LayerEdit: Disentangled Multi-Object Editing via Conflict-Aware Multi-Layer Learning

Fengyi Fu[1], **Mengqi Huang**[1] *, **Lei Zhang**[1], **Zhendong Mao**[1 2]

[1]University of Science and Technology of China, Hefei, China;
[2]Institute of Artificial intelligence, Hefei Comprehensive National Science Center, Hefei, China;
{ff142536f, huangmq}@mail.ustc.edu.cn, {leizh23, zdmao}@ustc.edu.cn

Abstract

Text-driven multi-object image editing which aims to precisely modify multiple objects within an image based on text descriptions, has recently attracted considerable interest. Existing works primarily follow the localize-editing paradigm, focusing on independent object localization and editing while neglecting critical inter-object interactions. However, this work points out that the neglected attention entanglements in inter-object conflict regions, inherently hinder disentangled multi-object editing, leading to either inter-object editing leakage or intra-object editing constraints. We thereby propose a novel multi-layer disentangled editing framework ***LayerEdit***, a training-free method which, for the first time, through precise object-layered decomposition and coherent fusion, enables conflict-free object-layered editing. Specifically, *LayerEdit* introduces a novel "decompose-editing-fusion" framework, consisting of: (1) *Conflict-aware Layer Decomposition module*, which utilizes an attention-aware IoU scheme and time-dependent region removing, to enhance conflict awareness and suppression for layer decomposition. (2) *Object-layered Editing module*, to establish coordinated intra-layer text guidance and cross-layer geometric mapping, achieving disentangled semantic and structural modifications. (3) *Transparency-guided Layer Fusion module*, to facilitate structure-coherent inter-object layer fusion through precise transparency guidance learning. Extensive experiments verify the superiority of *LayerEdit* over existing methods, showing unprecedented intra-object controllability and inter-object coherence in complex multi-object scenarios.

Code — https://github.com/fufy1024/LayerEdit

1 Introduction

Recent diffusion-based image generation models have revolutionized the Text-driven Image Editing (TIE) task (Wei et al. 2025). TIE aims to perform arbitrary text-driven image modifications with a good preservation of irrelevant regions, offering powerful applications for advertising, photography, social media, and so on. Given that real-world images typically consist of multiple objects with intricate spatial and attribute relationships, recent methods (Chakrabarty et al. 2024) extend TIE to more complex text-driven multi-object editing (TMOE) scenarios. Compared to simple single-object editing, the primary goal of TMOE is dual-faceted: (1) ***intra-object controllability***, *i.e.*, precise text-controlled editing for each target object, without being constrained by its original spatial boundaries; (2) ***inter-object coherence***, *i.e.*, coherent object-specific editing across multiple objects, without unintended modifications on non-target objects.

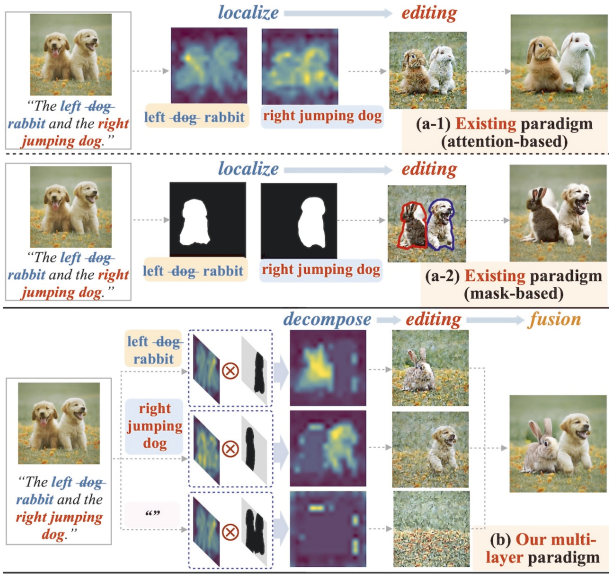

Figure 1: Illustration of motivation. (a) *Existing paradigm*: suffers from inaccurate disentanglement across conflict regions, resulting in (a-1) inter-object editing leakage or (a-2) intra-object editing artifacts. (b) *Our multi-layer disentangled editing paradigm*: by modeling conflict-aware object-layered decomposition and structure-coherent fusion, bringing conflict-free and accurate multi-object editing.

Recent works take their effort on two aspects toward these goals. The first relies on attention-based text-image alignment (Epstein et al. 2024; Guo and Lin 2024; Huang et al. 2024) for target object localization, by iterative object-specific editing at multi-step (Huang et al. 2024) or multi-turn (Brack et al. 2024) to maintain multi-object coherence. The other methods incorporate segmentation (Chakrabarty et al. 2024; Yang et al. 2024) or visual language models

* Mengqi Huang is the corresponding author.

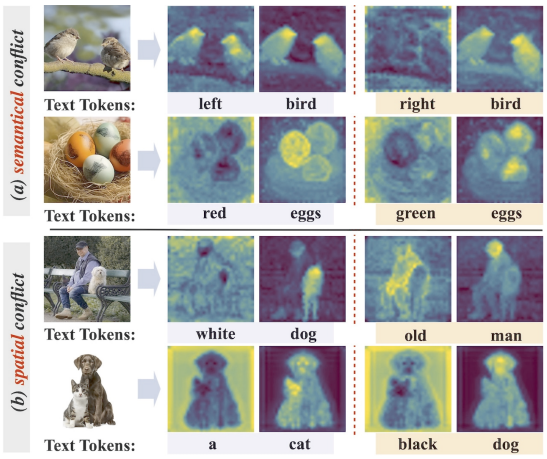

Figure 2: Visualization of cross-attention maps. Key observation is that model exhibits text-image attention misalignment in both: (a) *semantical*, and (b) *spatial conflict regions*.

