DesignEdit: Unify Spatial-Aware Image Editing via Training-free Inpainting with a Multi-Layered Latent Diffusion Framework

Yueru Jia[1], Aosong Cheng[1], Yuhui Yuan[2*], Chuke Wang[1], Ji Li[3], Huizhu Jia[1], Shanghang Zhang[1*]

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University;
[2]Microsoft Research Asia;
[3]Microsoft
jiayueru@stu.pku.edu.cn, yuyua@microsoft.com, shanghang@pku.edu.cn

Abstract

Spatial-aware image editing focuses on modifying the position and size of elements within a given image. However, previous works still struggle with maintaining background harmony in the original editing areas, as well as preserving the initial identity of the edited elements, making it difficult to achieve complex multi-object editing in a single pass. In this paper, we aim to perform flexible spatial editing in a simple yet straightforward manner. We propose to inpaint the background first and develop a two-stage multi-layered latent diffusion framework to edit each element independently. Specifically, we design a key-masking self-attention scheme alongside artifact suppression to achieve background inpainting within the denoising process, leveraging the powerful generative capabilities of the Latent Diffusion Model, Stable Diffusion XL-1.0. The latent decomposition and fusion framework is capable of unifying various spatial-aware operations, including removal, resizing, relocation, flipping, addition, camera panning, zooming out, occlusion-aware editing, and cross-image editing. Experiments demonstrate the superior inpainting quality for object removal, along with enhanced versatility and higher precision in spatial-aware editing achieved by our method.

Project Page — https://design-edit.github.io/
Code — https://github.com/design-edit/DesignEdit

Introduction

Spatial-aware image editing helps users adjust the position and size of elements in an image. For instance, Fig. 1 shows a "Three Little Pigs" storybook design generated by DALLE-3 (OpenAI 2023), the overall style and details are quite refined, but on closer inspection, there are four pigs instead of three, which means one needs to be removed. Spatial-aware editing allows us to remove or add pigs, resize and move houses, erase text and clouds, pan, zoom out, and even redesign the original image.

This type of complex spatial-aware editing faces two main challenges: how to fill in the background after something is removed and how to edit multiple objects while keeping their original appearance consistent. Latent diffusion models (LDMs) (Avrahami, Fried, and Lischinski 2023; Saharia et al. 2022) like Stable Diffusion (Rombach et al. 2022; Podell et al. 2023) have shown strong capabilities in image generation, and various inpainting models (Cao et al. 2024; Singh et al. 2023) based on LDMs have demonstrated impressive results. They fill regions based on text instructions, but focus on editing specific areas rather than spatial layout. Previous spatial-aware image editing methods use LDMs through classifier guidance, updating latent features during each denoising step, where the loss function constrains the position and size of elements to match the target. However, these methods often overlook background filling and struggle with removing large objects. Additionally, the soft constraints on the latent space can lead to a loss of detail in the objects, and these methods heavily depend on the type of operation and the objects being edited, limiting their ability to handle multiple objects with complex operations.

Our method is inspired by the *layer* concept in the design field, where each layer represents elements that can be manipulated independently. Instead of gradually updating the latent representation with loss guidance, our approach integrates removal-specific inpainting into the diffusion process and directly utilizes the flexible latent representation in LDMs to address the two main challenges mentioned above.

First, we propose a key-masking self-attention scheme applied during the initial K denoising steps. This training-free, straightforward method allows the Query to ignore masked regions and focus on the surrounding areas in the Key. To improve removal results, we introduce artifact suppression by adding an additional refinement mask to the Key, which helps reduce attention to areas that may cause artifacts. Additionally, we design a multi-layered framework with two sub-processes: latent decomposition and fusion, enabling flexible editing of multiple objects in a single denoising process. The original image is first broken down into separate latent layers with background inpainting, and then the layers are fused at a specific timestep $T - K$, following layer-wise instructions or a target layout, which can be generated by GPT-4V. Finally, we unify various spatial-aware image editing tasks within the multi-layered framework by designing task-specific masks for inpainting and assigning different latents for removal.

