# LEVERAGING DIFFUSION MLLMS FOR IMAGE EDITING VIA DYNAMIC MASKED GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Diffusion Multimodal Large Language Models (dMLLMs) have recently emerged as a powerful paradigm for multimodal generation. Their localized and progressive masked generation property naturally provides a promising foundation for training-free text-based image editing, *i.e.*, explicitly preserving editing-irrelevant regions from the source image, while treating regions to be edited as masked tokens for regeneration guided by the target prompt. Existing dMLLM-based editing method, however, relies on fixed edit masks, static source guidance, and pre-defined generation trajectories, resulting in inflexible edit localization and degraded source consistency. In this paper, we propose DynEdit, a general training-free framework that better exploits the intrinsic generation properties of dMLLMs for flexible and faithful image editing. Specifically, we introduce a parallel dual-branch framework that enables dynamic masked generation with three key designs: 1) Dynamic edit localization to automatically identify editable tokens throughout generation. 2) Adaptive source guidance to adjust source preservation according to specific editing areas and generation steps. 3) Edit-aware token decoding to dynamically refine the decoding schedule for more reliable generation. Extensive experiments on diverse image benchmarks have demonstrated that DynEdit consistently improves editing quality, editing efficiency, and source preservation under both mask-free and external mask-grounded settings, exhibiting strong generalization in various editing scenarios.

## 1 INTRODUCTION

Recent advances in Diffusion Multimodal Large Language Models (dMLLMs) (Shi et al., 2025; Xin et al., 2025; Yang et al., 2026) have emerged as a promising paradigm evolution for unified multimodal modeling. Built upon discrete diffusion architectures, dMLLMs represent both text and images as discrete tokens and formulate text-to-image generation as iterative masked token prediction in parallel, proceeding in a localized and progressive manner. This distinctive generation property thus provides a natural and potential foundation for the task of text-based image editing (*i.e.*, a source image is modified according to a target prompt) based on pre-trained dMLLMs. Specifically, editing-irrelevant regions can be explicitly preserved from the source image, while regions to be edited can be treated as masked tokens for regeneration guided by the target prompt.

The pioneering exploration GIDE (Zhu et al., 2026b) proposes to first employ an external grounding model (*e.g.*, SAM (Kirillov et al., 2023)) to obtain the mask of the edit area. It then directly preserves the source tokens outside the mask, and converts inside tokens into masked tokens for regeneration under the target prompt. While straightforward and training-free, this framework still faces notable limitations. **1) Inflexible Localization.** Since the edit area is determined by the grounding models at the beginning of the generation and then remains fixed, the editing quality becomes highly sensitive to the accuracy and suitability of such pre-defined masks. To mitigate this, GIDE further requires a post-refinement process to improve visual coherence and boundary alignment. However, such refinement remains constrained by the initial localization, making it difficult to handle edits that require large structural changes (*e.g.*, `bird` to `leopard` in Figure 1(a)). **2) Unfaithful Source Consistency.** To preserve the original context of the edit area, GIDE performs discrete inversion (Dao et al., 2025) for the source tokens inside the mask, which extracts source information as residual guidance and injects it into the prediction of masked tokens during regeneration. However, the source residual is injected with a fixed coefficient across all masked tokens and generation steps.



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

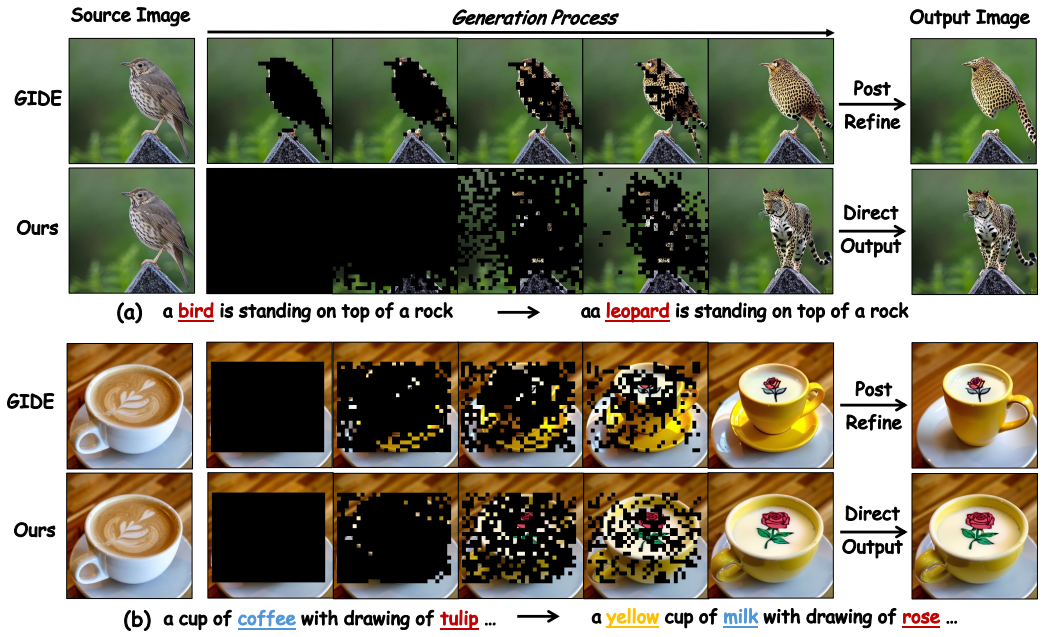


Figure 1: Comparison of GIDE and our method. The generation process visualizes steps 0, 16, 32, 48, 64 of a 64-step generation. (a) GIDE uses an extra mask from SAM, while we directly start from fully masked tokens. (b) Both use a bounding box mask for the start. Our method directly takes the final generated result as the output image, while GIDE requires further post-refinement process.

Such static guidance leads to degraded and unstable source consistency, making the model struggle with a wide range of editing scenarios, *e.g.*, complex edits, large edit areas, or global style changes. As shown in Figure 1(b), using a coarse bounding box can reduce the reliance on accurate segmentation and alleviate boundary constraints. Yet, the large mask leads to poor source consistency under complex editing targets (*e.g.*, multiple object and attribute changes). Moreover, the method cannot further identify fine-grained edit areas, while its post-refinement remains constrained by the initial editing results. **3) Rigid Generation Trajectory.** In addition, the existing method follows an inversion-then-generation pipeline, where complete inversion trajectory of source image is precomputed before editing. The decoding order of masked tokens is determined beforehand according to prediction confidence scores during inversion, regardless of the target prompt. Such pre-defined and fixed generation trajectories may further weaken the ability of dMLLMs to dynamically adapt the decoding process according to different editing targets. **Overall**, existing dMLLM-based image editing remains insufficiently flexible and faithful, restricting its practicality for broader usage.

To address these limitations, we propose **DynEdit**, a general training-free dMLLM-based framework that better exploits the intrinsic properties of dMLLMs with dynamic masked generation, enabling more flexible and effective editing. Instead of treating the edit area as a fixed mask with static source guidance and pre-defined generation order, our key idea is to let the dMLLMs *dynamically* decide *where* to edit, *how* to preserve source information, and *which* tokens to decode.

Specifically, our DynEdit builds upon a parallel dual-branch framework. The target branch performs masked generation conditioned on the target prompt, while the source branch simultaneously reconstructs the source image under the source prompt and extracts source guidance. Unlike conventional inversion-then-generation that precomputes and fixes the complete generation trajectory before editing, our framework synchronizes source residual extraction with the target generation process, which flexibly enables dynamic adaptation of the generation trajectory. Based on this, we introduce dynamic masked generation with three key designs. **1) Dynamic Edit Localization.** Instead of solely relying on a predefined edit mask, we identify editable tokens by comparing the prediction of two branches at each step. This enables automatic token-level edit localization based on target-aware semantic discrepancies, meanwhile naturally evolves with the generation process. **2) Adaptive Source Guidance.** Rather than injecting the source residual with a fixed coefficient, we dynamically adjust its strength across editable regions and generation steps, enabling faithful source preservation. **3) Edit-aware Token Decoding.** Instead of pre-defining the decoding order or

decoding masked tokens solely according to confidence scores, we dynamically adjust the decoding order according to specific editing targets, *e.g.*, prioritizing background or structural tokens earlier to provide richer contextual information, leading to more reliable and target-aware editing. Together, these designs form a dynamic editing framework that better aligns with the local and progressive generation nature of dMLLMs. As shown in Figure 1(a), without external masks, DynEdit can automatically localize the edit area during generation, first decoding the tokens of non-edit background, then generating the target object with flexible shape changes. Furthermore, given a coarse mask as start in Figure 1(b), DynEdit can further localize fine-grained edit areas, successfully achieving complex editing targets and faithful source preservation, without the need of any post-refinement.

We evaluated DynEdit across diverse benchmarks and settings. Experimental results have indicated: 1) Without external masks, DynEdit can automatically localize edit targets and achieve promising performance. 2) When external groundings are available, DynEdit can achieve more efficient editing with fewer steps. 3) Moreover, DynEdit yields improvements in more challenging scenarios, demonstrating its generalizability for broader applications. Our contributions are as follows:

- We propose DynEdit, a general training-free framework for dMLLM-based image editing that better exploits the intrinsic local and progressive generation mechanism of dMLLMs.
- We introduce a parallel dual-branch architecture that enables dynamic masked generation with three key designs, enabling flexible localization, faithful preservation, and reliable generation.
- Extensive experiments across diverse benchmarks and settings demonstrate that DynEdit consistently improves editing quality, efficiency, and robustness under various editing scenarios.

## 2 Related Work

**Diffusion Multimodal Large Language Models.** Recent advances in diffusion language models (dLLMs) (Ye et al., 2025; Nie et al., 2026; Zhu et al., 2026a) have demonstrated the effectiveness of iterative masked token prediction for efficient discrete generation. Building upon this paradigm, Diffusion Multimodal Large Language Models (dMLLMs) further unify text and image modeling within a shared discrete token space, enabling both multimodal understanding and generation in a single framework in parallel. This unique discrete formulation and non-autoregressive manner substantially enhance generation speed compared to previous unified autoregressive (AR) (Team, 2024; Liu et al., 2024; Wu et al., 2025c;b; Chen et al., 2025) or hybrid AR-diffusion (Xie et al., 2025; Zhou et al., 2025) models. Representative dMLLMs (Shi et al., 2025; Xin et al., 2025; Yang et al., 2026) have shown strong capabilities across text-to-image generation and visual understanding.

**Text-based Image Editing.** Training-free text-based image editing has been extensively studied for continuous diffusion and flow-based generative models (Rombach et al., 2022; Labs, 2024; Esser et al., 2024), where latent inversion and attention control have become standard paradigms (Hertz et al., 2023; Rout et al., 2024; Wang et al., 2024). In contrast, image editing for discrete dMLLMs remains largely underexplored. GIDE (Zhu et al., 2026b) presents the first training-free framework for dMLLM-based image editing by combining external grounding models with discrete inversion. Meanwhile, representative discrete editing methods (He et al., 2026; Dao et al., 2025; El-Ghoussani et al., 2026) also leverage discrete inversion or guidance for image editing. However, these methods are developed for masked generative models or next-scale visual autoregressive models (Tian et al., 2024; Voronov et al., 2024), therefore cannot fully exploit the capabilities of dMLLMs for editing.

## 3 Method

### 3.1 Preliminaries

**dMLLM-based Text-to-Image Generation.** It is formulated as iterative masked token prediction over a discrete vocabulary. Given a text prompt $\mathbf{P}$, the generation process starts from a fully masked image token sequence $\mathbf{x}_0$ and proceeds for $T$ iterative steps. At each step $t$, the current partially decoded image token sequence $\mathbf{x}_t$ contains both previously decoded tokens and remaining masked tokens. Following MaskGIT (Chang et al., 2022), the generation mainly consists of three operations. *1) Prediction.* The model predicts token logits for all masked positions conditioned on the prompt:

$$\hat{\mathbf{y}}_t = D_\theta(\mathbf{x}_t, \mathbf{P}). \tag{1}$$

*2) Sampling.* Candidate tokens are sampled from the predicted logits via Gumbel-Max sampling:

$$\hat{\mathbf{x}}_t = \arg\max\left(\hat{\mathbf{y}}_t + \mathbf{g}_t\right), \quad \mathbf{g}_t \sim \text{Gumbel}(0, 1), \tag{2}$$

where the probability of each sampled token is regarded as their confidence score $\mathbf{p}_t$.
*3) Remasking and Decoding.* Only high-confidence candidate tokens are fixed as decoded, while low-confidence candidate tokens are remasked for re-generation. The number of remasked tokens follows a cosine schedule $k_t = \left\lfloor N \cos\left(\frac{\pi t}{2T}\right) \right\rfloor$. The token sequence for the next step is updated by:

$$\mathbf{x}_{t+1} = \mathcal{R}(\hat{\mathbf{x}}_t, \mathbf{p}_t, k_t), \tag{3}$$

$\mathcal{R}(\cdot)$ remasks the $k_t$ lowest-confidence tokens while fixing the remaining candidate tokens together with all previously decoded tokens. Consequently, the number of masked tokens gradually decreases, yielding a localized and progressive generation process until all image tokens are decoded.

**Discrete Inversion.** Since standard continuous inversions (Song et al., 2021) cannot be directly applied, discrete inversion (Dao et al., 2025; He et al., 2026) aims to reconstruct the generation trajectory of an image within the discrete token space. Given source image token sequence $\mathbf{x}^s$ and source prompt $\mathbf{P}^{\mathbf{src}}$, the inversion progressively masks the source tokens, and extracts the residual for reconstructing the original image. At each inversion step $t$, the partially masked image token sequence $\mathbf{x}_t^s$ are fed into the dMLLM to predict the corresponding token logits for all masked tokens:

$$\hat{\mathbf{y}}_t^s = D_\theta(\mathbf{x}_t^s, \mathbf{P}^{\mathbf{src}}). \tag{4}$$

To faithfully reconstruct the source image, the location-aware argmax inversion (LAI) (Dao et al., 2025) is applied to construct the ground-truth logits $\mathbf{y}_t^s$ based on the original tokens. The discrepancy between the ground-truth logits and the model prediction is then stored as the inversion residual:

$$\mathbf{y}_t^s = \text{LAI}(\mathbf{x}_t^s, \hat{\mathbf{y}}_t^s, \mathbf{x}^s), \qquad \mathbf{z}_t = \mathbf{y}_t^s - \hat{\mathbf{y}}_t^s. \tag{5}$$

The extracted residual $\mathbf{z}_t$ captures the structural prior of the source image at each generation step and can subsequently be injected into the reverse generation process to preserve source information.

