GIDE: Unlocking Diffusion LLMs for Precise Training-Free Image Editing

Zifeng Zhu[1,2], Jiaming Han[2], Jiaxiang Zhao[1], Minnan Luo[1†], and Xiangyu Yue[2†]

[1] Xi'an Jiaotong University
[2] MMLab, The Chinese University of Hong Kong
zivenzhu@stu.xjtu.edu.cn

Abstract. While Diffusion Large Language Models (DLLMs) have demonstrated remarkable capabilities in multi-modal generation, performing precise, training-free image editing remains an open challenge. Unlike continuous diffusion models, the discrete tokenization inherent in DLLMs hinders the application of standard noise inversion techniques, often leading to structural degradation during editing. In this paper, we introduce **GIDE** (Grounded Inversion for DLLM Image Editing), a unified framework designed to bridge this gap. GIDE incorporates a novel **Discrete Noise Inversion** mechanism that accurately captures latent noise patterns within the discrete token space, ensuring high-fidelity reconstruction. We then decompose the editing pipeline into **grounding**, **inversion**, and **refinement** stages. This design enables GIDE supporting various editing instructions (text, point and box) and operations while strictly preserving the unedited background. Furthermore, to overcome the limitations of existing single-step evaluation protocols, we introduce GIDE-Bench, a rigorous benchmark comprising 805 compositional editing scenarios guided by diverse multi-modal inputs. Extensive experiments on GIDE-Bench demonstrate that GIDE significantly outperforms prior training-free methods, improving Semantic Correctness by 51.83% and Perceptual Quality by 50.39%. Additional evaluations on ImgEdit-Bench confirm its broad applicability, demonstrating consistent gains over trained baselines and yielding photorealistic consistency on par with leading models.[3].

Keywords: Image Editing · Diffusion LLMs · Discrete Inversion

1 Introduction

Diffusion Large Language Models (DLLMs) [31,35,54] have emerged as a powerful paradigm for unified multimodal modeling. Unlike autoregressive models constrained by sequential dependency [8, 33, 57, 69], DLLMs achieves superior sampling efficiency through parallel token generation [61,68]. Despite their potential, the application of DLLMs to image editing remains largely under-explored.

† Corresponding authors.
[3] Data and code are available at https://github.com/Zivenzhu/GIDE.

Fig. 1: Demonstration of our proposed GIDE framework. Given a source image (left of each pair) and an editing instruction guided by flexible spatial modalities (points, bounding boxes, or pure text), GIDE enables precise and localized image editing. The edited results (right of each pair) demonstrate high-quality visual synthesis and strict background preservation, all achieved in a completely training-free manner.

Existing editing approaches faces significant challenge due to the inherent discrete nature of DLLMs [7,42,71]. On the one hand, finetuning-based methods often struggle to balance editability and fidelity [68]. On the other hand, training-free methods are currently hindered by the lack of a principled inversion mechanism tailored for discrete token spaces. Unlike standard diffusion models where continuous noise inversion (e.g., DDIM inversion [13,55]) is well-established, the discrete, LLM-based diffusion process lacks a counterpart to accurately map images back to their initial noise states. This absence leads to poor semantic preservation and severe visual artifacts, ultimately limiting the practical utility of DLLMs for precise editing tasks and hindering their widespread adoption.

To bridge this critical gap, we present **Grounded Inversion for DLLM Image Editing (GIDE)**, the first training-free framework specifically designed for precise inversion and editing within the discrete token space of DLLMs. Operating entirely at inference time, GIDE adapts flexibly to diverse editing scenarios (see Fig. 1) without the need for additional data annotation or model retraining. The core philosophy of GIDE is to decompose the complex editing task into three synergistic stages, each addressing a distinct sub-problem:

- **Grounding for Localization**. To determine *where* to edit, we introduce a robust grounding module that accepts flexible user inputs, ranging from points to boxes to text descriptions, to precisely localize target regions, ensuring modifications are strictly confined to the intended area.
- **Discrete Inversion for Preservation**. To determine *how* to edit while preserving the original context, we propose a novel inversion mechanism tailored for discrete DLLMs. This cornerstone module projects the source image into a structure-aware latent representation, establishing a stable and reliable foundation for high-fidelity editing in complex visual scenarios.
- **Refinement for Harmonization**. Finally, to ensure visual coherence, a refinement module post-processes the edited regions, seamlessly blending them with the background preservation compared to prior attempted but also enables versatile editing capabilities across various modalities.

Beyond our methodological contributions, we identify a critical gap in current evaluation protocols. Existing benchmarks primarily focus on single-step text instructions [15, 53, 73, 77], failing to capture the complexity of compositional editing and spatial control that GIDE aim to solve. Furthermore, standard metrics like CLIPScore [21] often favor source preservation over editing faithfulness [26, 48], allowing models to "cheat" by ignoring instructions. To rigorously validate GIDE against these higher standards, we introduce **GIDE-Bench**, a challenging benchmark comprising 805 compositional editing scenarios guided by diverse multimodal inputs. Unlike prior datasets, GIDE-Bench emphasizes precise region control and multi-step reasoning. We couple this with a holistic evaluation protocol that combines dual-model (GPT and Gemini) assessments for semantic correctness and perceptual quality with strict masked-based metrics for background preservation, ensuring a faithful analysis of whether models can modify intended content while strictly preserving the surrounding context.

Extensive experiments on GIDE-Bench demonstrate that GIDE significantly outperforms state-of-the-art training-free methods, improving semantic correctness by 51.83% and perceptual quality by 50.39%. Notably, it achieves photo-realistic consistency comparable to leading models, effectively preserving depth and lighting interfaces often lost in prior works. Further evaluation on ImgEdit-Bench [72] confirm its robustness, surpassing the trained baseline model by up to 39.13% on complex editing tasks. In summary, our work makes **three key contributions**: **1)** The first discrete noise inversion mechanism tailored for DLLMs. **2)** A unified, training-free editing framework supporting diverse editing prompts and scenarios. **3)** A rigorous benchmark for compositional editing evaluation.

2 Related Work

2.1 Training-Free Image Editing

Training-free image editing enables controllable modifications without additional training [1,2,9,14,22,23,25,37,39,41,70,78,79], mainly via two paradigms: attention control and inversion. **Attention control methods**, designed for diffusion

models [50], manipulate attention maps to maintain consistency: Prompt-to-Prompt [20] reuses source cross-attention maps to preserve backgrounds, MasaCtrl [4] substitutes key and value matrices in deeper layers for region-aware control, and Add-it [59] dynamically weights source and target attention representations. **Inversion-based methods** reconstruct the generation trajectory for faithful editing: DDIM Inversion [13,55] reverses the diffusion ODE, Null-text Inversion [38] extends it to text-guided generation, and Direct Inversion [24] replaces this optimization with logit differences, while DICE [19] and VARIN [11] adapt inversion to masked generative and visual autoregressive models [6,56,60], respectively. However, a principled inversion algorithm tailored for the stochastic and discrete nature of DLLMs remains largely lacking, and methods like DICE still exhibit several notable limitations in practice when applied (Sec. 5.2).