We conduct extensive qualitative and quantitative experi-

*corresponding author

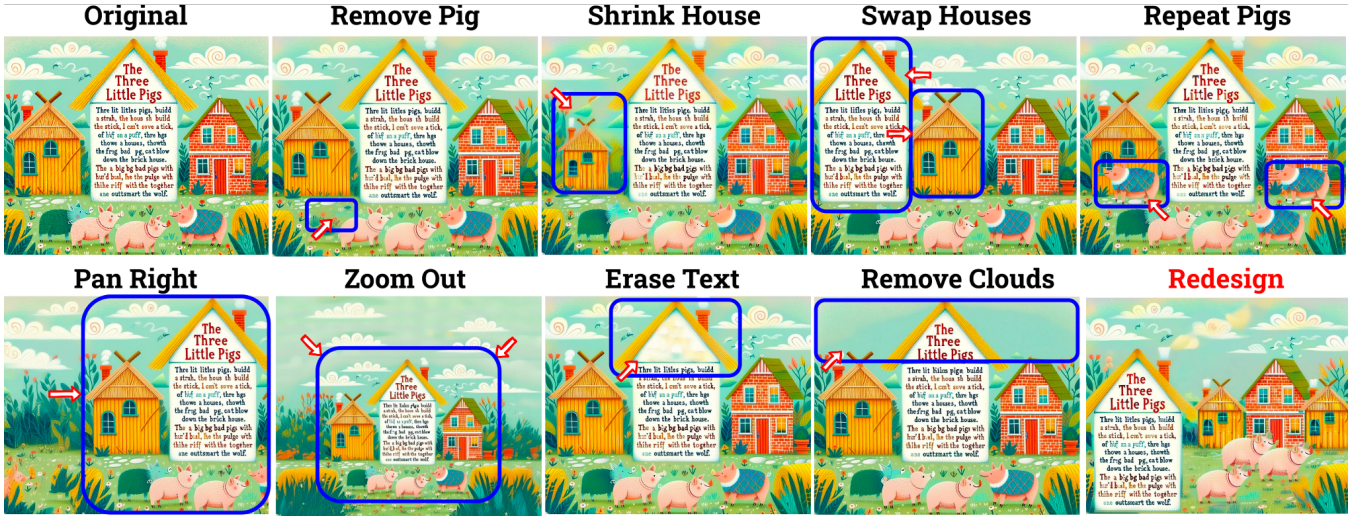


Figure 1: **Examples of our spatial-aware image editing.** Our approach facilitates a range of image editing operations through training-free background inpainting, using a unified multi-layered framework to achieve various spatial-aware edits.

ments. We compare our method with four existing inpainting methods on MagicBrush dataset (Zhang et al. 2023) for removal tasks, demonstrating comparable inpainting ability to other methods that are not training-free. Additionally, we present single-object editing results in comparison with spatial editing method, Self-Guidance (Epstein et al. 2023) and DiffEditor (Mou et al. 2024), and showcase further applications, particularly in design images where objects contain detailed features.

To summarize, we present the following contributions:

- We propose a key-masking self-attention scheme with artifact suppression refinement for training-free removal inpainting.
- We develop a multi-layered framework with two subprocesses, latent decomposition and fusion, for flexible object editing in a single diffusion denoising process.
- We unify spatial-aware editing through the multi-layered framework with inpainting to achieve complex multi-object editing, including basic operations like removal, resizing, relocation, flipping, and addition, as well as extended operations like camera panning, zooming out, occlusion-aware editing, and cross-image composition.

Related Work

Image Inpainting

Image inpainting aims to complete the missing elements in an image (Huang et al. 2024). LaMa (Suvorov et al. 2022), based on fast Fourier convolutions, is trained on real-world dataset to inpaint large masks. Recent diffusion-based inpainting models (Lugmayr et al. 2022; Ju et al. 2024; Zhang et al. 2024) like SDXL-inpainting and ControlNet-inpainting (Zhang, Rao, and Agrawala 2023a) introduce additional guidance to incorporate new context into the image by fine-tuning T2I diffusion models. Uni-paint (Yang, Chen, and Liao 2023) modifies attention with masks, to achieve multi-modal control but requires extra fine-tuning. These diffusion-based inpainting models tend to fill in new elements in their original positions, while our method aims for training-free inpainting, particularly for removal tasks.

Spatial-aware Image Editing

Spatial-aware editing modifies images by considering spatial context and relationships, including removing, moving, resizing, or adding elements. Previous methods have been inspired by classifier guidance strategies in diffusion models (Chen, Laina, and Vedaldi 2023; Chefer et al. 2023; Yu et al. 2023; Zhao et al. 2022). Self-guidance (Epstein et al. 2023) extracts object geometric information from cross-attention maps and computes loss with target positions and sizes to iteratively update latent features. DragDiffusion (Shi et al. 2024) and DiffEditor (Mou et al. 2024) incorporates dragging-based image editing tasks into diffusion models, expanding to more spatial-aware tasks like object movement and resizing using image prompts such as object masks. However, these guidance-driven methods neglect background inpainting and struggle with large object removal or significant changes. Additionally, the soft loss constraints make it challenging to maintain object identity, especially during resizing. Different operations can interfere with each other, complicating flexible editing. Our method aims for a more straightforward approach: we first perform background inpainting using a key-masking self-attention scheme, and then use a multi-layered framework to obtain separate latents for various flexible spatial-aware operations.