(Wang et al. 2024; Feng et al. 2024; Schouten et al. 2025) to obtain more accurate object location, combined with loss regularization (Zhu et al. 2025) or mask-weighted fusion (Chakrabarty et al. 2024; Yang et al. 2024) to minimize unintended modifications in non-target regions. In general, existing methods primarily follow a *localize-editing* paradigm, *i.e.*, independently performing intra-object localizing and editing, without considering inter-object interactions.

However, we argue that existing frameworks, fixated on intra-object editing while ignoring inter-object interactions, inevitably hinder accurate disentanglement across multiple objects, ultimately compromising either inter-object coherence or intra-object controllability. The fundamental limitation stems from text-conditioned diffusion models' reliance on text-image attention alignment for feature manipulation: while effective for single-object editing, they catastrophically fail in multi-object scenarios due to attention entanglement caused by unmodeled inter-object interactions. As analyzed in Fig.2, initial denoising cross-attention maps reveal two characteristic attention alignment conflict patterns: (1) semantic conflicts in conceptually related regions (*e.g.*, Fig.2.a, multi-directional "birds" or multi-color "eggs"), and (2) spatial conflicts in overlapping regions (*e.g.*, Fig.2.b, occluded "cat" and "dog"). Such regions exhibiting semantic/spatial conflicts with target regions constitute ***conflict regions***, fundamentally hindering object disentanglement, posing limitations to two core editing goals: (1) Attention-based methods suffer from inadequate attention disentanglement of inter-object conflicts, causing inter-object editing leakage (*e.g.*, Fig.1.a-1 misediting "right dog" to "rabbit"). (2) Mask-based methods suffer from non-adaptive disentanglement with fixed intra-object spatial constraints, suppressing intra-object editing in necessary structural changes (*e.g.*, Fig.1.a-2, artifacts during "rabbit" morphing). Therefore, a reasonable TMOE framework should establish comprehensive multi-object disentanglement, ensuring both precise and region-unconstrained intra-object disentanglement, and conflict-constrained inter-object disentanglement.

To this end, we present ***LayerEdit***, a novel and training-free framework that reformulates multi-object editing as a multi-layer disentangled learning problem, which for the first time to our knowledge, explicitly models object-layered decomposition and coherent fusion, to achieve disentangled object editing free from inter-object conflicts and intra-object region constraints (Fig.1.b). Departing from conventional paradigm, the *core idea of LayerEdit* lies in fundamentally shifting the focus from mere intra-object target regions to inter-object conflict regions, which enables: conflict-free object disentangled editing through inter-object conflict awareness and suppression; and structure-coherent object-layered fusion through inter-object structural modeling. Technically, *LayerEdit* introduces an innovative "decompose-editing-fusion" architecture, consisting of: (1) *Conflict-aware Layer Decomposition module*, which integrates cross-attention and panoptic segmentations via an adaptive IoU (Intersection over Union) scheme for accurate conflict region identification, followed by a time-dependent region removing scheme for conflict-free layer decomposition. (2) *Object-layered Editing module*, which establishes coordinated intra-layer text guidance and cross-layer geometric mapping, for disentangled semantic and structural modifications. (3) *Transparency-guided Layer Fusion module*, which implements layer transparency learning conditioned on inter-object spatial features, to ensure structure-coherent fusion of overlapping multi-layer.

Our contributions are summarized as follows: (1) For the first time, we point out that existing frameworks ignore inter-object interaction conflicts, resulting in either inter-object editing leakage or intra-object editing constraint. We propose *LayerEdit*, a *training-free* and *plug-and-play* framework that, through precise object-layered decomposition and coherent multi-layer fusion, achieves disentangled object editing free from inter-object conflicts and intra-object constraints. (2) We propose a novel decompose-editing-fusion framework, with three tailored modules to respectively enhance the conflict-free object layer decomposition, the disentangled editing for arbitrary semantic and structural modifications, and the coherent fusion for overlapping layers. (3) Extensive experiments validate *LayerEdit*'s superiority, demonstrating significant improvements in both editability (12.6% in CLIP-T) and global quality (15.3% in FID).

2 Related Work

2.1 Text-Driven Image Editing

Recent diffusion models (Fu et al. 2025; Zhou et al. 2025) revolutionize the text-driven image editing task. Hertz et al. (2023); Tumanyan et al. (2023) first achieve the image editing through text interaction directly. Kawar et al. (2023); Li et al. (2023); Song et al. (2024) enable non-rigid image editing with complex text embedding optimization. Brooks, Holynski, and Efros (2023); Chen et al. (2025) expand the task into instruction-guided editing scenarios. Other works inject attention-level (Cao et al. 2023; Fu et al. 2025) or task-specific (Epstein et al. 2024; Nguyen et al. 2025) control into generation process, facilitating precise image manipulations. Although effective in single-object editing, these models