2.2 Evaluation of Image Editing

Early image editing benchmarks evaluate semantic alignment via CLIP-based metrics [16, 29, 52, 53] and non-edited region preservation via pixel-level metrics (PSNR, SSIM, MSE), as in PIE-Bench [24] and MagicBrush [74]. However, CLIP-based evaluation is unreliable, as trivial solutions such as copying the original image still achieve high scores, failing to disentangle local modifications from global context [26]. Recent benchmarks thus adopt VLM-based evaluators [18, 45–47, 63], e.g., GPT-4o [43], which agree better with human judgments [72], as in ImgEdit-Bench [72], GEdit-Bench [34], and Omni-Edit-Bench [64]. Yet VLMs alone struggle to capture fine-grained structural distortions and pixel-level background preservation. GIE-Bench [48] combines both but relies on predefined edited regions that may not reflect the actual edits. In contrast, GIDE-Bench dynamically grounds edited regions and incorporates multi-modal instructions for a more challenging and comprehensive evaluation.

3 Approach

In this section, we introduce **Grounded Inversion for DLLM Image Editing (GIDE)**, a generalized framework for training-free image editing using DLLMs. To overcome the limitations of existing methods in discrete latent space, GIDE adopts a modular design comprising **Grounding**, **Inversion**, and **Refinement** (see Fig. 2). This architecture allows us to decouple the "where" (location via grounding) from the "how" (reconstruction via discrete inversion) and the "quality" (enhancement via refinement). Consequently, our framework naturally extends to a wide range of editing operations without requiring task-specific tuning. In the following section, we will introduce each component in detail.

3.1 Grounding-Aware Discrete Inversion

Different from continuous diffusion models that use deterministic ODEs for inversion, DLLMs rely on discrete, stochastic sampling. This means simply reversing

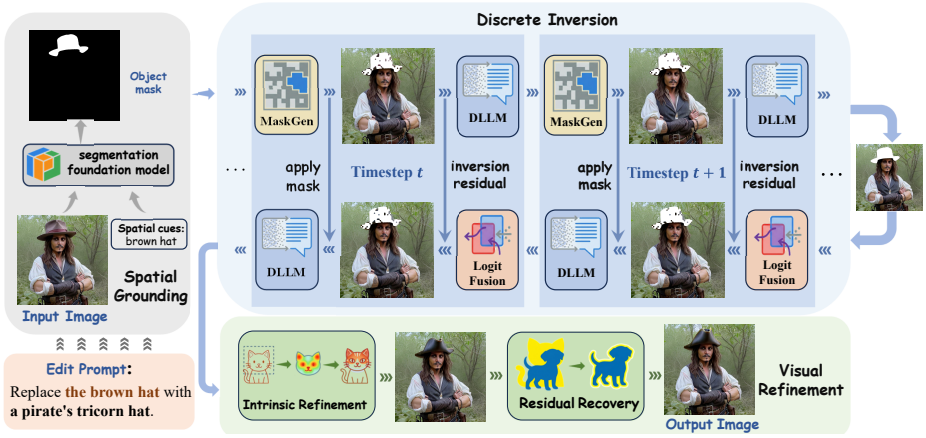

Fig. 2: The overall architecture of GIDE. It decouples the editing process into three sequential stages: **Grounding** locates the target region via a segmentation foundation model; **Inversion** executes the edit by reconstructing the discrete latent space; and **Refinement** enhances the visual coherence and fidelity of the final output.

the generation process is impossible: the same image could be generated by millions of different random paths. To solve this, we propose a **Grounding-Aware Discrete Inversion** algorithm (Algorithm 1). The core idea is to record the "errors" (residuals) the model makes when reconstructing the original image, and then replay these errors during editing to force the model to stay close to the original structure and prevent unintended visual deviations during editing.

Grounding-Constrained Masking.[4] To prevent edits from leaking into the background, we must strictly control where the model is allowed to generate new tokens. We employ a sinusoidal masking schedule to determine the number of tokens to mask at step t: $n_t = \lfloor N \cdot \sin(\frac{\pi t}{2T}) \rfloor$, where N is the total token count inside the grounding mask $\mathbf{M}$. This explicitly reverses the natural generation process: masking more tokens early on to capture fine-grained texture changes, and fewer tokens later to maintain global structural stability. To determine *which* tokens to mask, we calculate a confidence score $s^{(i)}$ for each token i based on the model's prediction probability. Crucially, to ensure progressive and strictly local masking, we enforce two constraints: 1) scores for tokens already masked in previous steps are set to $+\infty$, and 2) scores for tokens outside $\mathbf{M}$ are set to $-\infty$. The resulting step-specific binary mask $\boldsymbol{m}_t$ is then generated as:

$$m_t^{(i)} = \begin{cases} 1, & \text{if } s^{(i)} \in \text{top-}n_t(\mathbf{S}), \\ 0, & \text{otherwise.} \end{cases}$$

This ensures that the background $\boldsymbol{x}_{\text{bg}} = \boldsymbol{x}_0 \odot (\mathbf{1} - \mathbf{M})$ remains strictly preserved throughout the process, preventing any unintended visual alterations.

[4] Implementation details of the grounding module are provided in Sec. 3.2.

Algorithm 1 Grounding-Aware Inversion and Editing for DLLMs

Require: Input image $\boldsymbol{x}_0$; Grounding mask $\mathbf{M}$; Total timesteps T; Mask token id k_{mask}; DLLM $\mathcal{D}_\theta$; Source prompt $\boldsymbol{c}$, Target prompt $\boldsymbol{c}'$; Mixing coef. λ
Ensure: Edited image $\tilde{\boldsymbol{x}}_1$