**Problem Formulation: dMLLM-based Image Editing.** Given a source image represented as a discrete token sequence $\mathbf{x}^s$, a source prompt $\mathbf{P}^{\mathbf{src}}$, and a target prompt $\mathbf{P}^{\mathbf{tgt}}$, the goal is to generate a target image that follows the target prompt while preserving editing-irrelevant source content. Following dMLLM generation paradigm, the editing process starts from an initial target token sequence $\mathbf{x}_0$ and iteratively updates it for $T$ steps. The initial sequence $\mathbf{x}_0$ can be either a fully masked token sequence, or a partially masked source token sequence when external mask is available.

### 3.2 Framework Overview

As illustrated in Figure 2, our framework built upon a parallel dual-branch architecture. The target branch performs masked generation conditioned on target prompt, while source branch simultaneously reconstructs source image under source prompt. Instead of performing rigid inversion-then-generation, we synchronize source residual extraction with target generation process. This enables dynamic adaptation of the generation trajectory to evolve with specific editing targets. Consequently, we introduce dynamic masked generation with three key modules, where the dMLLM continuously determines *where* to edit, *how* to preserve source information, and *which* tokens should decode.
**Parallel Residual Extraction.** At each generation step $t$, given the current target token sequence $\mathbf{x}_t$, we construct a partially masked source sequence $\mathbf{x}_t^s$ by aligning the source branch with the current target generation state. Specifically, positions that remain masked in $\mathbf{x}_t$ are also kept masked in $\mathbf{x}_t^s$, while positions that have been decoded are replaced by their corresponding source image tokens. The source branch then reconstructs the source image under the source prompt, following the location-aware argmax inversion introduced in Sec. 3.1 to extract source residual as:

$$\hat{\mathbf{y}}_t^s = D_\theta(\mathbf{x}_t^s, \mathbf{P}^{\mathbf{src}}), \qquad \mathbf{z}_t = \text{LAI}(\mathbf{x}_t^s, \hat{\mathbf{y}}_t^s, \mathbf{x}^s) - \hat{\mathbf{y}}_t^s. \tag{6}$$

This residual captures source-specific structural information under the same masking pattern as the current target generation state, and is therefore suitable for step-wise source guidance.

### 3.3 Dynamic Edit Localization

To determine *where* to edit, a reliable localization strategy should satisfy two properties. 1) It should evolve with the progressive generation process. Since the semantic structure of the dMLLM-based

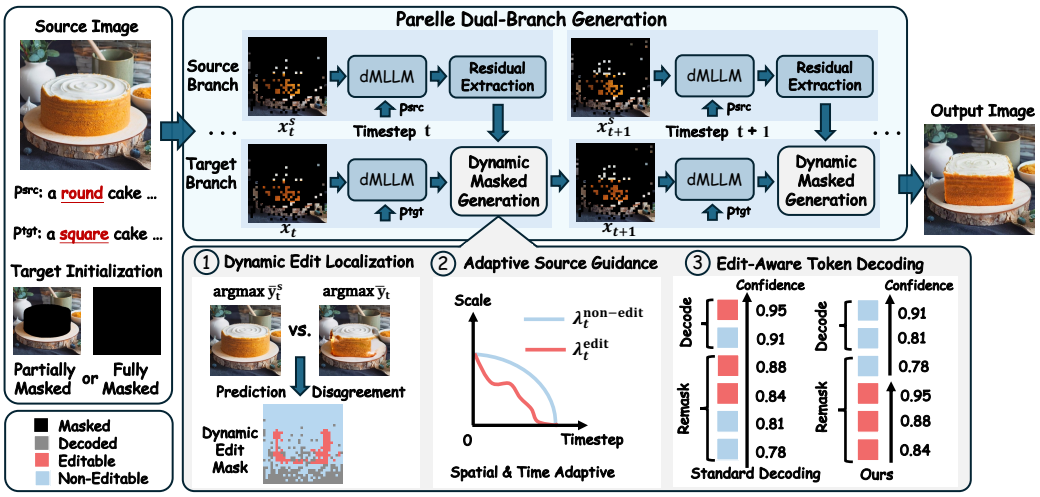


Figure 2: Framework Overview. It builds on a parallel dual-branch architecture with three key designs for dynamic masked generation. We take the fully masked initialization as an example here.

generation is continuously refined, it is infeasible to determine in advance which regions are editable, especially for structural edits. 2) Localization should be source-aware. For edits involving attribute or appearance changes, editable regions should be identified based on the semantic correspondence between the source image and the target prompt, thereby preserving the original object structure.

Therefore, instead of relying on a predefined edit mask, we dynamically identify editable tokens from the predictions of the source and target branches at each generation step. The key observation is that editing-irrelevant tokens tend to produce consistent predictions under the two prompts, whereas editable tokens are more likely to exhibit prediction disagreement due to their semantic differences. Specifically, given the current target token sequence $\mathbf{x}_t$, the target branch predicts:

$$\hat{\mathbf{y}}_t = D_\theta(\mathbf{x}_t, \mathbf{P}^{\mathbf{tgt}}). \tag{7}$$

We first inject the same residual guidance into both source and target predictions:

$$\bar{\mathbf{y}}_t = \hat{\mathbf{y}}_t + \lambda_t^{\text{loc}}\mathbf{z}_t, \qquad \bar{\mathbf{y}}_t^s = \hat{\mathbf{y}}_t^s + \lambda_t^{\text{loc}}\mathbf{z}_t, \tag{8}$$

where $\lambda_t^{\text{loc}}$ is the localization guidance scale. We then compare the source-aware sampled token predictions from the two branches at each masked position:

$$\mathbf{m}_t = \mathbb{I}\left[\arg\max \bar{\mathbf{y}}_t \neq \arg\max \bar{\mathbf{y}}_t^s\right]. \tag{9}$$

Here, $\mathbf{m}_t$ denotes the dynamic edit mask. A token with different source and target predictions is regarded as editable, while a token with consistent predictions is treated as a non-editable token. When an external mask is provided, the same strategy can be applied inside the candidate edit area, allowing the model to further distinguish fine-grained editable tokens within a coarse mask.

### 3.4 Adaptive Source Guidance

After identifying editable regions, we adapt source guidance accordingly. The adaptive guidance follows two principles. **1) Spatial-adaptive.** Non-editable tokens receive stronger guidance to preserve the source content, whereas editable tokens require relatively weaker guidance to maintain editing flexibility. **2) Time-adaptive.** The guidance strength gradually decreases throughout the generation. Stronger guidance at early stages helps establish reliable background and structural context, while weaker guidance at later stages better accommodates target modifications. We also let the source-aware localization guidance $\lambda_t^{\text{loc}}$ follow the same time-adaptive schedule, where early-stage stronger guidance stabilize source-target comparison, later-stage weaker guidance better reflects the evolving target generation. Following the above principles, we formulate the guidance scales as:

$$\lambda_t^{\text{loc}} = \lambda_0 f(t,T), \qquad \lambda_t^{\text{non-edit}} = \lambda_0 f(t,T), \qquad \lambda_t^{\text{edit}} = \lambda_0 f'(t,T), \tag{10}$$

where $\lambda_0$ denotes the initial guidance scale, $f(\cdot)$ and $f'(\cdot)$ are decay schedules for non-editable and editable regions, respectively. The final target logits are computed by combining source residual

guidance and Gumbel noise in an element-wise manner:

$$\tilde{\mathbf{y}}_t = \hat{\mathbf{y}}_t + \boldsymbol{\alpha}_t \odot \mathbf{z}_t + (1 - \boldsymbol{\alpha}_t) \odot \mathbf{g}_t, \qquad \boldsymbol{\alpha}_t = (1 - \mathbf{m}_t)\lambda_t^{\text{non-edit}} + \mathbf{m}_t \lambda_t^{\text{edit}}. \tag{11}$$

Compared with fixed residual injection, the proposed formulation makes the source guidance step-dependent, edit-aware, and adaptive to different generation processes, enabling a better balance between faithful source preservation and flexible target editing.

**Practical Choices.** The above formulation and design principles define a general adaptive guidance framework, which can be flexibly instantiated with specific schedules for different settings:

$$f(t,T) = \begin{cases} \cos\left(\frac{\pi t}{2T}\right), & \text{w/o mask,} \\ \exp\left(-\beta \frac{t}{T}\right), & \text{w/ mask,} \end{cases} \qquad f'(t,T) = f(t,T)(1 - r_t), \quad r_t = \frac{1}{|\mathcal{M}_t|} \sum_{i \in \mathcal{M}_t} m_{t,i}, \tag{12}$$

where $\mathcal{M}_t$ denotes the set of masked tokens at step $t$, and $r_t$ is the edit ratio estimated from the dynamic edit mask. For editing without external groundings, the generation starts from a fully masked token sequence without reliable visual context. Therefore, a relatively slow cosine decay is adopted following the standard decoding behavior of dMLLMs to preserve sufficient source information during early generation. The editable-token guidance is further modulated by $(1 - r_t)$ for spatial-adaptive. When external groundings are available for initialization, the preserved non-masked source tokens already provide reliable visual context. Therefore, we can reduce the initial guidance according to the mask size (*e.g.*, $\lambda_0' = r_m \lambda_0$, where $r_m$ denotes ratio of mask area), and adopt faster decay schedule to allow stronger or more efficient editing within the candidate edit area. It is worth noting that this framework is not restricted to these specific choices. As long as following the proposed design principles, other monotonic time-adaptive schedules (*e.g.*, linear decay) and different spatial-adaptive designs can also be readily adopted[1].

### 3.5 Edit-aware Token Decoding

Standard dMLLM decodes tokens according to prediction confidence, where high-confidence tokens are fixed and low-confidence tokens are remasked. Such confidence-only decoding may be suboptimal for editing, since editing-relevant and editing-irrelevant areas play different roles during progressive generation. In particular, stabilizing source-consistent background and structure at early stages can provide more reliable visual context for subsequently generating editable content. We thus introduce edit-aware token decoding to jointly consider prediction confidence and edit area. Based on the target logits $\tilde{\mathbf{y}}_t$, we sample candidate tokens $\hat{\mathbf{x}}_t = \arg\max(\tilde{\mathbf{y}}_t)$ and obtain their confidence scores $\mathbf{p}_t$. Instead of ranking all tokens solely by confidence, we perform a hierarchical ranking for remasking. Specifically, tokens are first partitioned into editable and non-editable groups according to the dynamic edit mask $\mathbf{m}_t$, where non-editable tokens are always ranked ahead of editable tokens. Within each group, tokens are further ordered by their prediction confidence. Consequently, editable tokens are preferentially remasked, while non-editable tokens are decoded earlier to establish reliable background and structural context for subsequent editing:

$$\mathbf{x}_{t+1} = \mathcal{R}(\hat{\mathbf{x}}_t, \mathbf{p}_t, k_t, \mathbf{m}_t), \tag{13}$$

This strategy jointly considers prediction confidence and editing awareness, allowing source-consistent content to decode earlier while keeping editable regions flexible for later refinement.

## 4 Experiments

We evaluate the effectiveness of our method by discovering three research questions:

- **Q1.** Can our method perform faithful dMLLM-based image editing *without* relying on external grounding models or edit masks?
- **Q2.** When external grounding information is available, what additional advantages does our method provide over existing dMLLM-based approaches?
- **Q3.** Can our method generalize to more challenging editing scenarios?

### 4.1 Experimental Setup

**Implementation Details.** We adopt Lumina-DiMOO (Xin et al., 2025) and MMaDA (Yang et al., 2026) as the pre-trained dMLLM backbone. We use the stronger Lumina-DiMOO for most experi-

Table 1: Quantitative comparison of different methods on PIE-Bench and PIE-Bench++. Dist., LPIPS, SSIM are scaled $\times 10^2$, and MSE is scaled $\times 10^3$ for readability. Best results in **bold** and second-best results are underlined within each dataset. Inference Time (IS) tested on the H100 GPU.

| Method | Target Alignment | | Background Preservation | | | | Global Preservation | | | | | IS(s)↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CLIP-G ↑ | CLIP-E ↑ | PSNR ↑ | LPIPS ↓ | MSE ↓ | SSIM ↑ | Dist. ↓ | PSNR ↑ | LPIPS ↓ | MSE ↓ | SSIM ↑ | |
| ***PIE-Bench*** | | | | | | | | | | | | |
| DICE | 24.70 | 21.76 | 9.01 | 48.89 | 160.70 | 47.00 | 17.84 | 6.71 | 74.41 | 234.50 | 21.64 | 24.5 |
| VARIN | 24.22 | 21.42 | 20.74 | 12.30 | 16.88 | 74.93 | 4.45 | 16.90 | 23.22 | 35.11 | 57.43 | 24.1 |
| MLN | 25.12 | 22.03 | 20.81 | 12.75 | 17.66 | 75.07 | 4.09 | 17.14 | 22.08 | 29.47 | 59.18 | 101.6 |
| GIDE | 24.70 | 22.17 | **24.37** | **7.86** | **8.36** | **79.24** | 3.84 | 17.89 | 19.85 | 30.48 | 61.45 | 31.1 |
| Ours | **25.27** | **22.21** | 22.52 | 9.67 | 9.45 | 77.34 | **2.81** | **18.24** | **18.06** | **19.24** | **61.80** | **21.3** |
| ***PIE-Bench++*** | | | | | | | | | | | | |
| DICE | 24.86 | 24.12 | 9.68 | 44.54 | 143.70 | 51.61 | 18.02 | 6.56 | 75.25 | 241.90 | 21.03 | 23.2 |
| VARIN | 23.99 | 23.24 | 21.14 | 11.35 | 18.52 | 77.01 | 5.33 | 15.86 | 26.37 | 42.74 | 54.82 | 23.1 |
| MLN | 25.11 | 24.28 | 21.89 | 10.93 | 13.07 | 77.72 | 4.25 | 16.62 | 23.66 | 30.84 | 57.82 | 105.1 |
| GIDE | **25.43** | **24.77** | 23.01 | 9.57 | 13.43 | 79.15 | 6.87 | 14.81 | 29.43 | 54.31 | 55.23 | 36.8 |
| Ours | 25.25 | 24.33 | **23.04** | **8.88** | **8.72** | **79.20** | **3.16** | **17.75** | **19.51** | **22.53** | **60.97** | **21.3** |

ments, and adopt 64 generation steps for editing. The initial source guidance scale is set to 0.7. [1]

**Benchmarks and Metrics.** We adopt PIE-Bench (Ju et al., 2024), PIE-Bench++ (Huang et al., 2024), and GIDE-Bench (Zhu et al., 2026b), covering both standard and complex image editing[1]. Following previous work, we evaluate editing quality from two perspectives: *target alignment*, measured by CLIP similarity (Radford et al., 2021) on full image (CLIP-G) and edited region (CLIP-E), and *source preservation*, measured by PSNR, LPIPS (Zhang et al., 2018), MSE, SSIM (Wang et al., 2004), and structure distance (Tumanyan et al., 2022) on both non-edited regions and full image.