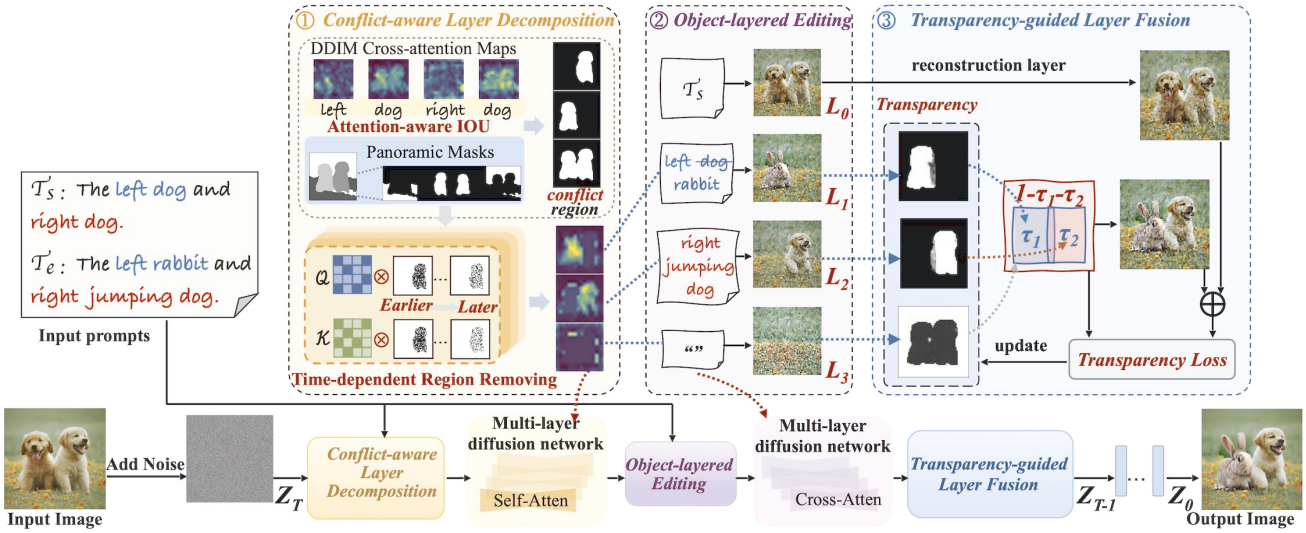

Figure 3: Overview of *LayerEdit*, consisting of: 1) Conflict-aware Layer Decomposition: precisely decompose object layers by identifying and constraining conflict regions; 2) Object-layered Editing: establish intra-layer text-guided editing (geometric editing detailed in Fig.5); 3) Transparency-guided Layer Fusion: enables structure-coherent fusion with transparency learning.

struggle with real-world multi-object scenarios, which have complex spatial and semantic inter-object relationships.

2.2 Multi-Object Image Editing

Following single-object editing work, Joseph et al. (2024); Brack et al. (2024) first propose an iterative editing framework to ensure multi-object coherent editing in a multi-turn editing process, but result in exponentially increasing calculation and time costs. Jia et al. (2024); Yang et al. (2024); Chakrabarty et al. (2024) adopt segmentation models to identify objects, based on mask-weighted fusion to concatenate various edited objects and unedited backgrounds, which impose strict region constraints and struggle with layout modifications. Patashnik et al. (2023); Guo and Lin (2024); Huang et al. (2024); Li et al. (2025) implicitly distinguish objects based on cross-attention maps, but easily entangle features of spatially overlapping or semantically related objects. However, different from existing methods, the core idea of *LayerEdit* is to explicitly model and resolve inter-object conflicts through a multi-layer decompose-editing-fusion framework, which is completely unexplored.

3 Preliminaries

Diffusion models consist of a forward process and a corresponding reverse process. The forward process gradually adds Gaussian noise to data, to generate the noisy sample with a predefined noise adding schedule α_t at timestep t:

$$z_t = \sqrt{\overline{\alpha}_t} z_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \quad \epsilon \in \mathcal{N}(0,1), \tag{1}$$

where $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, z_0 is a sample from data distribution.

The reverse process involves a parameterized noise prediction network ϵ_θ, to predict and iteratively remove the added noise. The training objective is to minimize the distance between two noises, based on the text condition $\mathcal{T}$:

$$\mathcal{L} = \mathbb{E}_{z_0,\mathcal{T},\epsilon\sim\mathcal{N}(0,1),t}[\| \epsilon - \epsilon_\theta(z_t, t, \mathcal{T}) \|]. \tag{2}$$

The feature incorporation in diffusion models is mainly based on a self-attention and cross-attention module, to update the spatial features $\phi(z_t)$ as $\hat{\phi}(z_t)$:

$$\hat{\phi}(z_t, \mathcal{T}) = \text{Softmax}(\frac{\mathcal{Q}\mathcal{K}^\top}{\sqrt{d}})\mathcal{V} = \mathcal{A}\mathcal{V}, \tag{3}$$

where $\mathcal{Q}$ is the query features projected from spatial features $\phi(z_t)$, and $\mathcal{K}$, $\mathcal{V}$ are the key and value features projected from the spatial features (in self-attention layers) or textual embedding (in cross-attention layers) with corresponding projection matrices. $\mathcal{A}$ is the calculated attention maps.

4 Method

Definition. Given a source image $\mathcal{I}$ and source text $\mathcal{T}_s$, the image editing task aims to synthesize the desired image with arbitrary rigid and non-rigid changes, based on editing prompt $\mathcal{T}_e$ (modified from $\mathcal{T}_s$). For multi-object editing scenarios, $\langle \mathcal{T}_s, \mathcal{T}_e \rangle$ can generally be represented as a unit of multiple editing pairs $\langle \mathcal{T}_s = \bigcup_i \mathcal{O}_s^i, \mathcal{T}_e = \bigcup_i \mathcal{O}_e^i \rangle_{i=1}^N$, where $\mathcal{O}_s^i$ and $\mathcal{O}_e^i$ indicate the i-th source-edited object in $\mathcal{T}_s$ and $\mathcal{T}_e$, and N is the total number of objects to be edited.

In this section, we detail the implementation of *LayerEdit*, an elegant decompose-editing-fusion framework shown in Fig.3, which first precisely decompose each object layer through intra-object conflict awareness and constraining (Sec.4.1), then explores the intra- and cross-layer feature processing to support semantic and structural editing functions (Sec.4.2), and finally realizes coherent layer fusion with structure-conditioned transparency learning (Sec.4.3).

4.1 Conflict-aware Layer Decomposition

Object-layered decomposition aims to construct exclusive editing layer for each object, free from attention entanglement and inter-object conflicts. By analyzing the attention

entanglement pattern in Fig.2, it stems from cross-attention misalignment in semantic or spatial conflict regions, hindering accurate localized editing. However, in contrast, this phenomenon also reveals that cross-attention facilitates conflict regions awareness. Thus, we design a novel attention-aware IoU scheme, by integrating cross-attention maps and panoptic segmentation to precisely localize conflict regions.