// Stage 1: Inversion (Forward Process)
for $t = 1$ **to** T **do**
  $n_t \leftarrow \texttt{SineSchedule}(t, T)$ ▷ Calculate mask ratio via sine function
  $\boldsymbol{m}_t \leftarrow \texttt{MaskGen}(\boldsymbol{x}_0, \mathbf{M}, n_t)$ ▷ Generate mask within $\mathbf{M}$
  $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_0 \odot (\mathbf{1} - \boldsymbol{m}_t) + k_{\text{mask}} \cdot \boldsymbol{m}_t$ ▷ Apply mask tokens
  $\hat{\boldsymbol{y}}_t \leftarrow \mathcal{D}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t)$ ▷ Predict token logits
  $\boldsymbol{y}_t \leftarrow \texttt{LAI}(\boldsymbol{x}_t, \hat{\boldsymbol{y}}_t, \boldsymbol{x}_0)$ ▷ Construct ground-truth logits
  $\boldsymbol{z}_t \leftarrow \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$ ▷ Store inversion residual
end for
// Stage 2: Editing (Reverse Process)
$\tilde{\boldsymbol{x}}_{T+1} \leftarrow \boldsymbol{x}_0$
for $t = T$ **to** 1 **do**
  $\boldsymbol{x}_t \leftarrow \tilde{\boldsymbol{x}}_{t+1} \odot (\mathbf{1} - \boldsymbol{m}_t) + k_{\text{mask}} \cdot \boldsymbol{m}_t$ ▷ Re-apply grounding mask
  $\hat{\boldsymbol{y}}_t \leftarrow \mathcal{D}_\theta(\boldsymbol{x}_t, \boldsymbol{c}', t)$ ▷ Predict logits under target prompt
  $\boldsymbol{g} \sim \text{Gumbel}(\mathbf{0}, \mathbf{I})$ ▷ Sample Gumbel noise
  $\tilde{\boldsymbol{y}}_t \leftarrow \hat{\boldsymbol{y}}_t + \lambda \cdot \boldsymbol{z}_t + (1 - \lambda) \cdot \boldsymbol{g}$ ▷ Logit fusion with stochasticity
  $\tilde{\boldsymbol{x}}_t \leftarrow \arg\max \tilde{\boldsymbol{y}}_t$ ▷ Token selection
end for
return $\tilde{\boldsymbol{x}}_1$

Stochastic Logit Rectification. To faithfully preserve the source object's appearance, we employ location-aware argmax inversion (LAI) [11] to extract a residual term $\boldsymbol{z}_t$, which encodes the structural priors of the original image. However, directly injecting this deterministic residual often overly restricts the generation process, leading to rigid artifacts and over-smoothed textures. To mitigate this, we introduce a stochastic fusion strategy that dynamically interpolates between the target semantics and the source structure. Specifically, we rectify the target logits $\hat{\boldsymbol{y}}_t$ by incorporating the inversion residual $\boldsymbol{z}_t$ along with a Gumbel noise term $\boldsymbol{g}$, controlled by a single tunable mixing coefficient λ:

$$\tilde{\boldsymbol{y}}_t = \hat{\boldsymbol{y}}_t + \lambda \cdot \boldsymbol{z}_t + (1 - \lambda) \cdot \boldsymbol{g}.$$

This formulation effectively balances the trade-off between semantic editability (driven by $\hat{\boldsymbol{y}}_t$) and structural fidelity (anchored by $\boldsymbol{z}_t$). Crucially, the controlled injection of Gumbel noise prevents the sampling distribution from collapsing into undesirable deterministic modes, thereby preserving fine-grained high-frequency details and ensuring a more natural and coherent synthesis.

3.2 Multimodal Spatial Grounding

To preserve non-edited regions, GIDE operates under a strict region-based constraint. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and an editing instruction $\mathcal{T}$, we

Algorithm 2 Spatially Diverse Point Sampling

Require: Mask $M \in \{0,1\}^{H\times W}$, Count $K = 4$
Ensure: Point set $\mathcal{S}$
$\mathcal{G} \leftarrow \{p \mid \forall q \in \mathcal{N}_p^{3\times 3}, M(q) = 1\}$ ▷ Interior pixels extraction
$\mathbf{c} \leftarrow \text{Mean}(\mathcal{G})$; Partition $\mathcal{G}$ into quadrants $\{Q_k\}_{k=1}^4$ centered at $\mathbf{c}$
$\mathcal{S} \leftarrow \bigcup_{k=1}^4 \{\arg\max_{p\in Q_k} \|p - \mathbf{c}\|_2 \text{ s.t. } Q_k \neq \emptyset\}$
while $|\mathcal{S}| < K$ **do**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{\arg\max_{p\in\mathcal{G}\setminus\mathcal{S}} \sum_{s\in\mathcal{S}} \|p - s\|_2\}$ ▷ Maximize dispersion
end while
return $\mathcal{S}$

first derive a binary grounding mask $\mathbf{M} \in \{0,1\}^{H\times W}$, where $\mathbf{M}_{i,j} = 1$ explicitly marks the editable foreground and 0 the preserved background region.

Prompt-Driven Segmentation. We implement the grounding function $\mathcal{G}$ with the zero-shot generalization of segmentation foundation models [5, 49], formulated as $\mathbf{M} = \mathcal{G}(\mathbf{I}, \mathcal{P})$, where $\mathcal{P}$ are the spatial cues derived from the instruction. For explicit spatial inputs (e.g., boxes or points) or descriptive text prompts, this unified multi-modal backbone directly extracts object masks $\mathbf{M}$ with pixel-level precision, thereby establishing a solid boundary for subsequent editing.

Attention-Guided Refinement. To mitigate the brittleness of text-only grounding, we add a fallback leveraging the DLLM's internal semantics. We compute a global heatmap $\mathbf{H} = \frac{1}{LK}\sum_{l,k} \mathbf{A}^{(l,k)}$ by averaging the visual-text cross-attention maps $\mathbf{A}^{(l,k)}$ across all layers and heads, and extract high-activation points $\mathcal{P}_{attn} = \{(x, y) \mid \mathbf{H}_{x,y} \in \text{top-}k(\mathbf{H})\}$ [59] as foreground prompts for the segmentation model, ensuring robust masks in complex scenarios.

3.3 High-Fidelity Visual Refinement

Building on the semantic fidelity established in the inversion stage, we introduce a unified refinement module to further improve visual coherence and boundary alignment, formulated as two complementary stages, **intrinsic refinement** and **residual recovery**, via simple set-theoretic operations on region masks. Let $\mathbf{M}_{\text{src}}$ denote the grounding mask of the original object (the region to be modified)[5] and $\mathbf{M}_{\text{tgt}}$ the mask of the newly generated entity; carefully manipulating their interplay resolves texture inconsistencies and ensures seamless integration.

Intrinsic Refinement. To resolve low-fidelity textures within generated regions, we compute the confidence map $\mathbf{C}$ of the edited image and identify unstable tokens $\mathcal{U} = \{(i, j) \mid \mathbf{C}_{i,j} < \tau\}$ with threshold τ. The refinement mask is their intersection with the target, $\mathbf{M}_{\text{conf}} = \mathbf{M}_{\mathcal{U}} \cap \mathbf{M}_{\text{tgt}}$, within which we re-sample to correct artifacts while keeping high-confidence structures intact.