Attention Modification in Diffusion-based Editing

Cross-attention and self-attention modules in diffusion models have been shown to have a strong relationship with the layout and geometric information of an image (Hertz et al. 2022) . Recent diffusion-based image editing works (Levin and Fried 2024; Mirzaei et al. 2023; Ren et al. 2024) have

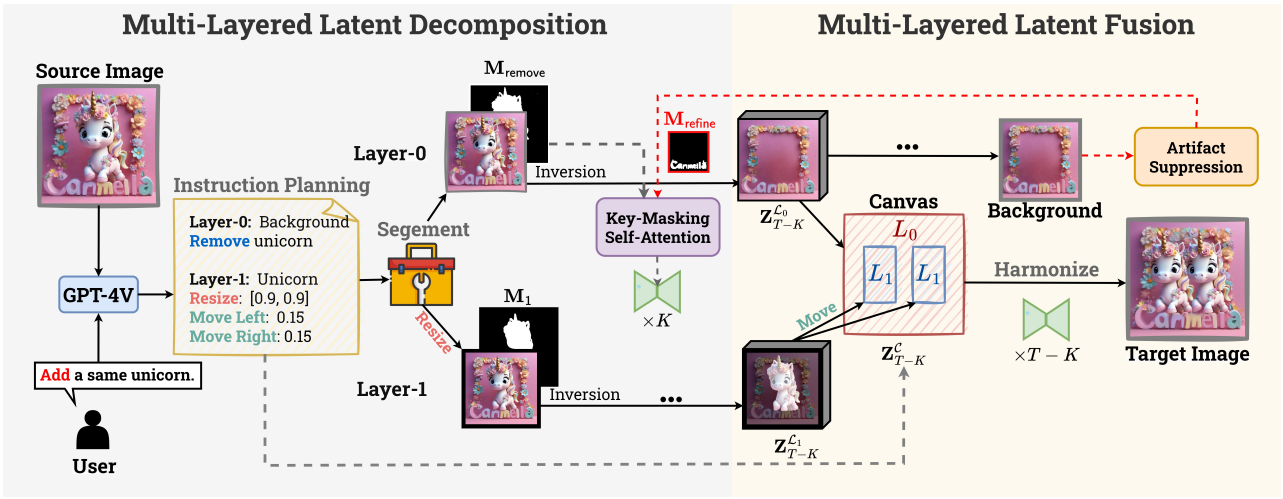


Figure 2: **Illustrating the multi-layered framework.** During the latent decomposition stage, GPT-4V is used to perform layer-wise instruction planning based on a vague input prompt. Key-masking self-attention scheme is applied to inpaint the background layer $\mathbf{Z}_t^{\mathcal{L}_0}$. In the latent fusion stage, the model follows layer-wise instructions sequentially to paste them onto the canvas latent $\mathbf{Z}_t^{\mathcal{C}}$.

achieved various types of editing effects through different uses of the attention mechanism. Prompt-to-Prompt (Hertz et al. 2022) exchanges cross-attention maps to achieve object replacement or editing, MasaCtrl (Cao et al. 2023) designs a mutual self-attention mechanism to achieve language-guided object pose changes, Style-Align (Hertz et al. 2024) maintains style consistency between generated images by designing a shared attention mechanism. These methods focus on *in-place* editing, which contrasts with spatial-aware editing. We modifies attention by examining the distinct roles of the query, key, and value in self-attention, and develops specialized techniques for removal inpainting, which provides a foundation for subsequent spatial editing.

Method

We propose our multi-layered framework in Fig.2, where a key-masking self-attention scheme is applied in the first K diffusion steps to achieve background inpainting during latent decomposition, followed by layer-wise latent fusion at $T-K$ with instruction-guided layout. Additionally, we will detail how this framework can be extended to various types of spatial-aware editing.

Key-Masking Self-Attention Scheme

Inspired by the role of self-attention in facilitating information exchange (Wu et al. 2023), we leverage the distinct roles of Query, Key, and Value to enable background inpainting. Query determines attention focus, Key matches relevant information, and Value holds content. To achieve removal, we suppress the Key, allowing the Query to *ignore* removal regions and gather surrounding information for inpainting. This key-masking self-attention scheme applies the removal mask $\mathbf{M}_{\text{remove}}$ to Key features during the first K steps, as shown in Fig. 3:

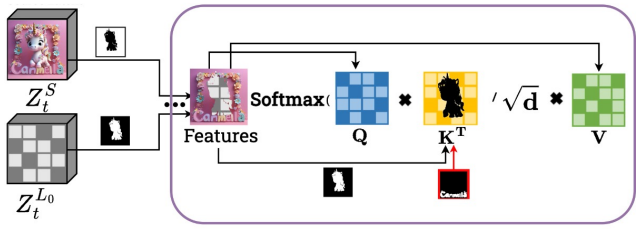


Figure 3: Illustrating the key-masking self-attention scheme at denoising time step **t**. The mask with a red border is denoted as $\mathbf{M}_{\text{refine}}$.

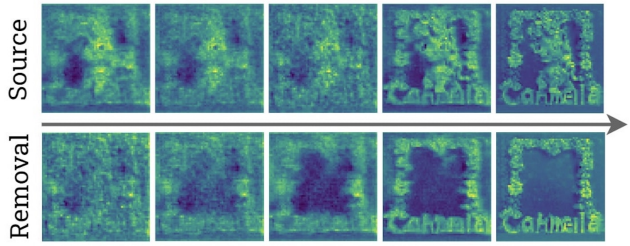


Figure 4: Visualization of the heatmap of self-attention output features. Denoising steps increase along the arrow.

$$\text{KMSAttn} := \text{Softmax}\left(\frac{\mathbf{Q}\left((1-\mathbf{M}_{\text{remove}})\odot\mathbf{K}\right)^{\text{T}}}{\sqrt{d}}\right)\mathbf{V}, \quad (1)$$

where Q, K, V come from the background latent features $\mathbf{Z}_T^{\mathcal{L}_0}$, projected by W_Q, W_K, W_V.