Attention-aware IoU. Specifically, following established practices (Guo and Lin 2024), we first aggregate the average cross-attention maps $\{\overline{\mathcal{A}}^i\}_{i=1}^N$ on object tokens $\{\mathcal{O}_s^i\}_{i=1}^N$ across all DDIM inversion timesteps. Then, by using segmentation models (SAM, (Kirillov et al. 2023); OneFormer, (Jain et al. 2023)), more precise object masks $\{\mathcal{M}_o^i\}_{i=1}^N$ and K panoramic segmentation masks $\{\mathcal{M}_{pan}^j\}_{i=1}^K$ are obtained. The attention-aware IoU is defined as:

$$\text{A-IoU}(i,j) = \frac{||\mathcal{M}_{pan}^j \bigcap \overline{\mathcal{A}}^i||}{||\mathcal{M}_{pan}^j \bigcup \overline{\mathcal{A}}^i||}, \forall i \in [1,N], j \in [1,K] \quad (4)$$

Intuitively, A-IoU(i,j) quantifies the affect weights of j-th panoptic region on the generation of i-th object. The conflict region for i-th object is defined as the high-affect regions (determined by threshold η) excluding object's own region:

$$\mathcal{M}_{con}^i = \bigcup_{\{j|\text{A-IoU}(i,j)>\eta\}} (\mathcal{M}_{pan}^j) - \mathcal{M}_o^i. \quad (5)$$

Time-dependent Region Removing. Having identified the conflict regions, decomposing the i-th object equates to suppressing its corresponding conflict regions. The *core challenge* lies in removing the conflict region without disrupting the global harmonious generation. Thus, we further design a time-dependent region removing strategy that dynamically suppresses conflict regions in self-attention features.

Specifically, we define a monotonically decreasing function $r(t)$ that varies with the image's Signal-to-Noise Ratio (SNR) at each timestep, along with a corresponding region feature weighting removal strategy on arbitrary feature $\mathcal{F}$:

$$r(t) = \text{Sigmoid}(k(\frac{\text{SNR}(t_{thres})}{\text{SNR}(t)} - 1)), \quad (6)$$

$$\text{Re}(\mathcal{F}, \mathcal{M}_{con}) = \mathcal{F} \odot (1 - \mathcal{M}_{con} \odot \text{Bernoulli}(r)), \quad (7)$$

where $\text{SNR}(t) = \sqrt{\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}}$ is obtained conditioned on noise scheduling coefficient $\overline{\alpha}_t$, which increases with denoising steps, causing $r(t)$ to progressively decrease. k controls the decrease rate, and t_{thres} determines the inflection point of decreasing curve. Bernoulli(r) represents the Bernoulli mask with probability $r(t)$. Empirically, applying feature removal to both query and key features in self-attention yields optimal performance (detailed analysis in Supp.F):

$$\mathcal{A}^i = \text{Softmax}(\frac{\text{Re}(\mathcal{Q}^i, \mathcal{M}_{con}^i) \cdot \text{Re}(\mathcal{K}^i, \mathcal{M}_{con}^i)^T}{\sqrt{d}}). \quad (8)$$

The design rationale behind it is: modulating the feature removal intensity of conflict region conditioned on time-dependent image information density (weighted by SNR). The visualization in Fig.4 validates the effectiveness of this design: (1) in early timesteps, with severe conflict interference, $r(t)$ tends to 1 to perform near-complete removal on conflict regions, preventing their leakage into subsequent editing. (2) In later timesteps, as conflict object features are suppressed, $r(t)$ tends to 0 for minimal process on conflict region, ensuring adaptive and harmonious complementation of the surrounding background (*e.g.*, "dog's grassland") or obscured objects (*e.g.*, "sofa"), thereby supporting precise and scalable object editing on an exclusive layer.

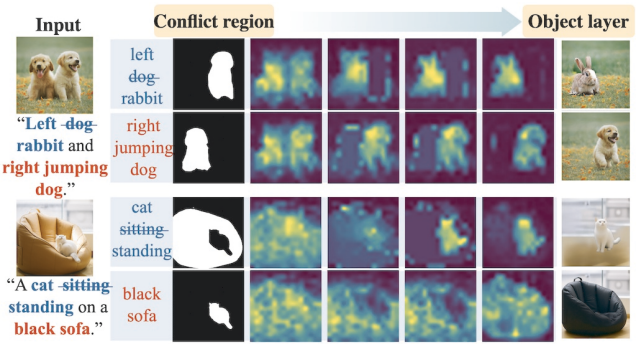

Figure 4: Visualization of attention features from earlier to later timesteps, driven by time-dependent region removing.

4.2 Object-layered Editing

To fully leverage the potential of multi-layer architecture, we further design intra-layer text guidance and cross-layer feature mapping, enabling compatible improvements for both textual semantic editing and geometric structural editing.

Textual Semantic Editing. We extend the diffusion network into an $N+2$ multi-diffusion network $\mathcal{L} = \{L_0, ..., L_{N+1}\}$, by reformulating the feature updating process (Eq.3) with object-layered text guidance:

$$\hat{\phi}^i(z_t^i, \mathcal{T}_e^i) = \mathcal{A}^i \mathcal{V}^i, \quad for\ i \in [0, N+1], \quad (9)$$

$$\mathcal{T}_e^i = \mathcal{T}_e - \bigcup_{\{j|\mathcal{M}_o^j\ in\ \mathcal{M}_{con}^i\}} (O_e^i),\ for\ i \in [1,N]. \quad (10)$$

The layer L_0 indicates the source image reconstruction layer with corresponding conflict mask $\mathcal{M}_{con}^0 = \emptyset$ and $\mathcal{T}_e^0 = \mathcal{T}_s$. The layer L_{N+1} is the canvas layer with conflict mask $\mathcal{M}_{con}^{N+1} = \bigcup_{i=1}^N (\mathcal{M}_o^i)$ and $\mathcal{T}_e^{N+1} = \emptyset$ to generate image background for fusion. Intuitively, the text guidance of each object layer is obtained by removing the object tokens located within its conflict region $\mathcal{M}_{con}^i$ from the global editing prompt $\mathcal{T}_e$, ensuring natural disentangled object editing guided by text. To simplify the subsequent discussions, we abbreviate the per-layer' spatial features $\hat{\phi}^i(z_t^i, \mathcal{T}_e^i)$ as $\hat{\phi}^i$.