Residual Recovery. To address shape mismatches between source and target objects, we recover the background over the *residual region* $\mathbf{M}_{\text{res}} = \mathbf{M}_{\text{src}} \setminus$

[5] For targets with substantially different geometry, we relax $\mathbf{M}_{\text{src}}$ to the bounding box of the original object for more spatial freedom.

$\mathbf{M}_{\text{tgt}}$. For *Replace* and *Remove*, $\mathbf{M}_{\text{res}}$ captures the exposed background gap for inpainting, whereas for *Add* it delineates the blending boundary with the original image, ensuring seamless integration across all editing modes.

Category	Value
Total Images	805
Editing Cases	805
Sub-instructions	1,610
Avg. Resolution	1193×831
Max. Resolution	1630×1420

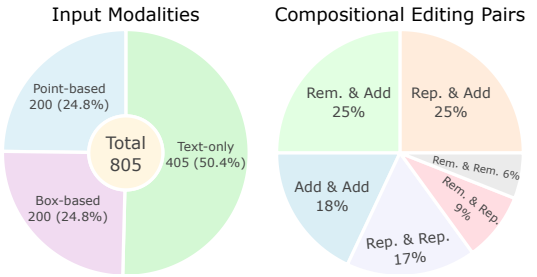

Fig. 3: Statistical overview of GIDE-Bench. It contains 805 cases and 1,610 sub-instructions (left), stratified across three input modalities (middle). The right chart shows the distribution of compositional editing pairs (Rem.: Remove, Rep.: Replace).

4 GIDE-Bench

We introduce GIDE-Bench to evaluate compositional image editing, comprising 805 high-quality cases each paired with sub-instructions. To assess robustness across grounding signals, the benchmark is further stratified into point-based, box-based, and text-only modalities. Detailed statistics are presented in Fig. 3.

4.1 Data Collection

We construct GIDE-Bench based on the OmniEdit dataset [64]. From an initial pool, we filter for high-quality images and leverage GPT-4o to generate coherent compositional instructions involving combinations of replace, add, and remove operations. We strictly enforce order-invariance by ensuring that executing sub-instructions in any sequence yields the same result. We then conduct rigorous human verification to eliminate ambiguous samples, resulting in 805 refined pairs.

To generate spatial grounding signals, we randomly sample subsets for point and box annotations. For point-based cases, we employ a spatial diversity strategy (Algorithm 2) to select four distinct foreground points, ensuring coverage of the object's extent. For box-based cases, we compute the minimal bounding box enclosing the target mask. This process yields a diverse benchmark covering point, box, and text modalities to facilitate comprehensive evaluations.

4.2 Evaluation Metrics

To rectify generation-induced spatial shifts, we follow prior work [48] and first align the edited image to the original one via SIFT keypoints [36], FLANN-based feature matching [40], and affine transformations. Within this aligned

Table 1: Performance comparison on GIDE-Bench. Best results within each group are highlighted in **bold**. Gray indicates our method. Notably, our GIDE integrated with Lumina-DiMOO achieves state-of-the-art performance among training-free methods, steadily narrowing the performance gap to top-tier fully supervised models.

Method	Non-edit			Edit_{GPT}		$\text{Edit}_{\text{Gemini}}$	
	MSE ↓	PSNR ↑	SSIM ↑	SC ↑	PQ ↑	SC ↑	PQ ↑
End-to-End Image Editing Models							
OneDiffusion [28]	1247.74	20.61	0.6831	1.81	1.79	1.94	1.81
InstructPix2Pix [3]	2540.06	17.71	0.7108	1.89	1.86	1.87	1.68
MagicBrush [74]	2410.44	19.36	0.7210	2.13	2.04	2.08	1.74
Lumina-DiMOO [68]	1208.22	19.80	0.6461	3.10	2.92	3.10	2.34
OmniGen2 [66]	1927.87	19.88	0.7368	3.65	3.36	3.67	3.13
FLUX.1-Kontext [27]	2321.48	16.91	0.5900	3.94	3.61	3.94	3.35
Qwen-Image [65]	1758.97	21.20	0.7902	4.38	4.15	4.33	3.98
Qwen-Image-Edit [65]	1741.50	21.29	0.7941	4.37	4.11	4.41	4.26
Qwen-Image-2.0 [76]	1445.26	19.66	0.7525	4.50	4.28	4.64	4.34
LongCat [58]	1166.27	20.26	0.7266	4.51	4.33	4.52	4.18
Edit-R1 [30]	1557.25	18.81	0.7083	4.64	4.46	4.55	4.19
Nano-Banana-1 [17]	**687.79**	**23.83**	**0.8338**	4.48	4.23	4.56	4.35
GPT-Image-1 [44]	5080.22	12.31	0.4714	**4.71**	**4.66**	**4.60**	**4.46**
Training-free Editing Methods applied to Base Models							
DirectInversion+P2P [24]	3008.64	14.54	0.5848	2.04	2.00	2.07	1.75
DirectInversion+PnP [24]	2126.71	16.33	0.6404	2.19	2.15	2.15	1.93
DICE+Lumina-DiMOO [19]	8323.89	9.38	0.3866	2.81	2.71	2.94	2.45
GIDE+MMaDA	3891.24	14.00	0.5522	2.96	2.80	2.74	2.54
GIDE+Lumina-DiMOO	**1224.89**	**20.40**	**0.7083**	**4.47**	**3.98**	**4.26**	**3.78**

space, we further distinguish edited and non-edited regions localized by our grounding module. Unlike the static or coarse masks commonly used in prior work [24, 48, 74], we dynamically define the edited region by operation type: the union of source and target subjects for *Replace*, the target for *Add*, and the source for *Remove*, with the non-edited region as its complement. Following ImgEdit [72], we independently leverage GPT-4o and Gemini-2.5-Pro [10] to assess *Semantic Correctness* (SC) and *Perceptual Quality* (PQ) within the edited region, enforcing $\text{PQ} \leq \text{SC}$ since visual quality is only meaningful once the instruction is semantically satisfied. For non-edited regions, we measure background preservation via standard pixel-level metrics (MSE, PSNR, SSIM).

5 Experiments

5.1 Experimental Setup

As a general training-free image editing method, we build GIDE upon two popular DLLMs, Lumina-DiMOO [68] and MMaDA [71], which are original developed for text-to-image generation. We compare GIDE with the following baselines:

- **Training-free methods**, including DICE [19][6], Direct Inversion with Plug-and-Play (PnP) [62] and Prompt-to-Prompt (P2P) [20][7].
- **Training-based methods**, including closed-source (GPT-Image-1 [44], Nano-Banana-1 [17]) and open-source models (Edit-R1 [30], LongCat [58], Qwen-Image [65], Qwen-Image-Edit [65], Qwen-Image-2.0 [76], FLUX.1-Kontext [27], OmniGen2 [66], MagicBrush [74], InstructPix2Pix [3], and OneDiffusion [28]), alongside Lumina-DiMOO's official pipeline as a strong baseline.