This scheme requires guidance from the original image's latent features $\mathbf{Z}_t^{\mathcal{S}}$ to ensure consistency between the fea-

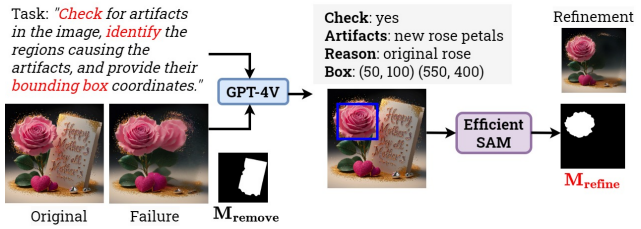


Figure 5: Illustrating the $\mathbf{M}_{\mathbf{refine}}$ generation with GPT-4V resoning in artifact suppression process.

tures outside the mask and the original image. At each timestep t, the features from the source latent, outside the masked region, are fused into the current latent $\mathbf{Z}_t^{\mathcal{L}_0}$, where $\mathbf{Z}_T^{\mathcal{L}_0} = \mathbf{Z}_T^{\mathcal{S}}$:

$$\mathbf{Z}_t^{\mathcal{L}_0} = \mathbf{Z}_t^{\mathcal{L}_0} \odot \mathbf{M}_{\text{remove}} + \mathbf{Z}_t^{\mathcal{S}} \odot (1 - \mathbf{M}_{\text{remove}}). \quad (2)$$

Fig.4 visualizes the heatmaps of the output features from the self-attention applied to the source latent $\mathbf{Z}_t^{\mathcal{S}}$ and the removal latent $\mathbf{Z}_t^{\mathcal{L}_0}$. We observe that information corresponding to the masked region $\mathbf{M}_{\text{remove}}$ is suppressed in the final output, receiving a lower attention score than $\mathbf{Z}_t^{\mathcal{S}}$, while maintaining a smooth transition with the surrounding background during the denoising process.

Artifact Suppression Inpainting results may be unsatisfactory due to artifacts, such as rose petals in Fig. 5. To address this, we introduce artifact suppression using $\mathbf{M}_{\mathbf{refine}}$, which guides the model away from artifact-prone regions. As shown in Fig. 5, GPT-4V identifies artifacts, reasons about their causes (e.g., the original rose), and generates segmentation bounding boxes. $\mathbf{M}_{\mathbf{refine}}$ is then applied alongside $\mathbf{M}_{\mathbf{remove}}$ in the key-masking self-attention scheme:

$$\text{Softmax}\left(\frac{\mathbf{Q}\left((1 - (\mathbf{M}_{\text{remove}} \vee \mathbf{M}_{\text{refine}})) \odot \mathbf{K}\right)^{\mathrm{T}}}{\sqrt{d}}\right)\mathbf{V}. \quad (3)$$

Multi-layered framework

Inspired by the concept of layers in design, we introduce a multi-layered latent decomposition and fusion framework to simplify complex editing. This framework decomposes image editing into a series of independent, manageable layer-wise operations for each element. Here, a *layer* refers to a singular basic visual element in the source image.

Latent Decomposition This stage acquires the latent representation $\mathbf{Z}_t^{\mathcal{L}_i}$ for each layer (Mokady et al. 2023; Miyake et al. 2024; Han et al. 2023). For vague instructions like "Add the same unicorn" in Fig. 2, GPT-4V generates layer-wise editing instructions and layer order, supporting "Resize" and "Move". Resizing in latent space may cause blurring and artifacts, which we address by encoding resized elements while maintaining their central positioning. Separate layer-wise latent features are obtained using key-masking self-attention in the initial K steps.

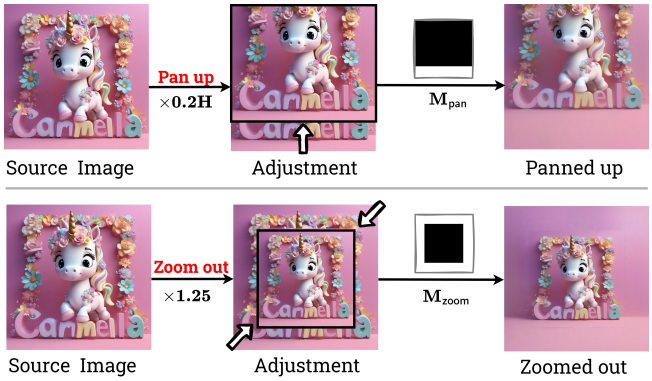


Figure 6: Illustration of the mask usage in camera panning and zooming out tasks.