Geometric Structural Editing. *LayerEdit* further supports geometric transformation, with a key insight: the multi-layer diffusion architecture inherently facilitates *multi-step*, *fine-grained* geometric feature mapping from reference object layers (L_i) to output canvas layer (L_{N+1}). Conventional diffusion frameworks often suffer from source feature overwriting during mapping, limiting them to coarse-grained single-step operation (Schouten et al. 2025) or multi-step cumulative errors (Epstein et al. 2024). *LayerEdit* overcomes these by maintaining persistent reference-to-canvas correspondence across all denoising timesteps, enabling iterative refinement of geometric transformations via multi-step

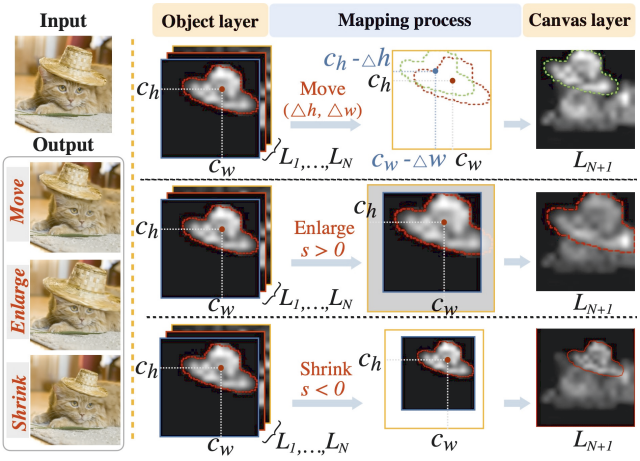

Figure 5: Diagram of *centroid-aligned mapping principle*. (c_h, c_w) is object centroid. Displacement $(\triangle h, \triangle w)$ and scale s are additional input controls for moving/resizing.

attention-based feature mapping. Our framework explores two critical geometric editing functions: object moving and resizing, following the centroid-aligned mapping principle applied to the attention output features $\hat{\phi}^i$. Fig.5 illustrates the diagram of centroid-aligned mapping principle within our multi-layer architecture via a representative multi-object case. More detailed formulas are provided in Supp.B.

All textual and geometric editing integrates seamlessly, as intra-layer textual guidance and cross-layer geometric mapping operate within disentangled operation domains.

4.3 Transparency-guided Layer Fusion

Object-layered editing enables individual objects to extend beyond their original boundaries into adjacent objects or background regions, creating potential region overlapping. To ensure coherent overlapping layer fusion, we draw inspiration from image matting techniques (Yao et al. 2024) and introduce *transparency* learning to capture more precise inter-object structure guidance. The transparency $\{\tau^i\}_{i=1}^N$ represents the contribution weights of object layers within each image patch. Object-layered fusion is performed on attention features in the canvas layer for each timestep:

$$\hat{\phi}^{N+1}_{fusion} = \sum_{i=1}^{N} \tau^i \odot \hat{\phi}^i + (1 - \sum_{i=1}^{N} \tau^i) \odot \hat{\phi}^{N+1}. \quad (11)$$

To achieve effective layer fusion, we formulate a novel constraint loss to calculate the optimal approximation of τ:

$$\mathcal{L}(\tau) = ((\hat{\phi}^{N+1}_{fusion} - \hat{\phi}^0) \odot \mathcal{M}_\tau)^2 + \sum_{i=1}^{N} (\max(0, -\tau^i))^2 + ((1 - \sum_{i=1}^{N} \tau^i) \odot \mathcal{M}_\tau)^2, \quad (12)$$

where $\mathcal{M}_\tau$ marks the overlapping region, *i.e.*, the τ of two layers are both greater than 0. Intuitively, the first term ensures the overlapping regions of fusion global image maintaining the spatial characteristics of source image, thereby guiding transparency to capture inter-object structure relationship. Physically, the second term imposes a non negativity constraint on τ, and the third term further constrains the global sum of overlapping layer region should be 1. We solve this minimization problem via gradient descent:

$$\frac{d}{d\tau^n}\mathcal{L}(\tau) = 2(\hat{\phi}^{N+1}_{fusion} - \hat{\phi}^0) \odot \mathcal{M}_\tau \odot (\hat{\phi}^n - \hat{\phi}^{N+1}) - 2\max(0, -\tau^n) - 2(1 - \sum_{i=1}^{N} \tau^i) \odot \mathcal{M}_\tau. \quad (13)$$

τ^i is initialized by object mask $\mathcal{M}^i_o$. As τ is independent transparency weights of denoising network, it quickly converges within iterative inference timesteps. This transparency learning, conditioned on source image spatial features, effectively captures inter-object structural relationships, promoting harmonious and coherent layer fusion.

Method	*editability* CLIP-T↑	*fidelity* CLIP-I↑	*fidelity* LPIPS↓	*quality* FID↓	*quality* KID↓
InfEdit$_{,LCM}$	0.2253	0.8105	0.2529	80.62	61.36
TurboEdit$_{,SDXL}$	0.2147	0.8128	0.2491	90.16	73.27
InstructCLIP$_{,IP2P}$	0.2314	0.8214	0.2307	76.27	62.40
OIR$_{,SDv1.4}$	0.2538	0.8205	0.2051	78.92	62.39
GenArtist$_{,SDXL}$	0.2372	0.8028	0.2205	88.64	66.17
ParallelEdits$_{,LCM}$	0.2457	0.8108	0.1924	74.47	57.81
LoMOE$_{,SDv2.0}$	0.2506	0.8365	0.1592	81.70	60.35
h-Edit$_{,SDv1.4}$	0.2478	0.8216	0.1815	68.38	52.28
Ours$_{,SDXL}$	0.2762	**0.8420**	**0.1358**	60.13	45.29
Ours$_{,FLUX}$	**0.2857**	0.8408	0.1372	**57.91**	**39.72**

Table 1: Quantitative comparisons with existing methods.