To align with their respective input formats, GIDE and training-free baselines adopt multi-instructions step-by-step, while end-to-end models process multi-instruction inputs jointly. Similarly, training-free methods use global descriptions extracted via GPT-4o [43], and instruction-guided end-to-end models directly use raw instructions. Closed-source models are accessed via API calls, all other models or methods are evaluated on a single A100 GPU. Additionally, all evaluated models employ their respective default hyperparameters in our experiments.

5.2 Experimental Results

Table 1 shows the quantitative evaluation of our proposed GIDE framework on GIDE-Bench, compared against three representative training-free methods and eleven supervised end-to-end models to demonstrate its robust capabilities.

Effectiveness of Editing on Target Regions. Our GIDE framework consistently demonstrates superior performance in SC while still maintaining a high PQ[8]. We detail its key advantages across several complementary dimensions:

- **Superiority over Training-Free Methods:** GIDE achieves remarkable gains over training-free baselines, surpassing DICE by 51.83% in SC and 50.39% in PQ. This margin extends further against other top-performing training-free methods (101.15% in SC, 90.20% in PQ), confirming that coupling discrete inversion with precise grounding is essential for robust editing.
- **Competitiveness with Supervised End-to-End Models:** Remarkably, GIDE surpasses the average performance of the 11 supervised models by 22.49% in SC and 17.54% in PQ. This demonstrates that a training-free framework can deliver highly competitive results, steadily closing the distance to top-performing closed-source models like Nano-Banana-1.
- **Addressing Base Model Limitations:** The image-to-image baseline of Lumina-DiMOO often misinterprets prompts and fails to accurately follow instructions (Fig. 4). By resolving these ambiguities through grounded inversion, GIDE significantly boosts SC by 40.81% and PQ by 47.53%.
- **Impact of Backbone Selection:** Weaker backbones like MMaDA struggle with fixed resolutions and complex operations (e.g., unnatural *Add*/*Remove* transitions). Adopting the powerful Lumina-DiMOO allows GIDE to overcome these bottlenecks, ensuring consistent, high-quality edits.

[6] Since DICE does not open-source its code, we implement it by ourself based on a DLLM Lumina-DiMOO.

[7] Both PnP and P2P are based on Stable-Diffusion-v1.4 [51].

[8] All SC and PQ scores reported herein are averaged between GPT and Gemini.

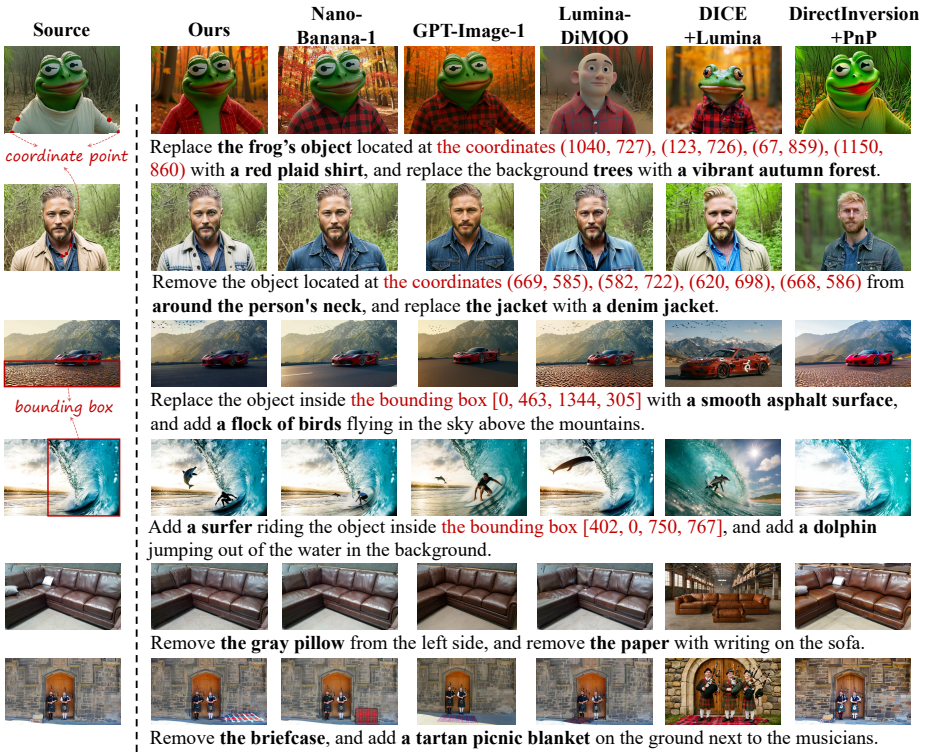

Fig. 4: Qualitative comparison on GIDE-Bench. GIDE accurately follows the given instructions and faithfully preserves structural fidelity, avoiding the unintended alterations (e.g., DICE) and aspect ratio distortions (e.g., GPT-Image-1) while consistently achieving photorealistic consistency. Zoom in for a better view.

Preservation of Non-edited Regions. Beyond pure editing accuracy, GIDE also exhibits an exceptional capability in faithfully preserving non-edited content, thereby effectively avoiding the common pitfalls of previous methods:

- **Advantage over Global Inversion:** Methods like DICE treat logits from a single forward step as the ground-truth $\boldsymbol{y}_0$, perturbing the token distribution and severely degrading background fidelity. By leveraging precise localization, GIDE reduces MSE by 85.28% against DICE, while improving PSNR by 117.48% and SSIM by 83.21%, ensuring superior visual consistency.
- **Optimal Performance Trade-off:** Compared to models heavily optimized for editing (such as GPT-Image-1), GIDE offers a much better overall balance. It effectively prevents background distortion by reducing MSE by 75.89% and boosting PSNR and SSIM by 65.72% and 50.25%, respectively.
- **Fidelity Constraints of DLLMs:** While GIDE effectively preserves visual semantics, its low-level fidelity still lags behind Nano-Banana-1. This is attributed to the inherent reconstruction loss of the VQModel in Lumina-DiMOO. As DLLM autoencoders evolve, this gap will naturally close.

Table 2: Evaluation on ImgEdit [72] benchmark. Rep.: Replace, Rem.: Remove.

