Ho et al., 2020; Rombach et al., 2022), this formulation significantly curtails the required sampling steps. Building upon this foundation, subsequent architectures have rapidly expanded the field: Paella (Rampas et al., 2022) optimized U-Net backbones with noise-robust objectives, while Muse (Chang et al., 2023) demonstrated scalability by integrating pre-trained LLMs for enhanced semantic control.

More recently, the field has witnessed a shift towards unified multimodal understanding and generation (You et al., 2025; Swerdlow et al., 2025; Xin et al., 2025; Li et al., 2025b,c; Zhou et al., 2025a). Notably, frameworks like Lumina-DiMOO (Xin et al., 2025) and LaVida-O (Li et al., 2025b) adopt a generalized discrete diffusion approach that treats visual and textual tokens as a shared sequence. While facilitating versatile generative capabilities across modalities, the iterative mask recovery process still imposes a non-negligible computational overhead. This underscores the necessity for efficient acceleration strategies that can expedite inference without compromising generative integrity.

B.2 Acceleration of Masked Discrete Diffusion

The efficacy of discrete diffusion (Wang et al., 2025b) hinges on iterative refinement, where multiple forward passes resolve the joint distribution of tokens. Unlike autoregressive models that benefit from causal masking and KV-caching (Li et al., 2024; Bai et al., 2023; Guo et al., 2025; Cai et al., 2024), masked diffusion relies on bidirectional attention with dynamically shifting mask states (Sa-

hoo et al., 2024; Xin et al., 2025). This characteristic precludes the reuse of historical computations, creating a distinct latency bottleneck.

Distillation-Based Approaches. To alleviate latency, research has gravitated towards model distillation (Hinton, 2014; Song et al., 2023; Deschenaux and Gulcehre, 2025; Yin et al., 2024; Hayakawa et al., 2025). While Consistency Models (Song et al., 2023; Kim et al., 2024) are effective in continuous pixel space, adapting them to discrete VQ space requires specialized formulations due to the absence of explicit ODE trajectories (Zhu et al., 2025a). Works such as DiMO (Zhu et al., 2025a) and Soft-DiMO (Zhu et al., 2025b) address this using policy gradients and soft embedding relaxations to compress multi-step trajectories. However, these methods necessitate computationally expensive retraining and student-teacher alignment, limiting their plug-and-play applicability.

Training-Free Heuristics. Advancements in DLMs have also explored architectural optimizations (Wu et al., 2025a; Hu et al., 2025; Wu et al., 2025b; Wang et al., 2025c) and adaptive sampling (Li et al., 2025a; Israel et al., 2025) primarily for text generation. While effective for 1D sequences, their direct adaptation to the visual domain is non-trivial. The inherent gap between the sequential dependencies of text and the 2D spatial correlations of images renders text-optimized heuristics suboptimal for visual generation. This discrepancy highlights the need for acceleration strategies explicitly tailored to the spatial redundancy and structural properties of images.