5 Experiments

5.1 Experimental Setups

Implementation and Datasets. As a plug-and-play framework, *LayerEdit* can be easily integrated into existing backbones. We adopt Stable Diffusion XL (Podell et al. 2023) as main backbone for a fair comparison, and FLUX (Labs 2024) to further validate generalization (detailed in Supp.C). We use DDIM sampler with 50 sampling steps and classifier-free guidance scale of 7.5. Experimentally, we obtain one of optimal settings with IoU threshold $\eta = 0.3$, rate $k = 5$, timestep threshold t_{thres} at 20 (query) and 40 (key). All experiments are conducted on two typical multi-object datasets: OIR-bench (Yang et al. 2024) and LoMOE-bench (Chakrabarty et al. 2024). Notably, the proposed multi-layer diffusion framework with parallelized layer-wise computation, incurs almost ***no additional time costs***, enabling efficient editing of more than 6-object on a single A40 GPU.

Baselines. We compare with state-of-the-art (SOTA) baselines for TMOE: h-Edit (Nguyen et al. 2025), ParallelEdits (Huang et al. 2024), LoMOE (Chakrabarty et al. 2024), GenArtist (Wang et al. 2024), and OIR (Yang et al. 2024). For comprehensive evaluation, other SOTA general editing methods (InfEdit, (Xu et al. 2024); TurboEdit, (Wu et al.

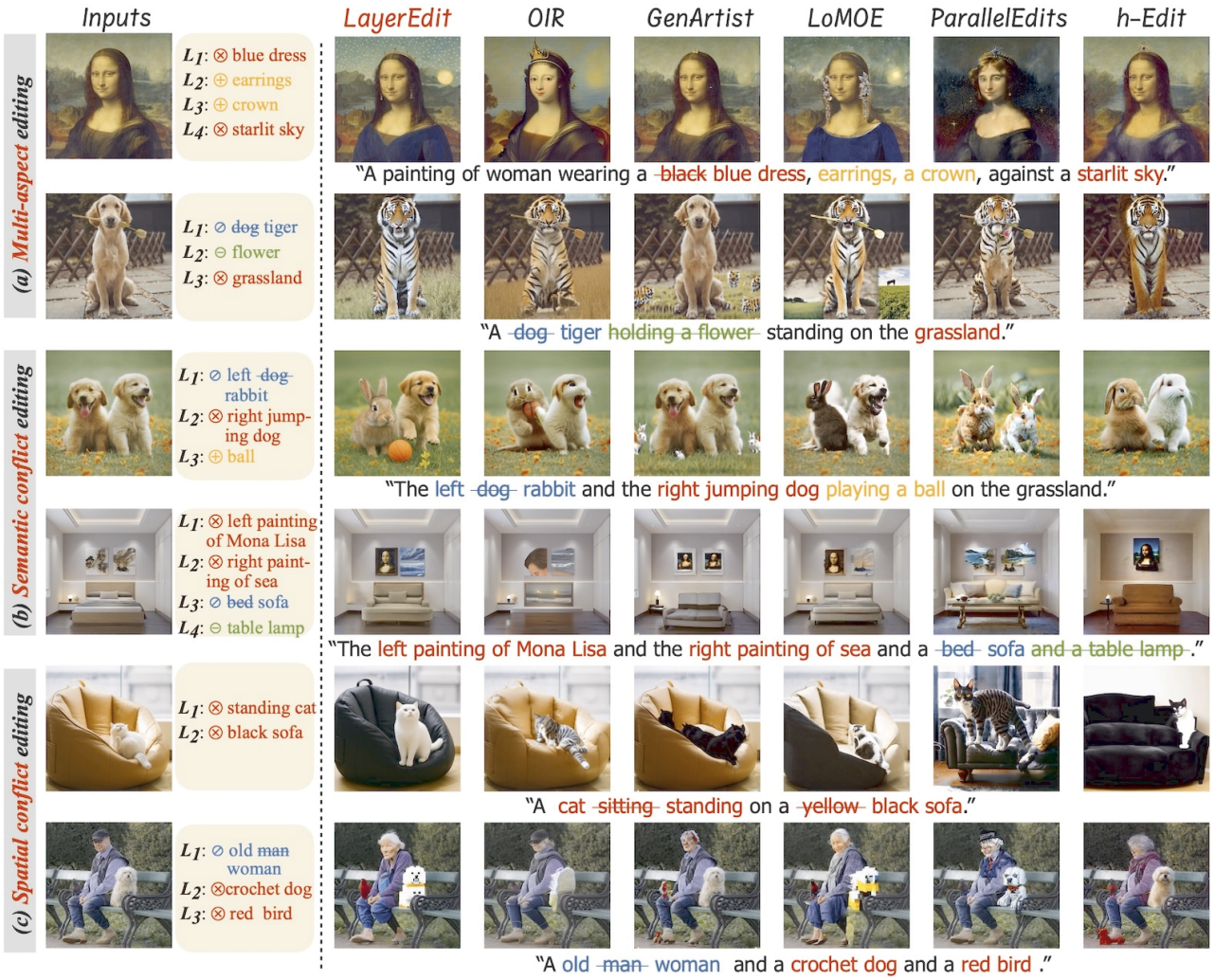

Figure 6: Qualitative comparison with SOTAs. $\{\oslash, \oplus, \ominus, \otimes\}$ denote different editing operations defined as Sec.5.2. *LayerEdit* exhibits unprecedented editability for multi-aspect/object editing. More comparisons with industrial-grade models in Supp.D.

2024); InstructCLIP, (Chen et al. 2025)) are also compared, processed as iteratively multi-turn editing for multi-object.

Evaluation Metrics. ***Image fidelity***: we evaluate with both CLIP embedding similarity (CLIP-I, (Radford et al. 2021)) and learned perceptual image patch similarity (LPIPS, (Zhang et al. 2018)). ***Editability***: we evaluate with CLIP embedding similarity between edited prompts and images (CLIP-T). ***Image quality***: we evaluate with Frechet Inception Distance (FID, (Heusel et al. 2017)) and Kernel Inception Distance (KID, (Lucic et al. 2018)).