Method	Rep.↑	Add↑	Rem.↑	Method	Rep.↑	Add↑	Rem.↑
MagicBrush [74]	1.97	2.84	1.58	OmniGen [67]	2.94	3.47	2.43
AnyEdit [73]	2.47	3.18	2.23	BAGEL [12]	3.30	3.56	2.62
UltraEdit [77]	2.96	3.44	1.45	UniWorld-V1 [32]	3.47	3.82	3.24
ICEdit [75]	3.15	3.58	2.93	Lumina [68]	3.83	3.82	2.76
Step1X-Edit [34]	**3.40**	**3.88**	**2.41**	**GIDE (Ours)**	**4.22**	**3.90**	**3.84**

Table 3: Sensitivity analysis of the mixing coefficient λ and the refinement scaling factor γ. The default setting is highlighted in gray and the best scores in **bold**.

λ	Non-edit			Edit_{GPT}		Edit_{Gemini}	
	MSE ↓	PSNR ↑	SSIM ↑	SC ↑	PQ ↑	SC ↑	PQ ↑
0.0	**1224.62**	20.35	**0.7086**	4.30	3.86	4.12	3.71
0.1	1306.47	20.22	0.7066	4.26	3.82	4.12	3.69
0.2	1224.89	20.40	0.7083	**4.47**	**3.98**	**4.26**	**3.78**
0.3	1308.57	20.33	0.7075	4.28	3.85	4.09	3.70
0.4	1225.40	**20.51**	0.7084	4.21	3.83	4.03	3.75

γ	Non-edit			Edit_{GPT}		Edit_{Gemini}	
	MSE ↓	PSNR ↑	SSIM ↑	SC ↑	PQ ↑	SC ↑	PQ ↑
0.3	1318.21	20.22	0.7063	4.46	3.91	4.20	3.52
0.4	1276.95	20.28	0.7077	4.47	3.95	4.22	3.74
0.5	1224.89	**20.40**	0.7083	4.47	**3.98**	4.26	3.78
0.6	1232.64	20.32	0.7064	**4.48**	3.97	4.26	3.79
0.7	**1217.27**	20.36	**0.7086**	4.47	**3.98**	**4.28**	**3.81**

5.3 Qualitative Comparison

Figure 4 compares GIDE against competing baselines across six different combinations of the *Replace*, *Add*, and *Remove* tasks. GIDE accurately executes the given instructions while strictly preserving unedited regions and overall structural fidelity, whereas training-free methods (e.g., DICE) and Lumina-DiMOO lack precise grounding and produce prominent visual artifacts, and GPT-Image-1 alters aspect ratios and introduces noticeable structural distortions. Moreover, by injecting source priors (e.g., lighting and texture) through its discrete inversion, GIDE further ensures strong photorealism and physical consistency: in Row 1 it faithfully retains the original shallow depth of field, thus avoiding the unnatural background sharpening of Nano-Banana-1, and in Row 4 it faithfully renders a realistic backlit silhouette for the inserted surfer, thereby mitigating the "copy-paste" artifacts produced by Nano-Banana-1 and GPT-Image-1.

5.4 Evaluation on ImgEdit Benchmark

To validate broad applicability, we further evaluate GIDE on the ImgEdit benchmark across three tasks (*Replace*, *Add*, *Remove*) with metrics computed by GPT-4o, comparing against Lumina-DiMOO's image-to-image pipeline and representative methods (MagicBrush, AnyEdit [73], UltraEdit [77], ICEdit [75], Step1X-Edit [34], OmniGen [67], BAGEL [12], and UniWorld-V1 [32]). As shown in Table 2, GIDE outperforms all baselines, improving over Lumina-DiMOO by 10.18%, 2.09%, and 39.13% on the respective tasks. The large gain on *Remove* highlights how our precise grounding and specialized inversion overcome the baseline's difficulty in cleanly erasing objects from the original scene.

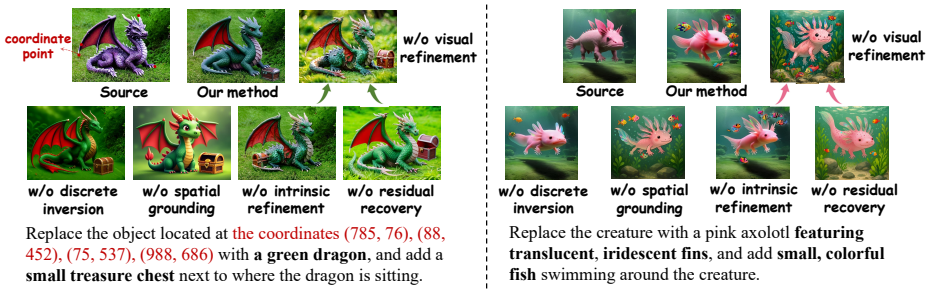

Fig. 5: Visual ablation on the proposed components. Removing *discrete inversion* distorts the overall object shape, dropping *spatial grounding* loses the surrounding background, and removing *visual refinement* blurs fine-grained details.

Table 4: Ablation experiments. We compare the full method against variants with specific modules removed. The results confirm that the full framework achieves the best performance in both background preservation and editing quality.

Method	Non-edit			Edit_{GPT}		$\text{Edit}_{\text{Gemini}}$	
	MSE ↓	PSNR ↑	SSIM ↑	SC ↑	PQ ↑	SC ↑	PQ ↑
Ours (Full)	**1224.89**	**20.40**	**0.7083**	**4.47**	**3.98**	**4.26**	**3.78**
w/o discrete inversion	$2196.29_{79\%\uparrow}$	$17.85_{13\%\downarrow}$	$0.6628_{6\%\downarrow}$	$4.30_{4\%\downarrow}$	$3.83_{4\%\downarrow}$	$4.08_{4\%\downarrow}$	$3.28_{13\%\downarrow}$
w/o spatial grounding	$8446.17_{590\%\uparrow}$	$9.42_{54\%\downarrow}$	$0.4161_{41\%\downarrow}$	$3.60_{19\%\downarrow}$	$3.43_{14\%\downarrow}$	$2.68_{37\%\downarrow}$	$2.21_{42\%\downarrow}$
w/o visual refinement	$2476.78_{102\%\uparrow}$	$17.23_{16\%\downarrow}$	$0.6444_{9\%\downarrow}$	$3.73_{17\%\downarrow}$	$3.25_{18\%\downarrow}$	$3.56_{16\%\downarrow}$	$2.67_{29\%\downarrow}$
w/o intrinsic refinement	$1424.29_{16\%\uparrow}$	$19.80_{3\%\downarrow}$	$0.6987_{1\%\downarrow}$	$3.96_{11\%\downarrow}$	$3.51_{12\%\downarrow}$	$3.87_{9\%\downarrow}$	$3.11_{18\%\downarrow}$
w/o residual recovery	$2316.94_{89\%\uparrow}$	$17.65_{13\%\downarrow}$	$0.6606_{7\%\downarrow}$	$4.16_{7\%\downarrow}$	$3.80_{5\%\downarrow}$	$3.85_{10\%\downarrow}$	$3.21_{15\%\downarrow}$

5.5 Sensitivity Analysis

We analyze the sensitivity of two key hyper-parameters in Table 3: the mixing coefficient λ controlling the inversion residual strength, and the scaling factor γ setting the refinement threshold $\tau = \gamma \cdot \max(\mathbf{C})$ as a fraction of the confidence map $\mathbf{C}$'s maximum value. The default setting $\lambda = 0.2$ strikes an effective balance, achieving the best overall editing quality while keeping the non-edited metrics near-optimal. Moreover, performance remains highly stable across $\lambda \in [0, 0.4]$ and $\gamma \in [0.3, 0.7]$ with only marginal fluctuations, clearly demonstrating that GIDE is robust to both and does not rely on careful manual tuning.