Latent Fusion After the first K steps of removal on the background layer L_0, we sequentially paste the prepared layered latent features onto the layout canvas latent $\mathbf{Z}_t^{\mathcal{C}}$ at timestep $T - K$ with layer-wise "Move" instructions V_i. At timestep $t = T - K$, first initialize layout canvas latent $\mathbf{Z}_t^{\mathcal{C}}$ with $\mathbf{Z}_t^{\mathcal{L}_0}$, and then for each Layer $L_i, i = 1, 2, ..., N$ and for each operating vector $\mathbf{v_j} \in V_i$, we denote $\hat{\mathbf{M}}_i = \text{Move}(\mathbf{M_i}; \mathbf{v_j})$, and the latent fusion process is described by the following equation:

$$\mathbf{Z}_t^{\mathcal{C}} = \mathbf{Z}_t^{\mathcal{C}} \odot (1 - \hat{\mathbf{M}}_i) + \text{Move}(\mathbf{Z}_t^{\mathcal{L}_i}; \mathbf{v_j}) \odot \hat{\mathbf{M}}_i. \quad (4)$$

Unifying Spatial-aware Image Editing Tasks

With the multi-layered framework, we can unify various basic spatial-aware editing operations in Tab.1 by designing task-specific masks, source, removal and target latents.

Object Removal, Movement, Resizing and Flipping These are basic editing operations. Resizing and flipping require pixel-level adjustments to the source image before encoding to avoid losing details. Movement is executed during the fusion stage. The $\mathbf{M}_{\text{remove}}$ is the union of masks for all objects needing manipulation, denoted as $\sum \mathbf{M}_{\text{obj}}$.

Camera Panning and Zooming Out Camera panning and zooming out are converted into removal tasks by adjusting the initial image and generating two masks, as shown in Fig. 6. The source image is panned or zoomed, pasted onto the original, and removal regions are initialized with adjacent areas to ensure smooth transitions. Regions corresponding to the original image are set to 0, and areas requiring completion are set to 1. During the $T \sim T - K$ latent decomposition, $\mathbf{M}_{\text{remove}}$ is replaced with $\mathbf{M}_{\text{zoom/pan}}$.

Occlusion-Aware Object Editing For objects partially occluded in the source image, direct editing with initial masks may cause loss of object details, as shown in Fig. 7, where a dog's leg is obscured by a ball, resulting in incomplete relocation. To address this, we propose the *Integrated Decomposition-Fusion Technique*, leveraging the inpainting capability of our key-masking self-attention scheme. As illustrated in Fig. 7, during each of the first K diffusion steps,

Editing Task	Adjust Image	Remove Mask $\mathbf{M}_{\text{remove}}$	Source, Removal, Target	Fusion Step t
Object Removal	-	$\sum \mathbf{M}_{\text{obj}}$	$\mathbf{Z}_t^{\mathcal{S}}, \mathbf{Z}_t^{\mathcal{L}_0}, \mathbf{Z}_t^{\mathcal{L}_0}$	-
Object Movement	-	$\sum \mathbf{M}_{\text{obj}}$	$\mathbf{Z}_t^{\mathcal{S}}, \mathbf{Z}_t^{\mathcal{L}_0}, \mathbf{Z}_t^{\mathcal{C}}$	$T-K$
Object Resizing, Flipping	Resize, Flip	$\sum \mathbf{M}_{\text{obj}}$	$\mathbf{Z}_t^{\mathcal{S}}, \mathbf{Z}_t^{\mathcal{L}_0}, \mathbf{Z}_t^{\mathcal{C}}$	$T-K$
Camera Panning	Pan and Paste	$\mathbf{M}_{\text{pan}}$	$\mathbf{Z}_t^{\mathcal{S}}, \mathbf{Z}_t^{\mathcal{L}_0}, \mathbf{Z}_t^{\mathcal{L}_0}$	-
Zooming Out	Zoom and Paste	$\mathbf{M}_{\text{zoom}}$	$\mathbf{Z}_t^{\mathcal{S}}, \mathbf{Z}_t^{\mathcal{L}_0}, \mathbf{Z}_t^{\mathcal{L}_0}$	-
Occlusion-Aware Editing	-	$\sum_{v_j \in V_i} \text{Move}(\mathbf{M}_{\text{occlude}}; \mathbf{v_j}))$	$\mathbf{Z}_t^{\mathcal{C}}, \hat{\mathbf{Z}}_t^{\mathcal{C}}, \hat{\mathbf{Z}}_t^{\mathcal{C}}$	$T-K \sim 0$
Cross-Image Composition	Layout-guided	$\mathbf{M}_{\text{BG}}$	$\mathbf{Z}_t^{\mathcal{BG}}, \mathbf{Z}_t^{\mathcal{L}_0}, \mathbf{Z}_t^{\mathcal{C}}$	$T-K$

Table 1: **Unified Overview of Spatial-Aware Image Editing Tasks.** "Source" represents the initial latent before removal, as defined in Eq.(1) and (2). "Removal" refers to the latent to apply key-masking self-attention in Eq.(1) and (2). "Target" latent is used to decode the final output. "Fusion Step t" is the range where Eq.(4) is implemented.