5.2 Main Results

Quantitative Results. As illustrated in Tab.1, *LayerEdit* outperforms existing methods in all metrics, improving editability ($\mathbf{12.6}\%$ in CLIP-T), fidelity ($\mathbf{14.7}\%$ in LPIPS) and global quality ($\mathbf{15.3}\%$ in FID). These significant improvements validate the distinctive superiority of our multi-layer disentangled editing paradigm, enabling high-quality intra-object modification with seamless inter-object coherence.

Qualitative Results. Fig.6 showcases *LayerEdit*'s revolutionary capabilities: (1) ***Wide spectrum of editing***, supporting object *replacement* ($\oslash$), *addition* ($\oplus$), *removal* ($\ominus$) and *appearance/posture/background manipulations* ($\otimes$, treating background as a regular object layer). Furthermore, *LayerEdit* also effectively supports the *multi-aspect editing*, through treating objects' aspects (*e.g.*, Fig.6.a, woman's "dress" and "earrings") as independent objects. (2) ***Inter-object conflict resolution*** across complex multi-object scenarios, enabling precise text-guided editing for semantically related objects (*e.g.*, Fig.6.b, multi-directional "dog" or "painting") and spatially overlapping objects (*e.g.*, Fig.6.c, occluded "cat" or "sofa"), with seamless and harmonious global texture. This conclusion verifies the effectiveness of multi-layer disentangled framework for unconstrained intra-object editing and coherent inter-object fusion, while existing baselines suffer from either strict intra-object boundaries (*e.g.*, LoMOE's artifacts) or inter-object attention leakage (*e.g.*, ParallelEdits' mis-editing on two "dog/painting").

Figure 7: Qualitative results of geometric editing.

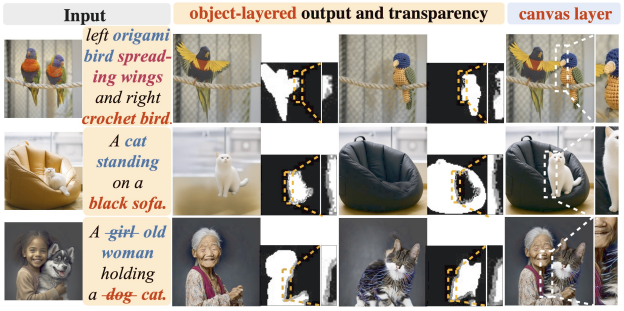

Figure 8: Visualization results of object-layered output and transparency. Boxes *highlight* the overlapping regions.

5.3 Results on Geometric Editing

As shown in Fig.7, *LayerEdit* demonstrates excellent geometric editing superiority, including: (1) ***Seamless dual-application*** of text- and geometric-guided modifications to single or multiple objects without constraints or conflicts, which originates from our multi-layer framework effectively disentangles and coordinates the operational domains of both editing functions. (2) ***Adaptive occlusion completion*** (*e.g.*, 1-st row, occluded "heart" or "bamboo"), demonstrating *LayerEdit*'s unique capability in region-unconstrained and scale-adaptive object manipulation. More qualitative results with SOTA geometric editing methods refer to Supp.E.

5.4 Visualization Results of Object Layers

Fig.8 visualizes the object-layered outputs and transparency maps. Intuitively, (1) for ***layer decomposition***, *LayerEdit* achieves precise removal of conflict regions (*e.g.*, semantically/spatially conflicting "bird"/"sofa") with adaptive region completion, thereby enabling unconstrained ***object-layered editing*** with arbitrary structural modifications and expansions (*e.g.*, "bird spreading wings" and "cat standing") within exclusive layer. (2) For ***layer fusion***, spatial features conditioned transparency learning effectively captures inter-object structure (*e.g.*, "cat on sofa", "woman hold cat"), further providing coherent overlapping region fusion for region-scalable object editing. These results verify that the ***essence of LayerEdit*** is to fundamentally model a precise decompose-editing-fusion pipeline, to achieve truly disentangled multi-object editing, free from both intra-object region constraints and inter-object conflicts.

ID	Settings	CLIP-T ↑	CLIP-I ↑
D-v1	w/o t-dependent r, w/ fixed r	0.2386	0.8103
D-v2	w/o t-dependent r, w/ linear r	0.2560	0.8194
E-v1	w/o intra-layer textual guidance	0.2489	0.8045
F-v1	w/o transparency, w/ mask	0.2415	0.8352
F-v2	w/o 1-st structure loss term	0.2536	0.8216
F-v3	w/o 2-nd non-negativity loss term	0.2508	0.8171
F-v4	w/o 3-th normalization loss term	0.2692	0.8375
– –	full model	**0.2762**	**0.8420**

Table 2: Ablation on effectiveness of different components.

5.5 Ablation Studies

We further ablate on model components to verify that our *LayerEdit* achieves one of the optimal designs for this decompose-editing-fusion architecture. We compare with: 1) D-v1/v2: model without **t**ime-dependent region removing but uses fixed (optimal at $r = 0.8$) or linear decreasing removal intensity (r from 1 to 0) on conflict region; 2) E-v1: model without intra-layer textual guidance but all use $\mathcal{T}_e$ in Eq.10. 3) F-v1: model without transparency learning but with mask-guided layer fusion; 4) F-v2/3/4: model without the $1/2/3$-th loss term in transparency learning.

As reported in Tab.2, we conclude that: (1) for decomposition, time-dependent region removing strategy excels at conflict region constraint, leading to conflict-free object-layered decomposition and editing, with a remarkable 15.8% improvement in CLIP-T. (2) For editing, the tailored intra-layer textual guidance better supports the multi-layer architecture for natural disentangled editing. (3) For fusion, model without transparency guidance severely impairs editability (CLIP-T dropped by 14.4%), proving transparency learning is essential for structure-coherent fusion in region-scalable object editing. More ablation results in qualitative comparisons, and aware/removal parameters η, k, t_{thres} in Supp.F.