**Baselines.** We compare our method with GIDE (Zhu et al., 2026b), the first training-free image editing framework designed for dMLLMs. To further provide a fair comparison with existing discrete editing methods, we additionally re-implement representative methods DICE (He et al., 2026), VARIN (Dao et al., 2025) and MLN (El-Ghoussani et al., 2026) on the same dMLLM backbone.

### 4.2 Q1: Editing without External Grounding

**Setting.** We first investigate whether our method can achieve effective editing without relying on external masks. Specifically, our method and all discrete editing baselines start from fully masked tokens. While GIDE utilizes SAM-generated masks to preserve non-edited regions for initialization.

**Results.** As shown in Table 1, we observe: **1)** Compared with existing discrete editing methods without external grounding, our method consistently achieves the best overall performance on both benchmarks. **2)** Compared with GIDE, which relies on external grounding models, our method achieves a better trade-off between editability and source preservation. On PIE-Bench, GIDE obtains stronger background preservation by freezing tokens outside the SAM mask, but its fixed edit mask limits target alignment and preservation within edited regions (*c.f.*, Figure 3(a), rows 1–2). In contrast, our method achieves the highest target alignment together with the best global preservation. The advantage becomes more evident on PIE-Bench++, where larger and more complex edits make GIDE more susceptible to inaccurate localization and over-editing (*c.f.*, Figure 3(a), rows 3–4). Although GIDE still achieves slightly better target alignment, our method consistently preserves both background and global content while maintaining comparable editing quality. **3)** Moreover, our method eliminates the need for external grounding and post-refinement, resulting in simpler and more efficient inference. Overall, these results demonstrate our framework can automatically localize editable regions and perform effective dMLLM-based image editing without external grounding.

### 4.3 Q2: Editing with External Grounding

**Setting.** We further investigate whether our framework can benefit from external grounding, which provides a natural initialization for dMLLM-based image editing. For a fair comparison, we use the same edit masks as GIDE: the official ground-truth masks on PIE-Bench and merged grounding

[1] Due to the limited space, more details are left in the Appendix.

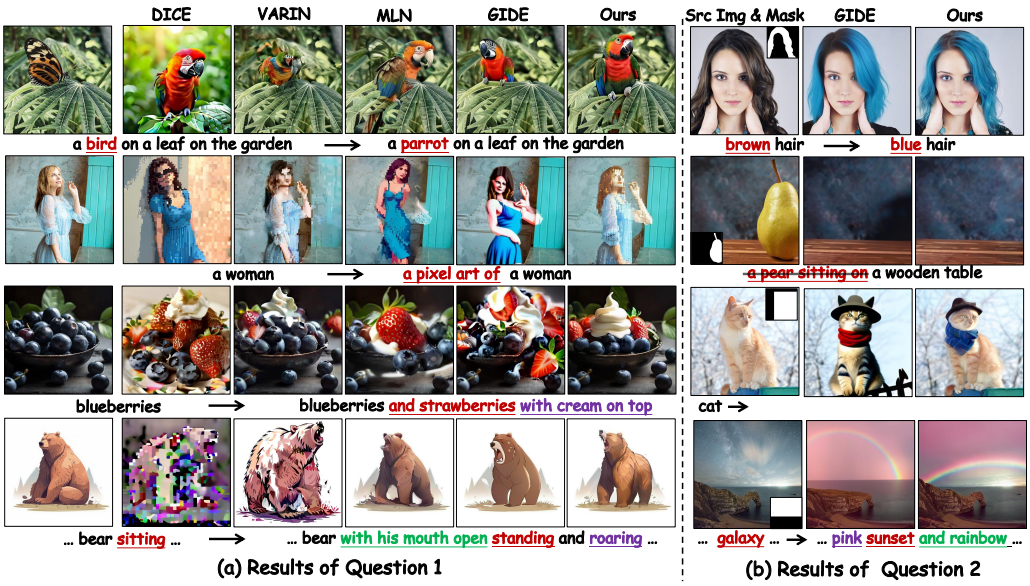


Figure 3: Qualitative comparisons on PIE-Bench and PIE-Bench++. (a) Q1: GIDE uses SAM masks while other methods don't use masks. (b) Q2: Both methods use the same external masks.

bounding boxes on PIE-Bench++. We further evaluate different generation budgets (8, 16, 32, and 64 steps) to study whether external grounding can accelerate the editing process.

**Results.** Figure 4 reports the results under different generation steps. We observe: **1)** With only 8 generation steps, both methods suffer from insufficient denoising. Benefiting from post-refinement stage, GIDE achieves slightly better target alignment, although the overall editing quality remains unsatisfactory. **2)** From 16 steps onward, our method consistently outperforms GIDE on both benchmarks. By progressively refining editable regions instead of strictly following the initial mask, our framework achieves more accurate localization while better preserving source content (*c.f.*, Figure 3(b), row 3). **3)** More importantly, our method with only 16 steps already surpasses GIDE with 64 steps on all metrics, while performance nearly saturates at 32 steps. These results demonstrate that external grounding provides a good initialization, whereas our framework further exploits the localized progressive generation mechanism of dMLLMs for more accurate and efficient editing.

### 4.4 Q3: Generalization to More Complex Editing Scenarios

**Setting.** Finally, we evaluate the generalization ability of our method on the more challenging GIDE-Bench (Zhu et al., 2026b), which involves compositional image editing with diverse multimodal inputs and various image resolutions. Following GIDE, we compare representative training-based editing methods (Labs et al., 2025; Wu et al., 2025a), and adopt Qwen3-VL (Bai et al., 2025) as the MLLM evaluator to assess semantic correctness (SC) and perceptual quality (PQ).

**Results.** As reported in Table 2, our method consistently outperforms GIDE on most evaluation metrics. More importantly, despite requiring no additional training, our method further narrows the performance gap with leading training-based approaches. These results demonstrate that our framework generalizes effectively to more complex editing scenarios.

Table 2: Comparison on the GIDE-Bench.

| Method | MSE ↓ | PSNR ↑ | SSIM ↑ | SC ↑ | PQ ↑ |
|---|---|---|---|---|---|
| *Training-based* | | | | | |
| FLUX-Kontext | 26.09 | 16.18 | 55.9 | 4.76 | 4.64 |
| Qwen-Image-Edit | **22.50** | **19.27** | **74.4** | **4.89** | **4.80** |
| *Training-free* | | | | | |
| GIDE | 20.99 | 17.70 | **63.5** | 4.39 | 4.29 |
| **Ours** | **16.95** | **18.00** | 63.3 | **4.56** | **4.50** |

### 4.5 Ablation Study

**Model Generalization.** We implement our framework on MMaDA. As shown in Table 3 (a), the results follow a similar trend to Lumina: GIDE achieves higher background preservation, while ours outperforms on both target alignment and global preservation, demonstrating good generalization across different dMLLM backbones. Besides, as backbone capabilities differ, the editing trade-off

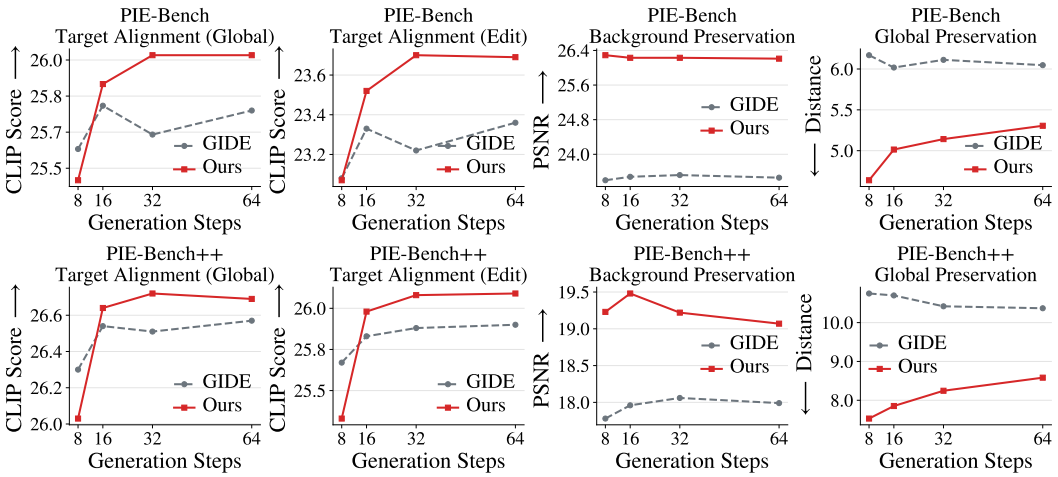


Figure 4: Comparison with GIDE with same external groundings. We use the official ground truth mask for PIE-Bench, and the grounded bounding box mask for PIE-Bench++.

varies, *i.e.*, MMaDA produces stronger target alignment but relatively weaker source preservation.

**Guidance Scale.** Table 3(b) shows larger scale favors source preservation, while smaller one improves target alignment, providing flexible control over editing trade-off. We use same scale (0.7) across all experiments and benchmarks, demonstrating our framework is robust to this parameter.

**Effectiveness of Key Modules.** Table 3(c) verifies the effectiveness of all three modules. Removing dynamic localization (**M1**) prevents identification of editable tokens for subsequent spatial-adaptive guidance and edit-aware decoding, reducing editability. Without adaptive guidance (**M2**), fixed strong guidance over-preserves the source content. Without edit-aware decoding (**M3**), editable tokens can be reconstructed too early under strong early-stage guidance, also limiting editability. Together, the three modules form a dynamic editing framework enabling faithful and flexible dMLLM-based image editing.

Table 3: Ablation Study. All ablations were conducted on the PIE-Bench without external groundings[1]. C-G/C-E denote global/edit CLIP scores, and P-bg denotes background PSNR.

**(a) Generalization on MMaDA**

| Method | C-G | C-E | P-bg | Dist. |
|---|---|---|---|---|
| GIDE | 24.96 | 22.29 | **24.07** | 4.56 |
| Ours | **25.48** | **22.48** | 22.00 | **3.27** |

**(b) Guidance Scale**

| $\lambda_0$ | C-G | C-E | P-bg | Dist. |
|---|---|---|---|---|
| 0.6 | **25.77** | **22.75** | 21.66 | 3.44 |
| 0.7 | 25.27 | 22.21 | 22.52 | 2.81 |
| 0.8 | 24.92 | 21.85 | **23.28** | **2.45** |

**(c) Key Modules**

| Method | C-G | C-E | P-bg | Dist. |
|---|---|---|---|---|
| Ours | 25.27 | 22.21 | 22.52 | 2.81 |
| w/o M1 | 22.83 | 20.07 | 27.12 | 0.90 |
| w/o M2 | 23.02 | 20.13 | 26.93 | 1.00 |
| w/o M3 | 22.83 | 20.14 | 27.11 | 0.93 |

**(d) Guidance Schedule**

| Schedule | C-G | C-E | P-bg | Dist. |
|---|---|---|---|---|
| $\text{Cosine}_{0.7}$ | 25.27 | 22.21 | 22.52 | 2.81 |
| $\text{Linear}_{0.7}$ | 26.16 | 23.07 | 20.04 | 4.65 |
| $\text{Linear}_{0.9}$ | 25.38 | 22.33 | 22.34 | 2.99 |
| $\text{Comb}_{0.7}$ | 25.25 | 22.23 | 22.39 | 2.95 |

**Guidance Schedule.** As discussed in Sec. 3.4, our proposed adaptive source guidance is compatible with different schedules. Table 3 (d) shows that, compared with default cosine schedule, a faster linear decay improves target alignment but weakens source preservation. Increasing the initial guidance scale recovers a more balanced trade-off. We also evaluate a combined schedule for spatial-adaptive guidance, where non-editable tokens use cosine decay while editable tokens use faster linear decay, which achieves performance close to the edit-ratio modulation. These results suggest that our framework provides a flexible and interpretable guidance principle.

## 5 Conclusion

In this paper, we present DynEdit, a general training-free framework for dMLLM-based image editing. DynEdit introduces dynamic masked generation that continuously determines where to edit, how to preserve source information, and which tokens to decode throughout the generation process. Extensive experiments demonstrate that our framework consistently improves editing quality, source preservation, and generation efficiency under both mask-free and external grounding settings, while further generalizing to more challenging compositional editing scenarios. We hope this work can serve as a simple yet effective dMLLM-based baseline, and plan to extend the dynamic masked generation framework to more general multimodal generation tasks in the future.

## AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing and improving the readability of the manuscript. All technical ideas, analyses, experiments, and conclusions were conceived, implemented, and validated by the authors. All AI-assisted text was carefully reviewed by the authors to ensure its accuracy and correctness. The authors take full responsibility for the final content of this work.

## ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. DynEdit is a training-free image editing framework built upon pretrained dMLLMs and evaluated on public benchmarks, without the use of private or user-specific data. Our research does not involve human subjects, private or sensitive data, or personally identifiable information. We are not aware of any direct ethical concerns specific to this work beyond those inherent to the underlying pretrained models and image editing applications.

## REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed implementation settings, hyperparameters, benchmarks, evaluation protocols, and baseline configurations in Appendix B. Additional ablation studies are reported in Appendix D, and the complete inference pipeline is provided in Appendix E.

## REFERENCES

Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, volume 2025, pp. 50726–50753, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, et al. Llada2. 0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*, 2025.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Quan Dao, Xiaoxiao He, Ligong Han, Ngan Hoai Nguyen, Amin Heyrani Nobar, Faez Ahmed, Han Zhang, Viet Anh Nguyen, and Dimitris Metaxas. Discrete noise inversion for next-scale autoregressive text-based image editing. *arXiv preprint arXiv:2509.01984*, 2025.

Amir El-Ghoussani, Marc Hölle, Gustavo Carneiro, and Vasileios Belagiannis. Prompt-guided image editing with masked logit nudging in visual autoregressive models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4810–4820, June 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Xiaoxiao He, Quan Dao, Ligong Han, Song Wen, Minhao Bai, Di Liu, Han Zhang, Felix Juefei-Xu, Chaowei Tan, Bo Liu, et al. Dice: Discrete inversion enabling controllable editing for masked generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 762–772, 2026.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

Mingzhen Huang, Jialing Cai, Shan Jia, Vishnu S Lokhande, and Siwei Lyu. Paralleledits: Efficient multi-aspect text-driven image editing with attention grouping. *Advances in Neural Information Processing Systems*, 37:22569–22595, 2024.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *International Conference on Learning Representations*, volume 2024, pp. 23395–23422, 2024.

Jeongsol Kim, Yeobin Hong, and Jong Chul Ye. Flowalign: Trajectory-regularized, inversion-free flow-based image editing. *arXiv preprint arXiv:2505.23145*, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng Yan. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Ye Tian, Ling Yang, Jiongfan Yang, Anran Wang, Yu Tian, Haochen Wang, Zhiyang Teng, Zhuochen Wang, Yinjie Wang, Yunhai Tong, et al. Mmada-parallel: Multimodal large diffusion language models for thinking-aware editing and generation. In *International Conference on Learning Representations*, volume 2026, pp. 36089–36123, 2026.

Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10748–10757, 2022.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1921–1930, 2023.

Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

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

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025b.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 93620–93638, 2025c.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

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

Yi Xin, Siqi Luo, Tianxiang Xu, Qi Qin, Haoxing Chen, Kaiwen Zhu, Zhiwei Zhang, Yangfan He, Rongchao Zhang, Jinbin Bai, et al. dmllm-tts: Self-verified and efficient test-time scaling for diffusion multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35726–35735, 2026.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026.

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11425–11460, 2026a.

Tianrui Zhu, Shiyi Zhang, Jiawei Shao, and Yansong Tang. Kv-edit: Training-free image editing for precise background preservation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16607–16617. IEEE, 2025.

Zifeng Zhu, Jiaming Han, Jiaxiang Zhao, Minnan Luo, and Xiangyu Yue. Gide: Unlocking diffusion llms for precise training-free image editing. *arXiv preprint arXiv:2603.21176*, 2026b.

# APPENDIX

The Appendix is organized as follows:

- In Sec. A, we provide more detailed related works.
- In Sec. B, we show more experimental setup details.
- In Sec. C, we discuss and compare with training-free flow-based image editing methods.
- In Sec. D, we provide more ablation studies about our framework.
- In Sec. E, we show the inference scheme.
- In Sec. F, we provide more visualization results.
- In Sec. G, we discuss the failure cases, limitations and potential future works.

## A DETAILED RELATED WORK

**Diffusion Multimodal Large Language Models.** Recent advances in diffusion language models (dLLMs) (Arriola et al., 2025; Ye et al., 2025; Bie et al., 2025; Nie et al., 2026; Zhu et al., 2026a) have demonstrated the effectiveness of iterative masked token prediction for efficient discrete generation. Instead of autoregressive left-to-right decoding, dLLMs iteratively predict masked tokens in parallel and progressively reveal reliable predictions through multiple denoising steps. Building upon this paradigm, Diffusion Multimodal Large Language Models (dMLLMs) (Shi et al., 2025; Xin et al., 2025; Yang et al., 2026; Tian et al., 2026; Xin et al., 2026) further unify text and image modeling within a shared discrete token space, enabling multimodal understanding and generation in a single framework. Compared with previous unified autoregressive (AR) (Team, 2024; Liu et al., 2024; Wu et al., 2025c;b; Chen et al., 2025) or hybrid AR-diffusion models (Xie et al., 2025; Zhou et al., 2025), dMLLMs substantially improve generation efficiency through parallel discrete generation. Representative models, including MUDDIT (Shi et al., 2025), Lumina-DiMOO (Xin et al., 2025), and MMaDA (Yang et al., 2026), demonstrate strong capabilities in both multimodal understanding and generation. During inference, dMLLM-based text generation typically follows the partially autoregressive decoding strategy of LLaDA (Zhu et al., 2026a), whereas dMLLM-based image generation adopts the MaskGIT-style (Chang et al., 2022) iterative prediction–sampling–remasking process, resulting in a localized and progressive generation mechanism.

**Text-based Image Editing.** Text-based image editing aims to modify a source image according to a target prompt while preserving editing-irrelevant visual content. Benefiting from the rapid development of large-scale diffusion and flow-based text-to-image models (Rombach et al., 2022; Labs, 2024; Esser et al., 2024), training-free image editing has attracted extensive attention by directly exploiting the priors of pre-trained generative models. Existing methods mainly fall into three categories: 1) Latent inversion (Mokady et al., 2023; Ju et al., 2024; Rout et al., 2024; Wang et al., 2024), which first inverts the source image into the generative trajectory before editing via ODE inversion (Song et al., 2021; Lipman et al., 2023). 2) Attention or feature control (Hertz et al., 2023; Tumanyan et al., 2023), which preserves source layouts by manipulating intermediate attention maps or features during generation, and more recently, 3) Inversion-free flow-based editing (Kulikov et al., 2024; Kim et al., 2025), which directly transports the source image toward the target through the learned velocity field. In contrast, image editing for discrete dMLLMs remains largely underexplored. GIDE (Zhu et al., 2026b) presents the first training-free framework for dMLLM-based image editing by combining external grounding models with discrete inversion, where source residuals extracted from inversion are injected during masked generation. Meanwhile, representative discrete editing methods such as DICE (He et al., 2026), VARIN (Dao et al., 2025), and MLN (El-Ghoussani et al., 2026) investigate discrete inversion, location-aware reconstruction, and logit-level guidance in the discrete token space. However, these methods are mainly developed for masked generative models or next-scale visual autoregressive models (Tian et al., 2024; Voronov et al., 2024), and therefore cannot fully exploit the capabilities of dMLLMs. In this paper, we instead focus on training-free image editing for dMLLMs by fully exploiting their intrinsic localized and progressive masked generation process, dynamically determining where to edit, how strongly to preserve source information, and which tokens to decode throughout generation.

## B Detailed Experimental Setup

**Implementation Details.** We adopt Lumina-DiMOO (Xin et al., 2025) and MMaDA (Yang et al., 2026) as the pre-trained dMLLM backbones. Following GIDE (Zhu et al., 2026b), we use the stronger Lumina-DiMOO for most experiments, while MMaDA is used to evaluate the generalization of our framework. Unless otherwise specified, all experiments use 64 generation steps for image editing. The initial source guidance scale is set to $\lambda_0 = 0.7$ across all benchmarks and settings without per-dataset tuning. For the mask-free setting, where editing starts from a fully masked image token sequence without external grounding, we adopt the default cosine decay schedule:

$$\lambda_t^{\text{non-edit}} = \lambda_0 \cos\left(\frac{\pi t}{2T}\right), \qquad \lambda_t^{\text{edit}} = \lambda_0 \cos\left(\frac{\pi t}{2T}\right)(1 - r_t), \tag{14}$$

where $r_t$ denotes the ratio of editable tokens estimated by the dynamic edit mask. The editable-token guidance is further modulated by $(1 - r_t)$ to provide weaker source guidance when larger editable regions are identified. When external grounding is available, editing starts from a partially masked source token sequence, where tokens outside the candidate edit region are directly preserved. Since these non-masked source tokens already provide reliable visual context, a weaker initial guidance together with a faster decay schedule is sufficient for faithful reconstruction while allowing more efficient editing inside the candidate edit region. Specifically, let $r_m$ denote the ratio of mask area in the external mask, we first adjust the initial guidance scale as:

$$\lambda_0' = r_m \lambda_0, \tag{15}$$

and then adopt an exponential decay schedule:

$$\lambda_t^{\text{non-edit}} = \lambda_0' \exp\left(-\beta \frac{t}{T}\right), \qquad \lambda_t^{\text{edit}} = \lambda_0' \exp\left(-\beta \frac{t}{T}\right)(1 - r_t), \tag{16}$$

where $\beta = 5$ in all experiments.

**Benchmarks and Metrics.** We evaluate our method on three representative image editing benchmarks. **PIE-Bench** (Ju et al., 2024) is the standard benchmark for text-based image editing, consisting of 700 editing samples covering diverse editing categories such as object replacement, attribute modification and style transfer. **PIE-Bench++** (Huang et al., 2024) further extends PIE-Bench with more challenging prompts involving multiple objects, multiple attributes edits. Both benchmarks contain images with a resolution of $512 \times 512$. Following GIDE (Zhu et al., 2026b), we further evaluate on **GIDE-Bench**, which contains more challenging compositional editing tasks with diverse multimodal instructions and images of various resolutions. Following previous work, we evaluate editing quality from two perspectives. *Target Alignment* is measured by CLIP similarity (Radford et al., 2021) on the entire image (CLIP-G) and the annotated edited region (CLIP-E). *Source Preservation* is evaluated by PSNR, LPIPS (Zhang et al., 2018), MSE, SSIM (Wang et al., 2004), and Structure Distance (Tumanyan et al., 2022) on both the preserved background region and the entire image. For GIDE-Bench, following GIDE's official evaluation protocol, we additionally adopt Qwen3-VL (Bai et al., 2025) as the MLLM evaluator to assess Semantic Correctness (SC) and Perceptual Quality (PQ). All inference times (IS) are measured on the same H100 GPU.

**Baselines.** We compare our method with GIDE (Zhu et al., 2026b), the first training-free image editing framework specifically designed for dMLLMs. GIDE first employs external grounding models (SAM2/SAM3) to obtain candidate editing masks, performs discrete inversion inside the masked region, injects inversion residuals during masked generation, and finally applies a post-refinement stage to improve visual consistency. To further provide a fair comparison with representative discrete image editing methods, we additionally re-implement DICE (He et al., 2026), VARIN (Dao et al., 2025), and MLN (El-Ghoussani et al., 2026) on the same dMLLM backbone. DICE and VARIN preserve source information through discrete inversion and source reconstruction, while MLN performs logit-level source guidance with attention-based localization. Since these methods are originally developed for masked generative models or next-scale visual autoregressive models, we adapt their guidance strategies to the dMLLM generation process while keeping their original designs unchanged. All methods are evaluated under the same backbone, generation steps, and evaluation protocol.

**Setting of Main Experiments.** We conduct three groups of experiments corresponding to the research questions in the main paper.

Table 4: Quantitative comparison of mask-free flow-based methods on PIE-Bench. Dist., LPIPS, SSIM are scaled $\times 10^2$, and MSE is scaled $\times 10^3$ for readability. Best results are in **bold** and second-best results are underlined.

| Method | Target Alignment | | Background Preservation | | | | Global Preservation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | CLIP-G ↑ | CLIP-E ↑ | PSNR ↑ | LPIPS ↓ | MSE ↓ | SSIM ↑ | Dist. ↓ | PSNR ↑ | LPIPS ↓ | MSE ↓ | SSIM ↑ |
| ***Continuous Flow-based*** | | | | | | | | | | | |
| Reconstruction | – | – | 37.68 | 0.79 | 0.39 | 96.47 | 0.02 | 34.23 | 1.31 | 0.64 | 94.28 |
| FlowEdit | 25.17 | **22.58** | 21.80 | 11.45 | 9.88 | **83.25** | 2.83 | 18.47 | 19.70 | 17.39 | **72.16** |
| RF-Inversion | 24.07 | 21.47 | 21.84 | 14.44 | 10.12 | 73.13 | 3.13 | **18.55** | 23.22 | **16.90** | 56.48 |
| ***Discrete dMLLM-based*** | | | | | | | | | | | |
| Reconstruction | – | – | 27.53 | 5.21 | 3.27 | 81.64 | 0.74 | 24.02 | 8.55 | 5.45 | 70.13 |
| GIDE | 24.70 | 22.17 | **24.37** | **7.86** | **8.36** | 79.24 | 3.84 | 17.89 | 19.85 | 30.48 | 61.45 |
| Ours | **25.27** | 22.21 | 22.52 | 9.67 | 9.45 | 77.34 | **2.81** | 18.24 | **18.06** | 19.24 | 61.80 |

- **Q1: Editing without External Grounding.** Our method and all discrete editing baselines (DICE, VARIN, and MLN) perform editing without using any external masks, starting from a fully masked image token sequence. GIDE follows its original setting and employs SAM2/SAM3 to generate editing masks.
- **Q2: Editing with External Grounding.** Both our method and GIDE are initialized using the same external masks. Specifically, we use the official ground-truth masks on PIE-Bench and the grounded bounding-box masks generated by SAM on PIE-Bench++, allowing a direct comparison under identical initialization.
- **Q3: Generalization to Complex Editing.** Following GIDE, we use the segmentation masks generated by SAM on GIDE-Bench. We further compare representative end-to-end training-based image editing models under the same evaluation protocol. Following GIDE, the training-free methods adopt multi-instruction step-by-step, while end-to-end training-based models process multi-instruction inputs jointly.