Methods	L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑	FID↓	LPIPS↓	Train or Finetune	Base Model
LaMa	0.014	0.004	0.985	0.979	0.307	25.077	0.053	Train	-
ControlNet-inpainting	**0.025**	0.010	0.951	0.905	0.300	59.322	0.095	Finetune	SD-v1.5
SDXL-inpainting	0.030	**0.005**	0.968	0.959	0.303	44.621	**0.069**	Finetune	SDXL-1.0
Uni-paint	0.064	0.020	0.920	0.837	0.298	103.846	0.156	Finetune	SD-v1.4
Ours	0.028	0.007	**0.969**	**0.971**	**0.306**	**38.883**	0.070	×	SDXL-1.0

Table 2: **Quantitative study on the MagicBrush test set for the mask-guided object removal task.** Underline and **Bold** represent the top-2 results. Our method is the only one that does not require training or finetuning, and it achieves results comparable to SDXL-inpainting across 7 metrics in 51 examples. Other methods are specifically trained or fine-tuned for mask-guided image inpainting.

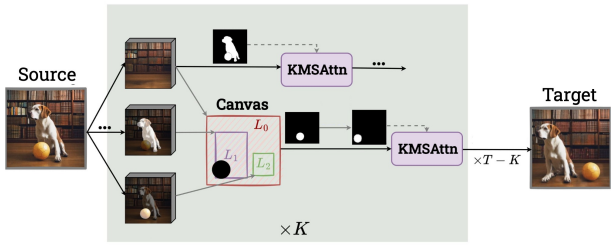


Figure 7: Illustrating the Integrated Decomposition-Fusion Technique in occlusion-aware object editing at timestep t.

background inpainting is performed on $\mathbf{Z}_t^{\mathcal{L}_0}$, followed by a fusion operation on the initial canvas latent $\mathbf{Z}_t^{\mathcal{C}}$ using Eq. (4). A new mask $\mathbf{M}_{\text{occlude}}$, such as the ball mask, is introduced, enabling inpainting on the updated canvas latent $\hat{\mathbf{Z}}_t^{\mathcal{C}}$, guided by the original latent $\mathbf{Z}_t^{\mathcal{C}}$:

$$\hat{\mathbf{Z}}_t^{\mathcal{C}} = \hat{\mathbf{Z}}_t^{\mathcal{C}} \odot \hat{\mathbf{M}}_{\text{occlude}} + \mathbf{Z}_t^{\mathcal{C}} \odot (1 - \hat{\mathbf{M}}_{\text{occlude}}). \tag{5}$$

We denote $\hat{\mathbf{M}}_{\text{occlude}} = \sum_{v_j \in V_i} \text{Move}(\mathbf{M}_{\text{occlude}}; \mathbf{v_j})$ to represent the sum of masks after moving with the occluded layer L_i. The key-masking self-attention scheme treats $\hat{\mathbf{M}}_{\text{occlude}}$ as $\mathbf{M}_{\text{remove}}$ in Eq.(1) for the new canvas latent $\hat{\mathbf{Z}}_t^{\mathcal{C}}$. The Integrated Decomposition-Fusion Technique is a more general fusion strategy, and in non-occluded image editing contexts, it is equivalent to the one-step fusion at $t = T - K$, with lower computational cost.

Cross-Image Composition Our multi-layered framework is naturally extended to support cross-image composition by encoding a background reference image $\mathbf{Z}_t^{\mathcal{BG}}$ that needs inpainting, along with a set of foreground images containing the target elements. GPT-4V can be used to design new layouts and determine the layer sequences.

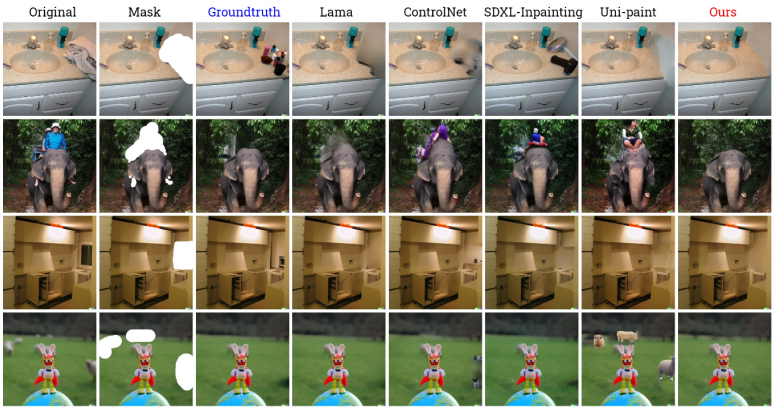
Experiments

Implementation Details

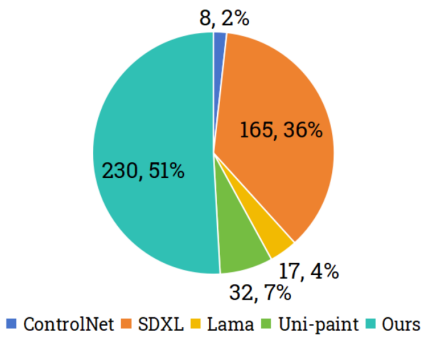
We made structural modifications to SDXL-1.0 (Podell et al. 2023) using the frozen weights and generated images at a resolution of 1024×1024. As a latent diffusion model, the resolution of SDXL-1.0's latent space is 128×128. We utilized a 50-step DDIM (Song, Meng, and Ermon 2022) denoising procedure, which means $T = 50$. We selected the most effective value for K, which is $K = 40$. The key-masking self-attention is applied across the 70 self-attention blocks in SDXL-1.0, with a range of $[50 \sim 10]$.