6 Conclusion

In this paper, we present *LayerEdit*, a training-free multi-layer disentanglement editing framework which, for the first time, through precise object-layered decomposition and fusion, enhances text-driven multi-object editing free from inter-object conflicts and intra-object constraints. Different from existing methods fixated on intra-object localize-editing, our method shifts the focus into inter-object interactions, by explicitly identifying and constraining inter-object conflict regions to achieve *conflict-aware layer decomposition*, and through inter-object structural modeling to facilitate structure-coherent *transparency-guided layer fusion*. Moreover, the tailored *object-layered editing* module further supports the constructed multi-layer architecture, establishing novel intra-layer textual guidance and cross-layer feature mapping, to achieve comprehensive improvements for arbitrary semantic and structural editing functions. Extensive experiments demonstrate the superiority of our *LayerEdit*.

Acknowledgements

This research is supported by Artificial Intelligence National Science and Technology Major Project 2023ZD0121200, and National Natural Science Foundation of China under Grant 62222212, 623B2094 and 62336001.

References

Brack, M.; Friedrich, F.; Kornmeier, K.; Tsaban, L.; Schramowski, P.; Kersting, K.; and Passos, A. 2024. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8861–8870.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.

Chakrabarty, G.; Chandrasekar, A.; Hebbalaguppe, R.; and AP, P. 2024. Lomoe: Localized multi-object editing via multi-diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3342–3351.

Chen; X, S.; Sra, M.; and Sen, P. 2025. Instruct-CLIP: Improving Instruction-Guided Image Editing with Automated Data Refinement Using Contrastive Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28513–28522.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A.; and Holynski, A. 2024. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36.

Feng, Y.; Gong, B.; Chen, D.; Shen, Y.; Liu, Y.; and Zhou, J. 2024. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4744–4753.

Fu, F.; Zhang, L.; Huang, M.; and Mao, Z. 2025. FeedEdit: Text-Based Image Editing with Dynamic Feedback Regulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2661–2670.

Guo, Q.; and Lin, T. 2024. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6986–6996.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-prompt image editing with cross attention control. *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Huang, M.; Cai, J.; Jia, S.; Lokhande, V. S.; and Lyu, S. 2024. ParallelEdits: Efficient Multi-Aspect Text-Driven Image Editing with Attention Grouping. In *Advances in Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada.

Jain, J.; Li, J.; Chiu, M. T.; Hassani, A.; Orlov, N.; and Shi, H. 2023. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2989–2998.

Jia, Y.; Yuan, Y.; Cheng, A.; Wang, C.; Li, J.; Jia, H.; and Zhang, S. 2024. DesignEdit: Multi-Layered Latent Decomposition and Fusion for Unified & Accurate Image Editing. *arXiv preprint arXiv:2403.14487*.

Joseph, K.; Udhayanan, P.; Shukla, T.; Agarwal, A.; Karanam, S.; Goswami, K.; and Srinivasan, B. V. 2024. Iterative multi-granular image editing using diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8107–8116.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Li, P.; Huang, Q.; Ding, Y.; and Li, Z. 2023. Layerdiffusion: Layered controlled image editing with diffusion models. In *SIGGRAPH Asia 2023 Technical Communications*, 1–4.

Li, Y.; Chan, K.; Sun, Y.; Lam, C.; Tong, T.; Yu, Z.; Fu, K.; Liu, X.; and Tan, T. 2025. MoEdit: On Learning Quantity Perception for Multi-object Image Editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2683–2693.

Lucic, M.; Kurach, K.; Michalski, M.; Gelly, S.; and Bousquet, O. 2018. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31.

Nguyen, T.; Do, K.; Kieu, D.; and Nguyen, T. 2025. h-Edit: Effective and Flexible Diffusion-Based Editing via Doob's h-Transform. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28490–28501.

Patashnik, O.; Garibi, D.; Azuri, I.; Averbuch-Elor, H.; and Cohen-Or, D. 2023. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23051–23061.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schouten, M.; Kaya, M. O.; Belongie, S.; and Papadopoulos, D. P. 2025. POEM: Precise Object-level Editing via MLLM control. *arXiv preprint arXiv:2504.08111*.

Song, X.; Cui, J.; Zhang, H.; Chen, J.; Hong, R.; and Jiang, Y.-G. 2024. Doubly Abductive Counterfactual Inference for Text-based Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9162–9171.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. 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*, 1921–1930.

Wang, Z.; Li, A.; Li, Z.; and Liu, X. 2024. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*.

Wei, C.; Xiong, Z.; Ren, W.; Du, X.; Zhang, G.; and Chen, W. 2025. Omniedit: Building image editing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations*.

Wu, Z.; Kolkin, N.; Brandt, J.; Zhang, R.; and Shechtman, E. 2024. TurboEdit: Instant text-based image editing. *arXiv preprint arXiv:2408.08332*.

Xu, S.; Huang, Y.; Pan, J.; Ma, Z.; and Chai, J. 2024. Inversion-free image editing with natural language. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yang, Z.; Ding, G.; Wang, W.; Chen, H.; Zhuang, B.; and Shen, C. 2024. Object-aware inversion and reassembly for image editing. *The International Conference on Learning Representations*.

Yao, J.; Wang, X.; Yang, S.; and Wang, B. 2024. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 103: 102091.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhou, J.; Li, J.; Xu, Z.; Li, H.; Cheng, Y.; Hong, F.-T.; Lin, Q.; Lu, Q.; and Liang, X. 2025. Fireedit: Fine-grained instruction-based image editing via region-aware vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13093–13103.

Zhu, H.; Liu, H.; Fu, B.; and Wang, Y. 2025. MDE-Edit: Masked Dual-Editing for Multi-Object Image Editing via Diffusion Models. *arXiv preprint arXiv:2505.05101*.