## C Comparison with Continuous Flow-based Methods

Our work primarily focuses on how to better exploit and empower pre-trained dMLLMs for training-free text-based image editing. Meanwhile, continuous diffusion and flow-based generative models remain the dominant paradigm for image generation and editing, with substantial progress achieved in recent years. To provide a broader comparison across different generative paradigms, we further compare our method with representative flow-based training-free editing approaches operating in the continuous latent space, including the mask-free methods FlowEdit (Kulikov et al., 2024) and RF-Inversion (Rout et al., 2024), as well as KV-Edit (Zhu et al., 2025), which additionally supports mask-grounded editing. We organize the comparison following the same three research questions as in the main experiments, covering mask-free editing, mask-grounded editing, and more challenging editing scenarios. All flow-based methods are evaluated using their official implementations with FLUX.1-dev (Labs, 2024) as the pre-trained backbone.

### C.1 Q1: Editing without External Grounding

**Setting.** We first compare with mask-free flow-based methods on PIE-Bench (Ju et al., 2024). Specifically, our method, FlowEdit, and RF-Inversion perform editing without using any external grounding or edit masks. Our method starts from a fully masked visual token sequence and dynamically identifies editable regions throughout generation, while the flow-based methods directly operate in the continuous latent space. For reference, we also report GIDE (Zhu et al., 2026b), which follows its original setting and uses SAM-generated masks to preserve non-edited regions.

**Results.** As shown in Table 4, we make the following observations. 1) Despite operating under a substantially different discrete generation paradigm, our method achieves performance comparable to representative flow-based approaches. Compared with FlowEdit, our method achieves slightly better global target alignment (25.27 vs. 25.17 CLIP-G) and comparable alignment for the edit area (22.21 vs. 22.58 CLIP-E), while providing stronger preservation on several key metrics, including background PSNR (22.52 vs. 21.80), background LPIPS (9.67 vs. 11.45), and global structure

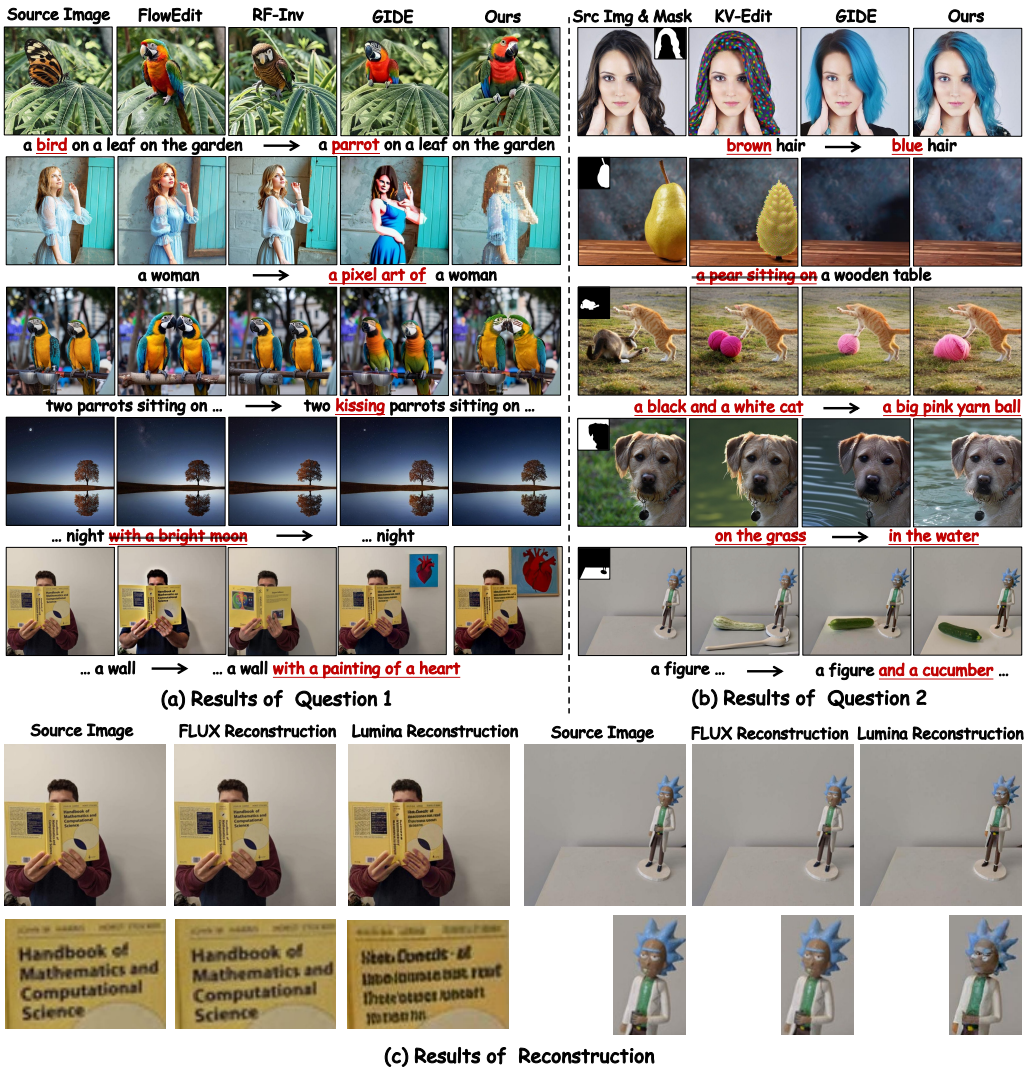


Figure 5: Qualitative comparisons with flow-based methods on PIE-Bench. (a) Q1: GIDE uses SAM masks while other methods don't use masks. (b) Q2: Both methods use the same external masks. GIDE and our method use 64 generation steps. (c) Reconstruction results of different backbones.

distance (2.81 vs. 2.83). These results demonstrate that effectively exploiting the localized and progressive generation mechanism of dMLLMs can provide competitive editing capability.

2) It is important to note that the underlying generative backbones have substantially different capacities and visual representations. The flow-based methods are built upon FLUX.1-dev, a 12B model specialized for text-to-image generation, whereas our method uses the 8B Lumina-DiMOO, a unified multimodal model in which text-to-image generation constitutes only one of multiple capabilities. More importantly, FLUX employs a continuous autoencoder, whereas Lumina-DiMOO represents images using discrete visual tokens. To quantify this representation gap, we additionally report *Reconstruction* performance by directly encoding/tokenizing and decoding each source image without performing any editing. The gap is substantial: FLUX achieves a global reconstruction PSNR of 34.23 and LPIPS of 1.31, compared with only 24.02 and 8.55 for Lumina-DiMOO, respectively. A similar gap can also be observed in background reconstruction (37.68 vs. 27.53 PSNR). Such differences reveal a considerable representation-level reconstruction bottleneck for the current dMLLM backbone, which inherently places it at a disadvantage on reconstruction-sensitive source preservation metrics. As shown in Figure 5(c), in the example of "a man holding a book", Lumina-DiMOO struggles to accurately reconstruct the rendered text, which inherently limits the achievable source preservation fidelity.

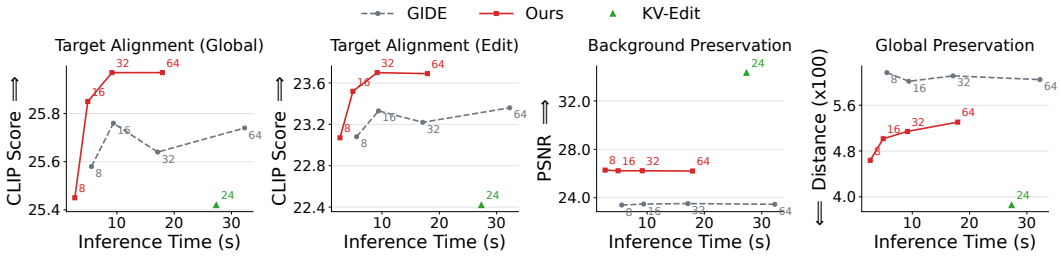


Figure 6: Comparison with flow-based method with the same external groundings on PIE-Bench. Structure Distance scaled $\times 10^2$ for readability. Numbers beside each node denote generation step.

Nevertheless, after editing, this large reconstruction gap is substantially narrowed: our method achieves a global PSNR of 18.24 compared with 18.47 for FlowEdit, while obtaining better global LPIPS (18.06 vs. 19.70) and structure distance (2.81 vs. 2.83). This contrast suggests that the editing performance of our method is not simply inherited from a stronger underlying image representation; instead, the proposed dynamic masked generation effectively preserves source information despite the more restrictive discrete representation. We expect that further advances in dMLLM backbones and discrete visual tokenizers, particularly improved reconstruction fidelity and increased model capacity, can further strengthen dMLLM-based image editing.

### C.2 Q2: Editing with External Grounding

**Setting.** We further compare with KV-Edit (Zhu et al., 2025), a representative flow-based editing method that can also utilize external masks for background preservation. For a fair comparison, all methods use the official ground-truth masks provided by PIE-Bench as the same editing initialization. Since different methods employ different generation budgets, editing procedures, and pre-trained backbones, we additionally report their inference time to provide a comprehensive comparison of editing quality and efficiency.

**Results.** As shown in Figure 6, we make the following observations. 1) Our method demonstrates substantially stronger target editing capability than KV-Edit. Even with only 16 generation steps, our method already achieves better target alignment than KV-Edit with 24 steps on both the full image and the edited region. With more generation steps, the advantage remains consistent, demonstrating that the progressive masked generation of dMLLMs can effectively exploit the external mask as an initialization while dynamically refining the editable content.

2) KV-Edit achieves stronger source preservation on pixel-level metrics. As discussed in Sec. C.1, this gap should be interpreted together with the substantial difference in reconstruction fidelity between the underlying visual representations. FLUX exhibits considerably stronger reconstruction quality than Lumina-DiMOO before any editing is performed (*e.g.*, 34.23 vs. 24.02 in global PSNR). This representation gap becomes more pronounced in the mask-grounded setting: since source content outside the editing mask is directly preserved, the final background fidelity is strongly affected by how accurately the respective continuous autoencoder or discrete tokenizer can reconstruct the original source image. As shown in Figure 5(b) and (c), in the example of adding "a cucumber" next to the "figure", even though the external mask directly preserves the content of the "figure", the reconstruction of Lumina-DiMOO still loses some fine-grained pixel-level details, resulting in degraded source preservation. Therefore, the stronger preservation of KV-Edit is partly attributable to the substantially stronger reconstruction capability of the FLUX autoencoder. Despite this inherent representation disadvantage, our method achieves substantially stronger target alignment (*c.f.*, Figure 5(b)), demonstrating the effectiveness of our framework for mask-grounded editing.

### C.3 Q3: Generalization to More Complex Editing Scenarios

**Setting.** We further evaluate training-free flow-based methods on the more challenging GIDE-Bench (Zhu et al., 2026b), which contains compositional editing tasks with diverse multimodal inputs and images of various resolutions. We include the mask-free methods FlowEdit (Kulikov et al., 2024) and RF-Inversion (Rout et al., 2024), as well as KV-Edit (Zhu et al., 2025), which uses SAM-generated segmentation masks for localized editing.

Table 5: Comparison with training-free flow-based methods on the GIDE-Bench.

| Method | MSE ↓ | PSNR ↑ | SSIM ↑ | Semantic Correctness ↑ | Perceptual Quality ↑ |
|---|---|---|---|---|---|
| *Continuous Flow-based* | | | | | |
| FlowEdit | 52.35 | 11.97 | 42.0 | 4.47 | 4.43 |
| RF-Inversion | 78.50 | 9.90 | 34.3 | 4.29 | 4.18 |
| KV-Edit | 28.88 | 15.60 | 53.1 | 3.90 | 3.82 |
| *Discrete dMLLM-based* | | | | | |
| GIDE | 20.99 | 17.70 | **63.5** | 4.39 | 4.29 |
| **Ours** | **16.95** | **18.00** | 63.3 | **4.56** | **4.50** |

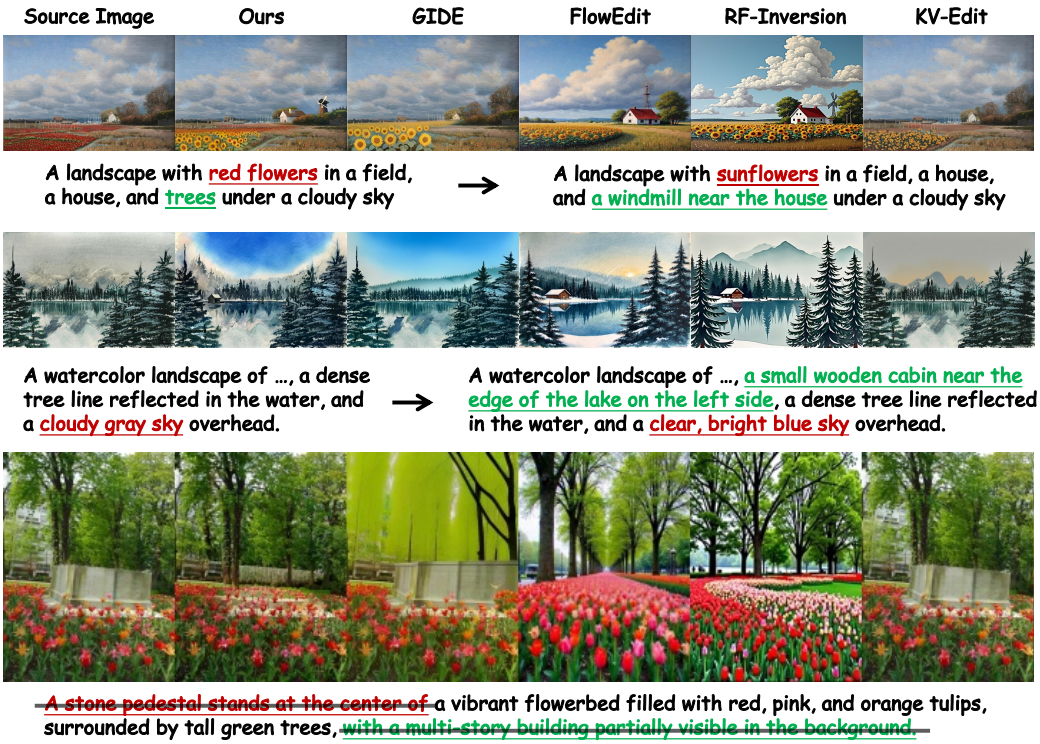


Figure 7: Qualitative comparisons with the flow-based method on GIDE-Bench.