Comparison to State-of-the-art

Object Removal We compare the removal-specific inpainting ability of our methods with other 4 inpainting methods: LaMa (Suvorov et al. 2022), ControlNet-inpainting (Zhang, Rao, and Agrawala 2023b), SDXL-inpainting, and Uni-paint (Yang, Chen, and Liao 2023) on the MagicBrush benchmark (Zhang et al. 2023). In our experiments, we utilize data with instructions that are only about removal, with a total of 51 examples, each including a ground-truth image for evaluation.



(a) Qualitative comparison on MagicBrush dataset.



(b) The vote count and percentage in user study

Figure 8: Removal ability comparison with other inpainting models.

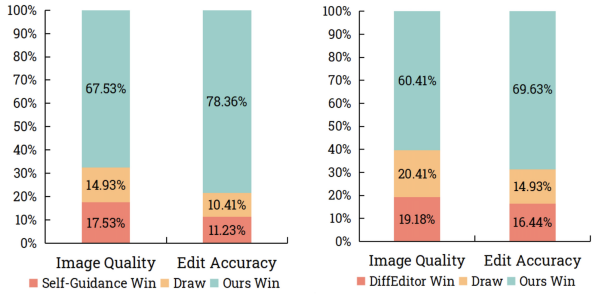


Figure 9: Win rates of spatial-aware editing compared with DiffEditor and Self-Guidance in terms of image quality and edit accuracy.

Figure 10: Qualitative comparison between our method against Self-Guidance and DiffEditor.

We evaluated inpainting performance on 7 metrics from the MagicBrush benchmark, as shown in Tab. 2. While LaMa, specifically trained for inpainting, performed best, our method achieved comparable results to SDXL-inpainting without fine-tuning. A user study with 452 votes from 113 participants (Fig. 8 (b)) showed our method had a 51% preference rate, outperforming LaMa, which produced blurring artifacts in large-area removal (Fig. 8 (a)).

Object Spatial-aware Editing We compared our spatial editing abilities—resizing and moving—with Self-Guidance (Epstein et al. 2023) and DiffEditor (Mou et al. 2024) on single-object editing. As shown in Fig. 9, a user study with 73 participants and 1,460 votes across 10 examples evaluated image quality and edit accuracy. Fig. 10 highlights our method's superior inpainting performance and object consistency.

Ablation Study

Effect Range of Key-Masking Self-Attention By implementing Eq. (2) across the entire range $[50 \sim 0]$ to maintain the surrounding features consistent with the source image, we investigate which effect range results in the most effective removal. As illustrated in the second row of Fig.12, significant removal can be achieved within the first 10 steps of key-masking self-attention scheme, while the range $[50 \sim 10]$ more effectively integrates the edges and blends better with the background. Therefore, we use $K = 40$ as our optimal setting across all editing tasks.

Mask Positioning within Self-Attention We compared different mask positions for removal, as shown in Fig.12. Masking the Query blurs the removal area by reducing the importance of the masked region during attention calculation, thus lowering pixel clarity. Masking the Value distorts the image by assigning incorrect pixel values. In contrast, Masking the Key ensures a smooth transition and a coherent integration with the surrounding information.

Multi-Object Complex Editing Results

Fig.11 shows additional multi-object editing results with complex operations such as removal, swapping, relocation, resizing, addition, flipping, and cross-image composition. All results are generated in one round.

Failure Analysis for GPT-4V planning GPT-4V assigns layer sequences and generates instructions or layouts, but its planning can fail in complex cross-image editing. Fig. 13 (a) shows GPT-4V misestimating object size and position, causing the cup to obscure the dog. Fig. 13 (b) illustrates an incorrect layer order, placing the crown beneath the cat. Subsequent attempts can be made to use GPT-4V for correcting and adjusting the generated images in a closed loop.