5.6 Ablation Study

We ablate our components in Fig. 5 and Table 4. First, the **discrete inversion** is critical for structural consistency: replacing it with standard inpainting degrades semantic metrics (SC/PQ) and distorts fine-grained geometry, e.g., the dragon and the axolotl. Second, the **spatial grounding** preserves the background; its absence forces global editing, causing a 589.55% MSE increase and severe hallucination (e.g., the grassland). Finally, the **visual refinement** ensures high fidelity and seamless integration: omitting *intrinsic refinement* blurs textures,

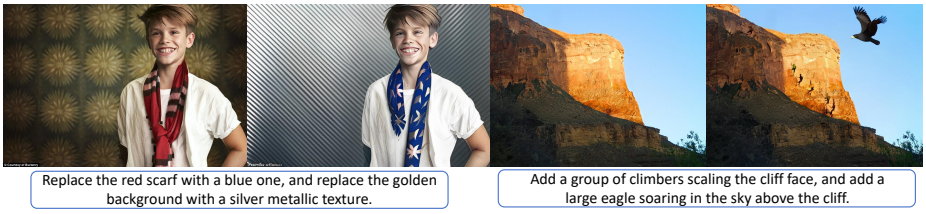

Fig. 6: Current limitations. VQModel reconstruction loss degrades the boy's eyes and teeth (left), while the DLLM struggles with the "group of climbers" (right).

while removing *residual recovery* causes boundary artifacts, and disabling the whole module incurs a 102.20% MSE penalty and a 23.71% PQ decline.

6 Conclusions and Limitations

In this work, we present GIDE, the first training-free image editing framework specifically tailored for DLLMs. By coupling a principled discrete inversion mechanism with a compositional grounding and refinement pipeline, GIDE effectively bridges the gap between discrete representations and high-fidelity editing. We further establish GIDE-Bench, a benchmark emphasizing compositional instructions and diverse grounding modalities. Extensive experiments show that GIDE significantly outperforms existing training-free baselines and achieves photorealistic consistency comparable to state-of-the-art supervised methods.

Its performance is nonetheless bounded by the reconstruction loss of VQ-based tokenizers, which can distort high-frequency details (Fig. 6, left), and the DLLM's limited capacity for fine-grained generation, which often blurs tiny added objects (Fig. 6, right). As a generic framework, GIDE is naturally expected to overcome these bottlenecks as DLLM architectures continue to evolve.

Acknowledgements

This work is supported by the New Generation Artificial Intelligence–National Science and Technology Major Project (No. 2025ZD0123001), the National Natural Science Foundation of China (No. 62272374, No. 62192781), the Natural Science Foundation of Shaanxi Province (No. 2024JC-JCQN-62), the State Key Laboratory of Communication Content Cognition under Grant No. A202502, and the Key Research and Development Project in Shaanxi Province (No. 2023GXLH-024). This work is also partially supported by the National Natural Science Foundation of China (Grant No. 62306261) and the SHIAE Grant (No. 8115074), as well as by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. This work is also partially supported by the Hong Kong RGC Strategic Topics Grant STG1/E-403/24-N and the CUHK–CUHK(SZ)–GDST Joint Collaboration Fund YSP26-4760949.

References

1. Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., Cohen-Or, D.: Stable flow: Vital layers for training-free image editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7877–7888 (2025)
2. Bai, C., Chen, J., Bai, X., Chen, Y., She, Q., Lu, M., Zhang, S.: Uniedit-i: Training-free image editing for unified vlm via iterative understanding, editing and verifying. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29750–29759 (2026)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
4. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K.P., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. In: International Conference on Machine Learning. pp. 4055–4075. PMLR (2023)
7. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
8. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
9. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8795–8805 (2024)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
11. Dao, Q., He, X., Han, L., Nguyen, N.H., Nobar, A.H., Ahmed, F., Zhang, H., Nguyen, V.A., Metaxas, D.: Discrete noise inversion for next-scale autoregressive text-based image editing. arXiv preprint arXiv:2509.01984 (2025)
12. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
14. Fu, F., Zhang, L., Huang, M., Mao, Z.: Feededit: Text-based image editing with dynamic feedback regulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2661–2670 (2025)

15. Ge, Y., Zhao, S., Li, C., Ge, Y., Shan, Y.: Seed-data-edit technical report: A hybrid dataset for instructional image editing. arXiv preprint arXiv:2405.04007 (2024)
16. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
17. Google: Gemini 2.5 flash image (nano banana). `https://aistudio.google.com/models/gemini-2-5-flash-image` (2025)
18. Gu, X., Li, M., Zhang, L., Chen, F., Wen, L., Luo, T., Zhu, S.: Multi-reward as condition for instruction-based image editing. In: The Thirteenth International Conference on Learning Representations (2025)
19. He, X., Han, L., Quan Dao, S.W., Bai, M., Di Liu, H.Z., Juefei-Xu, F., Tan, C., Liu, B., Min, M.R., Li, K., et al.: Dice: Discrete inversion enabling controllable editing for masked generative models (2025)
20. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023)
21. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
22. Hu, T., Li, L., Wang, K., Wang, Y., Yang, J., Cheng, M.M.: Anchor token matching: Implicit structure locking for training-free ar image editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18166–18176 (2025)
23. Hu, Y., Tang, Z., Jiang, X., Li, W., Zhang, J.: Talkphoto: A versatile training-free conversational assistant for intelligent image editing. arXiv preprint arXiv:2601.01915 (2026)
24. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In: The Twelfth International Conference on Learning Representations (2024)
25. Kim, J., Park, J., Song, Y., Kwak, N., Rhee, W.: Reflex: Text-guided editing of real images in rectified flow via mid-step feature extraction and attention adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15939–15948 (2025)
26. Krojer, B., Vattikonda, D., Lara, L., Jampani, V., Portelance, E., Pal, C., Reddy, S.: Learning action and reasoning-centric image editing from videos and simulation. Advances in Neural Information Processing Systems **37**, 38035–38078 (2024)
27. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
28. Le, D.H., Pham, T., Lee, S., Clark, C., Kembhavi, A., Mandt, S., Krishna, R., Lu, J.: One diffusion to generate them all. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2671–2682 (2025)
29. Li, T., Ku, M., Wei, C., Chen, W.: Dreamedit: Subject-driven image editing. arXiv preprint arXiv:2306.12624 (2023)
30. Li, Z., Liu, Z., Zhang, Q., Lin, B., Yuan, S., Yan, Z., Ye, Y., Yu, W., Niu, Y., Yuan, L.: Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. arXiv preprint arXiv:2510.16888 (2025)
31. Liang, Z., Li, Y., Yang, T., Wu, C., Mao, S., Nian, T., Pei, L., Zhou, S., Yang, X., Pang, J., et al.: Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. arXiv preprint arXiv:2508.20072 (2025)

32. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
33. Liu, D., Zhao, S., Zhuo, L., Lin, W., Xin, Y., Li, X., Qin, Q., Qiao, Y., Li, H., Gao, P.: Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. arXiv preprint arXiv:2408.02657 (2024)
34. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
35. Liu, X., Song, Y., Liu, Z., Huang, Z., Guo, Q., He, Z., Qiu, X.: Longllada: Unlocking long context capabilities in diffusion llms. arXiv preprint arXiv:2506.14429 (2025)
36. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
37. Mo, S., Mu, F., Lin, K.H., Liu, Y., Guan, B., Li, Y., Zhou, B.: Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7465–7475 (2024)
38. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: 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)
39. Morita, R., Frolov, S., Moser, B.B., Shirakawa, T., Watanabe, K., Dengel, A., Zhou, J.: Tkg-dm: Training-free chroma key content generation diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13031–13040 (2025)
40. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: International conference on computer vision theory and applications. vol. 1, pp. 331–340. Scitepress (2009)
41. Nguyen, T., Do, K., Kieu, D., Nguyen, T.: h-edit: Effective and flexible diffusion-based editing via doob's h-transform. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28490–28501 (2025)
42. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., ZHOU, J., Lin, Y., Wen, J.R., Li, C.: Large language diffusion models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
43. OpenAI: Hello gpt-4o. News announcement by OpenAI (2024), `https://openai.com/index/hello-gpt-4o/`
44. OpenAI: Gpt image 1: State-of-the-art image generation model. `https://platform.openai.com/docs/models/gpt-image-1` (2025)
45. Pan, Y., He, X., Mao, C., Han, Z., Jiang, Z., Zhang, J., Liu, Y.: Ice-bench: A unified and comprehensive benchmark for image creating and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16586–16596 (2025)
46. Pathiraja, B., Patel, M., Singh, S., Yang, Y., Baral, C.: Refedit: A benchmark and method for improving instruction-based image editing model on referring expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15646–15656 (2025)
47. Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., Han, C., Ge, Z., Zhang, X., Xia, S.T.: Dreambench++: A human-aligned benchmark for personalized image generation. In: The Thirteenth International Conference on Learning Representations (2025)

48. Qian, Y., Lu, J., Fu, T.J., Wang, X., Chen, C., Yang, Y., Hu, W., Gan, Z.: Giebench: Towards grounded evaluation for text-guided image editing. arXiv preprint arXiv:2505.11493 (2025)
49. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: The Thirteenth International Conference on Learning Representations (2025)
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
52. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
53. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8871–8879 (2024)
54. Shi, Q., Bai, J., Zhao, Z., Chai, W., Yu, K., Wu, J., Song, S., Tong, Y., Li, X., Li, X., et al.: Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. arXiv preprint arXiv:2505.23606 (2025)
55. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
56. Tang, H., Wu, Y., Yang, S., Xie, E., Chen, J., Chen, J., Zhang, Z., Cai, H., Lu, Y., Han, S.: Hart: Efficient visual generation with hybrid autoregressive transformer. In: The Thirteenth International Conference on Learning Representations (2025)
57. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
58. Team, M.L., Ma, H., Tan, H., Huang, J., Wu, J., He, J.Y., Gao, L., Xiao, S., Wei, X., Ma, X., et al.: Longcat-image technical report. arXiv preprint arXiv:2512.07584 (2025)
59. Tewel, Y., Gal, R., Samuel, D., Atzmon, Y., Wolf, L., Chechik, G.: Add-it: Training-free object insertion in images with pretrained diffusion models. In: The Thirteenth International Conference on Learning Representations (2025)
60. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
61. Tian, Y., Yang, L., Yang, J., Wang, A., Tian, Y., Zheng, J., Wang, H., Teng, Z., Wang, Z., Wang, Y., et al.: Mmada-parallel: Multimodal large diffusion language models for thinking-aware editing and generation. arXiv preprint arXiv:2511.09611 (2025)
62. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: 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)
63. Wang, Y., Yang, S., Zhao, B., Zhang, L., Liu, Q., Zhou, Y., Xie, C.: Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. arXiv preprint arXiv:2507.21033 (2025)

64. Wei, C., Xiong, Z., Ren, W., Du, X., Zhang, G., Chen, W.: Omniedit: Building image editing generalist models through specialist supervision. In: The Thirteenth International Conference on Learning Representations (2025)
65. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
66. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
67. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13294–13304 (2025)
68. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al.: Lumina-dimoo: An omni diffusion large language model for multimodal generation and understanding. arXiv preprint arXiv:2510.06308 (2025)
69. Xin, Y., Zhuo, L., Qin, Q., Luo, S., Cao, Y., Fu, B., He, Y., Li, H., Zhai, G., Liu, X., et al.: Resurrect mask autoregressive modeling for efficient and scalable image generation. arXiv preprint arXiv:2507.13032 (2025)
70. Xu, R., Xi, W., Wang, X., Mao, Y., Cheng, Z.: Stylessp: Sampling startpoint enhancement for training-free diffusion-based method for style transfer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18260–18269 (2025)
71. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025)
72. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
73. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
74. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
75. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
76. Zhao, B., Wu, C., Li, D., Meng, H., Li, J., Zhang, J., Zhou, J., Lin, J., Gao, K., Cao, K., et al.: Qwen-image-2.0 technical report. arXiv preprint arXiv:2605.10730 (2026)
77. Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. Advances in Neural Information Processing Systems **37**, 3058–3093 (2024)
78. Zhu, H., Zhu, Z., Zhang, K., Gong, Y., Liu, Y., Bai, X.: Training-free geometric image editing on diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19130–19140 (2025)
79. Zhu, T., Zhang, S., Shao, J., Tang, Y.: Kv-edit: Training-free image editing for precise background preservation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16607–16617 (2025)