**Results.** As examples shown in Figure 7 and results reported in Table 5, the mask-free FlowEdit and RF-Inversion achieve reasonable semantic correctness (4.47 and 4.29), but substantially degrade source consistency, with MSE increasing to 52.35 and 78.50 and PSNR decreasing to 11.97 and 9.90, respectively. Introducing external masks considerably improves source preservation for KV-Edit (28.88 MSE and 15.60 PSNR), but its target editing capability becomes notably limited, achieving only 3.90 semantic correctness and 3.82 perceptual quality. In contrast, our method achieves the best overall performance across both aspects: it obtains the strongest source preservation in terms of MSE (16.95) and PSNR (18.00), while simultaneously achieving the highest semantic correctness (4.56) and perceptual quality (4.50), with SSIM (63.3) also comparable to the best result (63.5).

Compared with its stronger performance on the standard $512 \times 512$ PIE-Bench, the relative advantage of KV-Edit does not transfer as effectively to GIDE-Bench, where editing involves more diverse resolutions, multimodal inputs, and compositional modifications. This suggests that relying primarily on a fixed external mask for localized continuous-space editing can become restrictive in more complex scenarios. In comparison, our dynamic masked generation continuously adapts edit localization, source guidance, and token decoding throughout generation, enabling a more favorable balance between target editing and source preservation. These results further demonstrate the generalization ability of our framework and the potential of effectively exploiting the intrinsic generation mechanism of dMLLMs for broader image editing scenarios.

Table 6: Ablation studies on PIE-Bench ***without*** extra grounding. C-G/C-E denote global/edit CLIP scores, and P-bg denotes background PSNR.

**(a) Dynamic Edit Localization**

| Method | C-G↑ | C-E↑ | P-bg↑ | Dist.↓ |
|---|---|---|---|---|
| Ours | 25.27 | 22.21 | 22.52 | 2.81 |
| w/o source-aware | 25.99 | 22.94 | 20.86 | 4.40 |

**(b) Adaptive Source Guidance**

| Method | C-G↑ | C-E↑ | P-bg↑ | Dist.↓ |
|---|---|---|---|---|
| Ours | 25.27 | 22.21 | 22.52 | 2.81 |
| w/o spatial-adaptive | 25.20 | 22.12 | 22.29 | 3.10 |
| w/o time-adaptive | 23.95 | 20.96 | 25.91 | 1.39 |

# D More Ablation Study

We provide additional ablation studies under both mask-free and mask-grounded settings on PIE-Bench (Ju et al., 2024).

## D.1 Ablations of Editing without External Grounding

While the main paper verifies the effectiveness of the three key modules as a whole, here we further investigate the specific design choices within each module.

**Design of Dynamic Edit Localization.** As discussed in Sec. 3.3, a reliable edit localization should be source-aware. In particular, for attribute or appearance modifications, the editable region should be determined according to the semantic discrepancy between the source image and the target prompt, rather than solely by the target generation itself. We therefore inject the same source residual into both branches before comparing their predictions, allowing the localization process to retain source-specific structural information.

Table 6(a) verifies the importance of this design. Removing the source-aware guidance and directly comparing the original predictions of the two branches increases target alignment (25.99 vs. 25.27 C-G and 22.94 vs. 22.21 C-E), but substantially degrades source preservation (20.86 vs. 22.52 P-bg and 4.40 vs. 2.81 Dist.). Without source information, prediction disagreement tends to identify broader semantic differences between the two prompts as editable, leading to excessive modifications of source-consistent structures and layouts. In contrast, the source-aware design provides a better balance between identifying target-relevant changes and preserving the original visual structure.

**Design of Adaptive Source Guidance.** As discussed in Sec. 3.4, our adaptive source guidance follows two complementary principles: *spatial-adaptive*, which applies different guidance strengths to editable and non-editable tokens, and *time-adaptive*, which gradually decreases the overall source guidance throughout generation. We separately ablate these two designs in Table 6(b).

Removing spatial-adaptive and applying the same time-varying guidance to all tokens consistently degrades the overall performance, including both target alignment and source preservation. This confirms the benefit of distinguishing editable from non-editable regions: source-consistent tokens require stronger preservation, while editable tokens should retain greater flexibility to accommodate the target modification. In contrast, removing time-adaptive guidance and maintaining a fixed strong guidance throughout generation results in substantially stronger source preservation (25.91 P-bg and 1.39 Dist.) but considerably weaker target alignment (23.95 C-G and 20.96 C-E). This indicates that spatial-adaptive alone is insufficient when the overall source constraint remains strong throughout the entire generation process. Gradually relaxing source guidance is necessary to provide sufficient flexibility for target content at later stages. Together, spatial and time-adaptive complement each other to achieve a balanced trade-off between faithful source preservation and effective target editing.

## D.2 Ablations of Editing with External Grounding

We further conduct ablation studies on PIE-Bench using the official ground-truth masks as external grounding. Overall, the results exhibit consistent trends with the mask-free setting, further validating our design under different editing initialization strategies.

**Guidance Scale.** We further study the effect of the initial source guidance scale $\lambda_0$ under external grounding in Table 7(a). Similar to the mask-free setting studied in the main paper, a smaller guidance scale generally favors target editing, particularly within the editable region, while a larger scale provides stronger source preservation. For example, increasing $\lambda_0$ from 0.6 to 0.8 improves P-bg from 26.20 to 26.30 and reduces Dist. from 5.46 to 5.11, while C-E decreases from 23.72 to 23.49.

Table 7: Ablation studies on PIE-Bench ***with*** extra grounding. C-G/C-E denote global/edit CLIP scores, and P-bg denotes background PSNR.

**(a) Guidance Scale**

| $\lambda_0$ | C-G↑ | C-E↑ | P-bg↑ | Dist.↓ |
|---|---|---|---|---|
| 0.6 | 25.95 | 23.72 | 26.20 | 5.46 |
| 0.7 | 25.97 | 23.69 | 26.21 | 5.30 |
| 0.8 | 25.78 | 23.49 | 26.30 | 5.11 |

**(b) Exp. Decay Hyperparameter**

| $\beta$ | C-G↑ | C-E↑ | P-bg↑ | Dist.↓ |
|---|---|---|---|---|
| 3 | 25.92 | 23.57 | 26.35 | 4.57 |
| 5 | 25.97 | 23.69 | 26.21 | 5.30 |
| 7 | 25.94 | 23.63 | 26.14 | 5.66 |

**(c) Key Modules**

| Method | C-G↑ | C-E↑ | P-bg↑ | Dist.↓ |
|---|---|---|---|---|
| Ours | 25.97 | 23.69 | 26.21 | 5.30 |
| w/o M1 | 24.88 | 22.56 | 26.16 | 3.30 |
| w/o M2 | 22.84 | 20.10 | 27.48 | 0.90 |
| w/o M3 | 24.90 | 22.60 | 26.22 | 3.06 |

**(d) Dynamic Edit Localization**

| Method | C-G↑ | C-E↑ | P-bg↑ | Dist.↓ |
|---|---|---|---|---|
| Ours | 25.97 | 23.69 | 26.21 | 5.30 |
| w/o source-aware | 26.01 | 23.76 | 26.19 | 5.58 |

**(e) Adaptive Source Guidance**

| Method | C-G↑ | C-E↑ | P-bg↑ | Dist.↓ |
|---|---|---|---|---|
| Ours | 25.97 | 23.69 | 26.21 | 5.30 |
| w/o spatial-adaptive | 25.84 | 23.52 | 26.14 | 5.62 |
| w/o time-adaptive | 24.79 | 22.36 | 26.81 | 2.50 |

**(f) Guidance Schedule**

| Schedule | C-G↑ | C-E↑ | P-bg↑ | Dist.↓ |
|---|---|---|---|---|
| Exp Decay ($\lambda_0 = 0.7$) | 25.97 | 23.69 | 26.21 | 5.30 |
| Linear Decay ($\lambda_0 = 0.7$) | 25.16 | 22.79 | 26.63 | 2.90 |
| Linear Decay ($\lambda_0 = 0.3$) | 25.93 | 23.67 | 26.22 | 5.32 |

This consistent trend confirms that the guidance scale provides a simple and interpretable control over the editing–preservation trade-off. Meanwhile, the overall performance remains relatively stable across different scales, indicating that our method is not sensitive to this hyperparameter. We therefore consistently use $\lambda_0 = 0.7$ across both mask-free and mask-grounded settings without setting-specific tuning.

**Hyperparameter of Exponential Decay.** As described in Sec. B, when external grounding is available, we adopt a faster exponential decay schedule controlled by $\beta$. We study the effect of $\beta$ in Table 7(b). Increasing $\beta$ from 3 to 5 leads to slightly stronger target alignment (25.92 to 25.97 C-G and 23.57 to 23.69 C-E), accompanied by a moderate decrease in source preservation, reflecting the expected editing–preservation trade-off. Further increasing $\beta$ to 7 provides no additional improvement in target alignment while further degrading source preservation. We therefore use $\beta = 5$ as the default choice, which provides a favorable balance between target editing and source preservation.

**Effectiveness of Key Modules.** Table 7(c) evaluates the three key modules under external grounding. Consistent with the mask-free results in the main paper, removing any module (M1: dynamic localization, M2: adaptive guidance, M3: edit-aware decoding) reduces target alignment and makes generation more biased toward reconstructing the source content. These results demonstrate that even with an external mask providing a candidate edit region, dynamically identifying editable tokens, adapting source guidance, and controlling the decoding order remain important for effective editing. Together, the three modules enable the model to move beyond simple masked reconstruction toward flexible target-conditioned generation.

**Design of Dynamic Edit Localization.** We further ablate the source-aware design of dynamic localization in Table 7(d). The results are consistent with the mask-free setting in Sec. D.1. Directly comparing the predictions of the two branches without source-aware guidance slightly increases target alignment (26.01 vs. 25.97 C-G and 23.76 vs. 23.69 C-E), but leads to weaker global source preservation (5.58 vs. 5.30 Dist.). This again indicates a tendency toward over-editing when source-specific information is not incorporated into localization. Although the effect is less pronounced than in the mask-free setting, since the external mask already constrains the candidate editing area, source-aware localization still provides a more balanced editing–preservation trade-off.

**Design of Adaptive Source Guidance.** Table 7(e) further verifies the spatial- and time-adaptive designs. Removing spatial adaptation consistently degrades both target alignment and source preservation, confirming the benefit of distinguishing editable and non-editable tokens even within an externally grounded editing process. Removing time adaptation instead results in substantially stronger source preservation (26.81 P-bg and 2.50 Dist.) but considerably weaker target alignment (24.79 C-G and 22.36 C-E), indicating that fixed strong guidance again biases generation toward source reconstruction. These observations are consistent with those under mask-free editing: spatial-adaptive

guidance provides region-specific source constraints, while time-adaptive guidance progressively relaxes these constraints to accommodate the target modification. Combining both achieves a more favorable balance between editability and source consistency.

**Guidance Schedule.** Our adaptive source guidance is not restricted to a particular decay schedule. In Table 7(f), we further replace the default exponential decay with a linear schedule. With the same initial scale $\lambda_0 = 0.7$, linear decay retains stronger source guidance for a longer portion of the generation process, resulting in stronger source preservation but insufficient target editing (25.16/22.79 C-G/C-E). Reducing its initial scale to $\lambda_0 = 0.3$ compensates for the slower decay and recovers a similar editing–preservation trade-off to the default exponential schedule (25.93 vs. 25.97 C-G, 23.67 vs. 23.69 C-E, and 5.32 vs. 5.30 Dist.). These results demonstrate that the effectiveness of adaptive source guidance does not rely on a specific functional form. Instead, different schedules can achieve comparable performance when following the same principle of progressively relaxing source constraints throughout generation. We adopt exponential decay as a simple and effective default for the mask-grounded setting.

**Discussion on Mask-grounded Trade-off.** Compared with mask-free editing, the default mask-grounded setting adopts a faster exponential decay, since the preserved source tokens outside the external mask already provide reliable visual context. This intentionally relaxes source constraints more aggressively within the candidate edit region, favoring target alignment at the cost of a larger global structure distance. Notably, this behavior is not inherent to mask-grounded editing. With the same initial guidance scale ($\lambda_0 = 0.7$), using a slower linear decay reduces Dist. from 5.30 to 2.90 (*c.f.*, Table 7(f)), which is comparable to the mask-free setting (2.81), while achieving better background preservation and higher editing-region alignment than mask-free editing. These results demonstrate that mask grounding provides additional flexibility to balance source preservation and target editability, while the guidance schedule controls the desired operating point.

Overall, the ablation results under mask-free and mask-grounded settings exhibit highly consistent behaviors. The three key modules remain complementary under different initialization strategies, while the source-aware, spatial-adaptive, and time-adaptive designs consistently provide a better balance between editability and source preservation. Moreover, our framework remains stable across different guidance scales and is compatible with different decay schedules. These results demonstrate that the effectiveness of DynEdit stems from its underlying dynamic editing principles rather than a particular initialization, hyperparameter, or guidance schedule, supporting its robustness and generality across different editing settings.

## E Inference Pipeline

We show the full inference pipeline in Algorithm 1. Initialize returns a fully masked sequence when $\mathbf{M}$ is unavailable, and a partially masked source sequence when external grounding is provided. Finally, the edited image tokens are fed into the visual decoder to reconstruct the final edited image.

**Algorithm 1** DynEdit Inference Pipeline

**Require:** Source image tokens $\mathbf{x}^s$; Source prompt $P^{\text{src}}$; Target prompt $P^{\text{tgt}}$; Total steps $T$; Optional external mask $\mathbf{M}$
**Ensure:** Edited image tokens $\mathbf{x}_T$

$\mathbf{x}_0 \leftarrow \textsc{Initialize}(\mathbf{x}^s, \mathbf{M})$ ▷ Fully or partially masked initialization
**for** $t = 0$ **to** $T - 1$ **do**
  $\mathbf{x}_t^s \leftarrow \textsc{AlignSource}(\mathbf{x}^s, \mathbf{x}_t)$ ▷ Construct parallel source branch
  $\hat{\mathbf{y}}_t^s \leftarrow D_\theta(\mathbf{x}_t^s, P^{\text{src}})$
  $\mathbf{z}_t \leftarrow \text{LAI}(\mathbf{x}_t^s, \hat{\mathbf{y}}_t^s, \mathbf{x}^s) - \hat{\mathbf{y}}_t^s$ ▷ Parallel residual extraction
  $\hat{\mathbf{y}}_t \leftarrow D_\theta(\mathbf{x}_t, P^{\text{tgt}})$ ▷ Target prediction
  $\mathbf{m}_t \leftarrow \textsc{DynamicEditLocalization}(\hat{\mathbf{y}}_t, \hat{\mathbf{y}}_t^s, \mathbf{z}_t, t)$ ▷ Identify editable tokens
  $\tilde{\mathbf{y}}_t \leftarrow \textsc{AdaptiveSourceGuidance}(\hat{\mathbf{y}}_t, \mathbf{z}_t, \mathbf{m}_t, t)$ ▷ Adapt source preservation
  $\hat{\mathbf{x}}_t, \mathbf{p}_t \leftarrow \textsc{Sample}(\tilde{\mathbf{y}}_t)$ ▷ Sample tokens and confidence
  $\mathbf{x}_{t+1} \leftarrow \textsc{EditAwareDecoding}(\hat{\mathbf{x}}_t, \mathbf{p}_t, \mathbf{m}_t, t)$ ▷ Decode and remask
**end for**
**return** $\mathbf{x}_T$

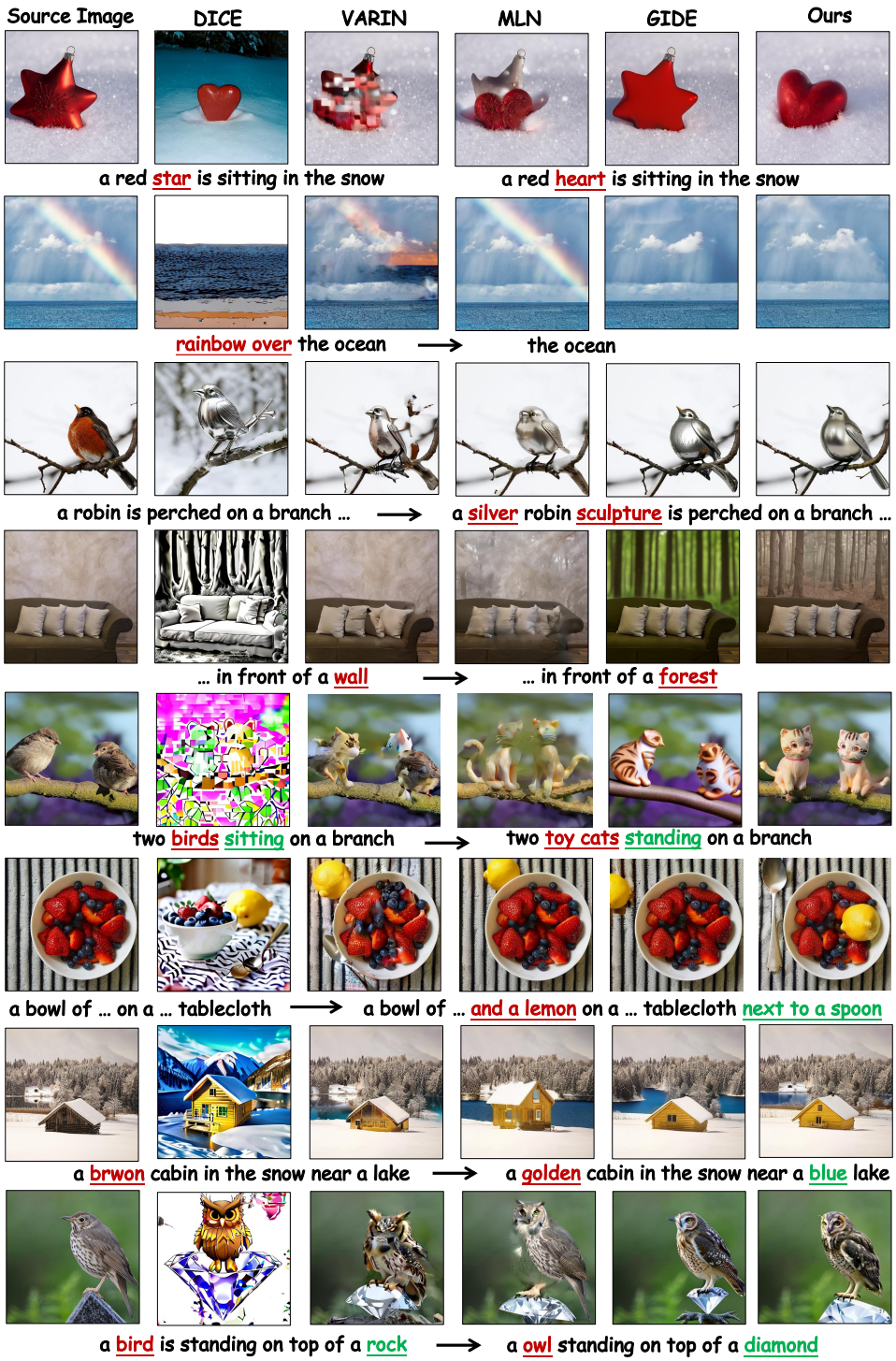


Figure 8: Qualitative comparisons on PIE-Bench (rows 1-4) and PIE-Bench++ (rows 5-8). GIDE uses external masks from SAM, other methods don't use external masks. All methods use 64 steps.

## F More Visualization Results

**Q1: Editing without External Grounding.** Figure 8 presents additional qualitative comparisons. Figure 9 further demonstrates the model generalization of our framework across different dMLLM backbones. Both Lumina-DiMOO and MMaDA achieve effective editing, while MMaDA tends to favor stronger target alignment in the editing–preservation trade-off.

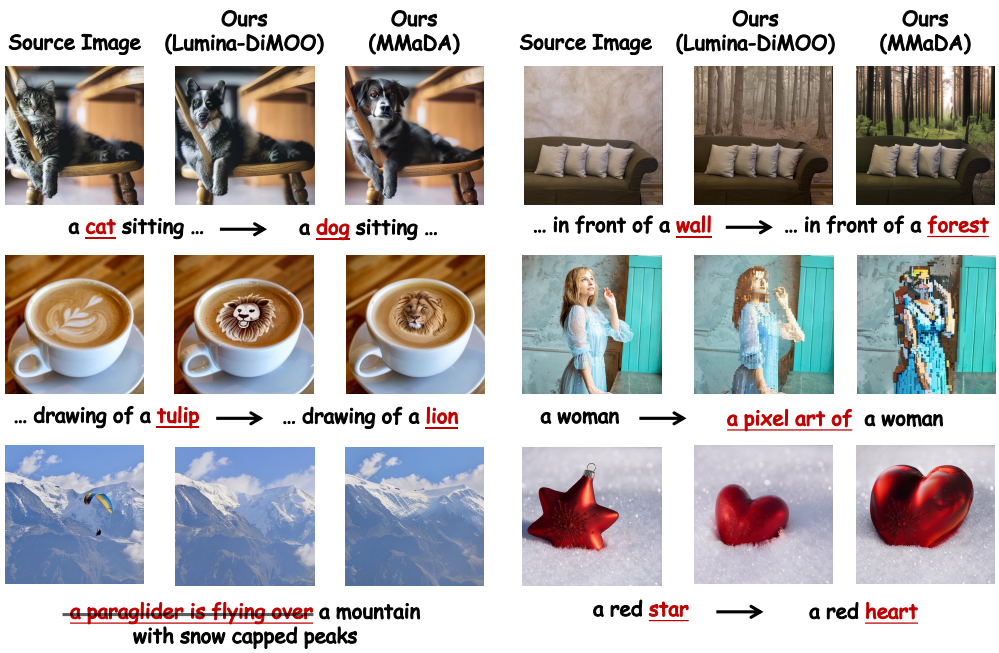


Figure 9: Qualitative results on PIE-Bench with different backbones. Both methods use 64 steps.

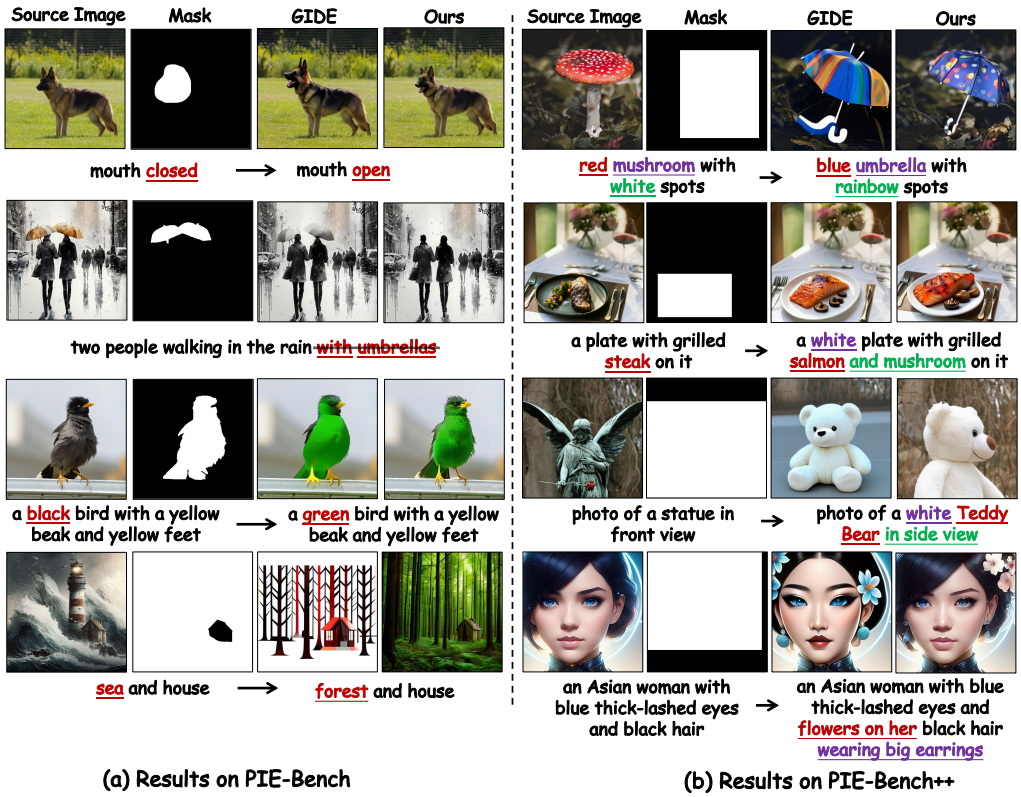


Figure 10: Qualitative comparisons on PIE-Bench and PIE-Bench++ with the same external masks. (a) Both methods use the official ground-truth masks on PIE-Bench. (b) Both methods use the grounded bounding-box masks generated by SAM on PIE-Bench++. Both methods use 64 steps.

**Q2: Editing with External Grounding.** Figure 10 presents additional qualitative comparisons under external grounding. Figure 11 further visualizes the same editing samples with different generation budgets (8, 16, 32, and 64 steps). Our method produces stable and consistent editing

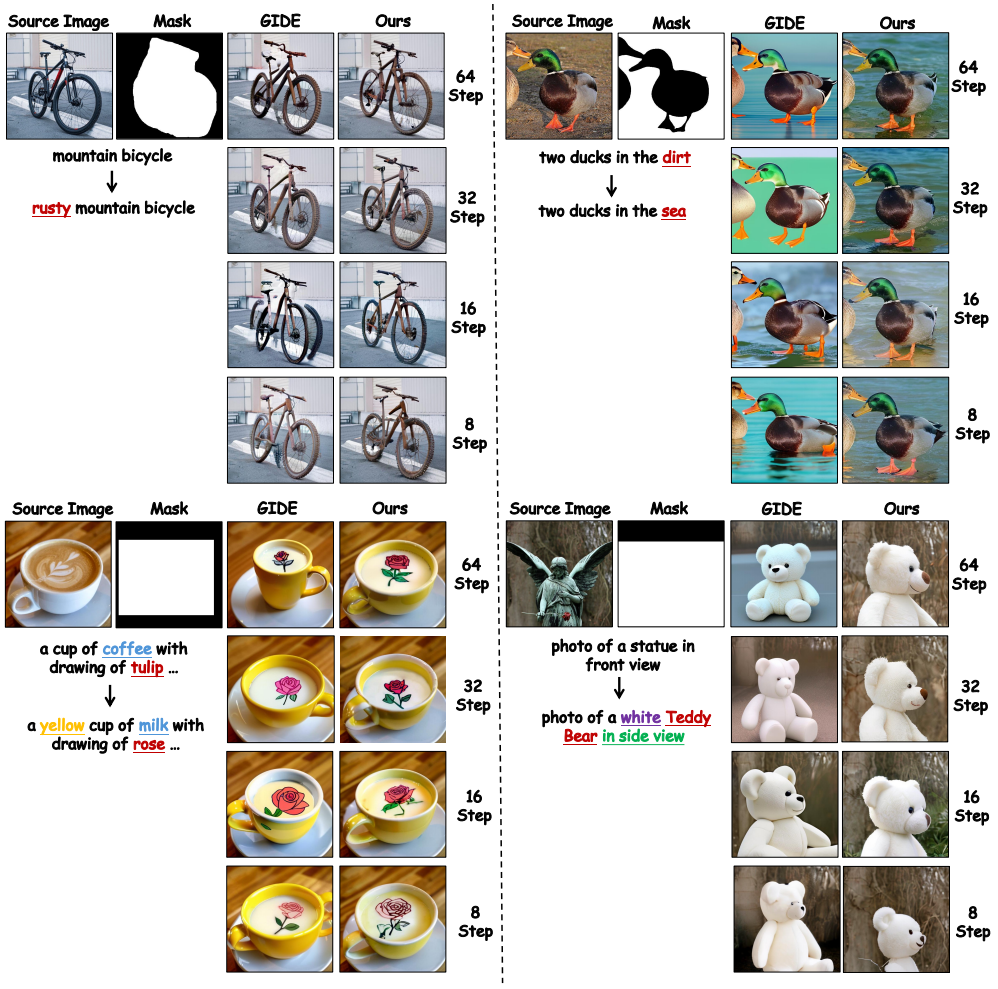


Figure 11: Qualitative comparisons with different generation steps. Both methods use the same ground-truth masks on PIE-Bench, and the bounding-box masks on PIE-Bench++.

results across different numbers of generation steps. Although 8-step generation may lack some fine-grained visual details, 16 steps are already sufficient to achieve faithful and visually coherent editing, with further increasing the generation steps mainly refining local details. In contrast, GIDE shows larger variations in both edited content and visual quality across different generation budgets. These results further demonstrate that our framework can effectively exploit external grounding as a reliable initialization to accelerate the dynamic editing process, achieving stable and high-quality editing with substantially fewer generation steps.

**Generation Process.** Figure 12 visualizes the mask-free generation process, including intermediate decoding results and dynamic edit masks at different steps. Across diverse edits, including object replacement, addition/removal, and fine-grained attribute changes, our method progressively identifies the editable regions while prioritizing the decoding of source-consistent background and structures. Interestingly, at early steps, although most masked tokens are initially regarded as non-editable, source-consistent tokens typically receive higher confidence after source guidance and are decoded first, while tokens with larger source–target semantic discrepancies remain masked due to their relatively lower confidence. As more source-consistent context is revealed, the editable regions become progressively better localized and regenerated toward the target. At late steps, although the decayed source guidance may cause most remaining masked tokens to be classified as editable, the already decoded source structure provides sufficient visual context, together with the minimum guidance constraint, to maintain source consistency. Overall, this progressive process demonstrates how our framework naturally exploits the localized and iterative masked generation of dMLLMs to coordinate source preservation and target editing without requiring external localization.

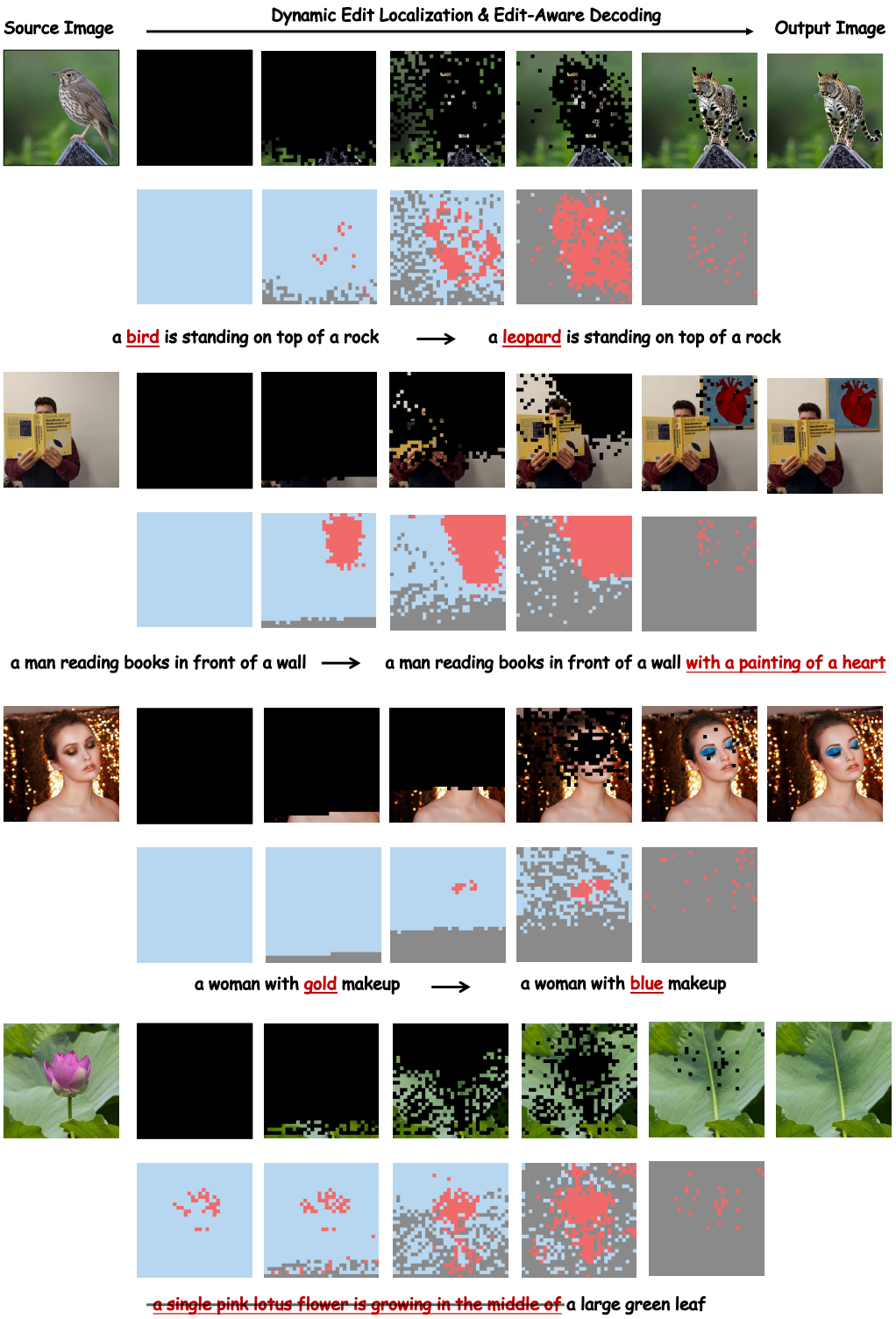


Figure 12: Visualization of the Generation Process. It visualizes the input and corresponding dynamic edit mask of steps 0, 16, 32, 48, 64 of a 64-step generation. For the edit mask, grey denotes decoded tokens, red denotes editable tokens, and blue denotes non-editable tokens.

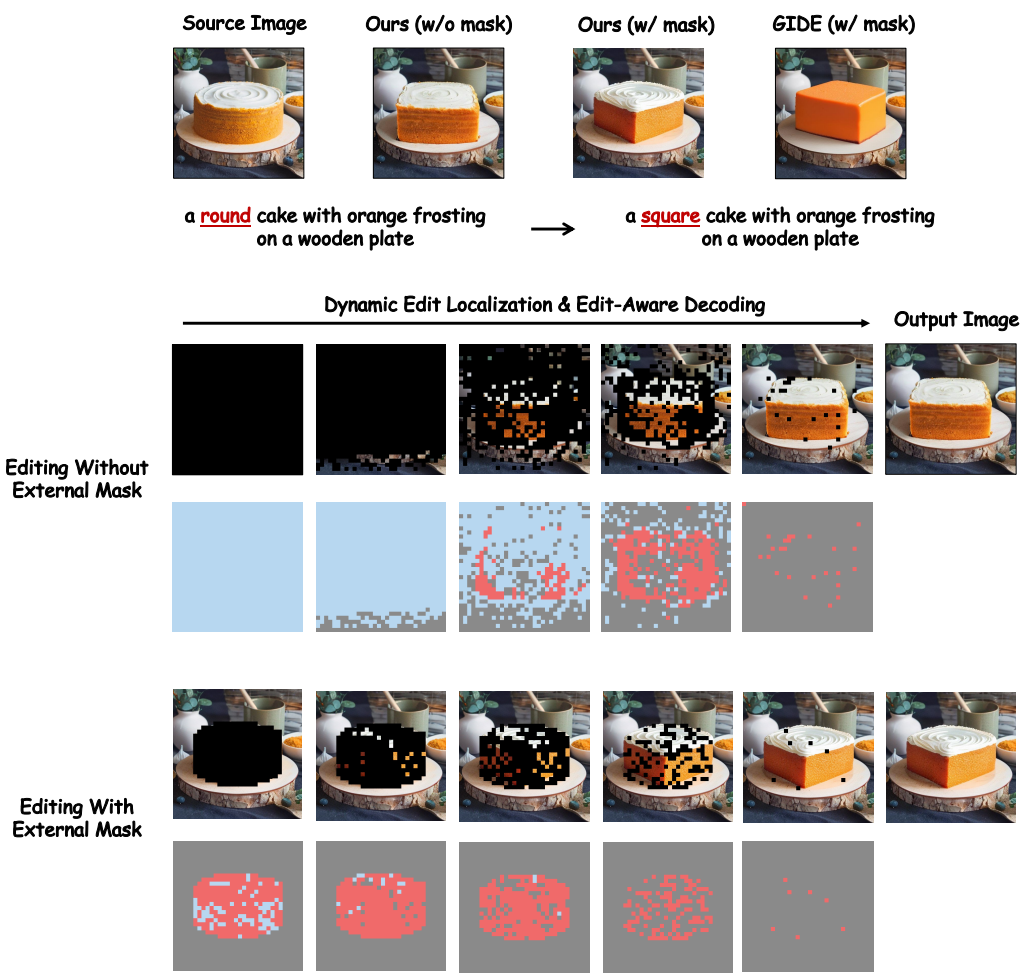


Figure 13: Visualization of the Generation Process. It visualizes the input and corresponding dynamic edit mask of steps 0, 16, 32, 48, 64 of a 64-step generation. For the edit mask, grey denotes decoded tokens, red denotes editable tokens, and blue denotes non-editable tokens.

Figure 13 further compares the generation process of the same sample with and without external grounding, initialized from partially and fully masked tokens, respectively. When an external mask is provided, our method can further refine the candidate edit region and identify source-consistent content within the mask, avoiding the over-editing observed in GIDE and better preserving distinctive source details, such as the original texture of the cake. Due to the faster guidance decay and fewer masked tokens under external grounding, most remaining tokens may be classified as editable at intermediate steps. Nevertheless, the already decoded background and structural content provides sufficient visual context, while the minimum source guidance further maintains source consistency. This demonstrates that external grounding serves as a useful initialization rather than a fixed editing constraint, allowing our dynamic framework to achieve more efficient editing while faithfully preserving source-specific details.

**Q3: Generalization to More Complex Editing Scenarios.** Figure 14 presents qualitative comparisons with end-to-end training-based methods, including FLUX-Kontext (Labs et al., 2025) and Qwen-Image-Edit (Wu et al., 2025a), on GIDE-Bench. Despite being entirely training-free, our method further narrows the gap with these dedicated editing models and, in some fine-grained localized editing scenarios (*e.g.*, adding a small windmill or a wooden cabin), achieves a more favorable trade-off between target editing and source preservation.




Figure 14: Qualitative comparisons with the end-to-end training-based methods on GIDE-Bench.




Figure 15: Visualization of the failure cases and limitations.

## G Limitations and Future Work

As a training-free framework built upon pre-trained dMLLMs, our method inevitably inherits the representation and generation limitations of the underlying models. First, as discussed in Sec. C.1, the discrete visual tokenizers of current dMLLMs introduce non-negligible reconstruction loss, particularly for fine-grained visual details. This limits editing scenarios that require precise reconstruction or regeneration, such as text rendering, faces, and hands. As shown in the first row of Figure 15, removing the glasses from a woman may introduce facial distortions during regeneration, while fine details of the hand and the held book can also be degraded.

Second, current dMLLM backbones remain limited in generating complex human interactions, crowded scenes, and compositional actions. For example, adding a crowd to a conference room can result in distorted or overlapping people. Similarly, when changing a walking fox to "yawning and rolling", both backbones successfully capture the simpler "yawning" attribute but fail to realize the more complex rolling action. For human pose editing, changing a man with hands down to "hands in pockets and standing sideways" requires simultaneous modification of both local hand poses and global body orientation, which remains challenging for both Lumina-DiMOO and MMaDA. These examples suggest that the performance of our framework is ultimately bounded by the semantic and visual generation capabilities of the pre-trained backbone, especially for complex instructions requiring coordinated changes across multiple regions.

Beyond developing stronger dMLLM backbones and higher-fidelity visual tokenizers, an interesting future direction is to integrate dMLLM-based editing into more systematic agent-level editing frameworks. Complex instructions could be decomposed through multimodal reasoning and external grounding into a sequence of simpler, localized editing operations. Such a framework could further exploit the localized generation capability of dMLLMs for efficient multi-step editing, while providing more systematic and controllable handling of complex editing instructions.