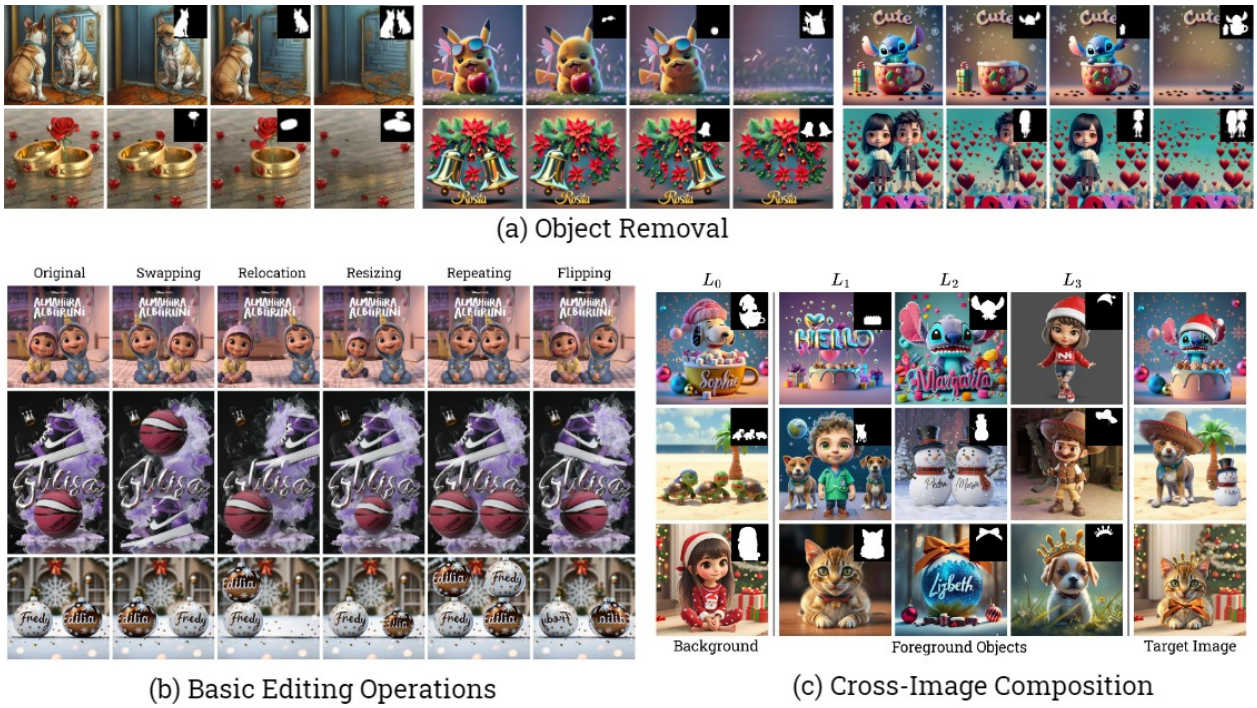


Figure 11: **Qualitative results of multi-object complex editing.** (a) shows removal results for small to large objects, (b) presents flexible editing for two-object manipulation, (c) shows cross-image composition with layers ordered from left to right.

Figure 12: Image removal quality comparison under different masking strategies and effect ranges.

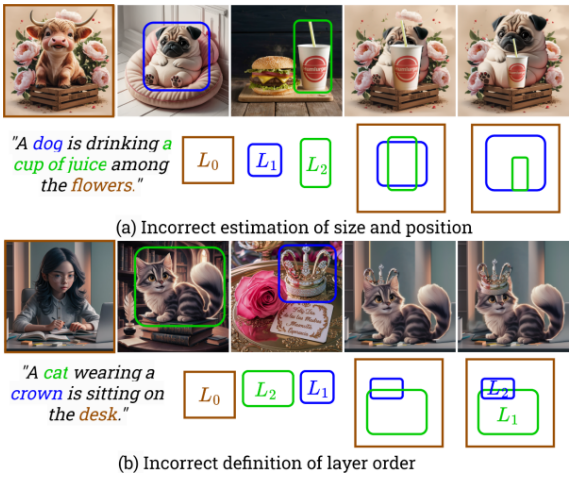


Figure 13: Two failure case studies of GPT-4V's layer-wise instruction planning.

Conclusion

In this study, we focus on spatial-aware image editing using the latent diffusion model, Stable Diffusion XL-1.0. We propose a key-masking self-attention scheme integrated into the diffusion process to enable training-free background inpainting. We also introduce an artifact suppression process to further enhance removal-specific inpainting results. Additionally, we develop a multi-layered latent decomposition and fusion framework that leverages the planning capabilities of GPT-4V to generate layer-wise editing instructions, achieving flexible editing while ensuring the identity consistency of existing elements. Within this multi-layered framework, we unify various spatial editing tasks, including removal, resizing, moving, camera panning, zooming out, occlusion-aware object editing, and cross-image composition. Notably, our training-free background inpainting achieves removal results comparable to diffusion-based inpainting techniques that require fine-tuning, and our spatial editing quality surpass state-of-the-art editing methods. We showcase numerous results across a range of complex, multi-object, spatial-aware image editing tasks.

Acknowledgements

This work was supported by the National Science and Technology Major Project (No. 2022ZD0117800).

References

Avrahami, O.; Fried, O.; and Lischinski, D. 2023. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4): 1–11.

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 (ICCV)*, 22560–22570.

Cao, P.; Zhou, F.; Song, Q.; and Yang, L. 2024. Controllable Generation with Text-to-Image Diffusion Models: A Survey. arXiv:2403.04279.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. arXiv:2301.13826.

Chen, M.; Laina, I.; and Vedaldi, A. 2023. Training-Free Layout Control with Cross-Attention Guidance. arXiv:2304.03373.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A. A.; and Holynski, A. 2023. Diffusion Self-Guidance for Controllable Image Generation. arXiv:2306.00986.

Han, L.; Wen, S.; Chen, Q.; Zhang, Z.; Song, K.; Ren, M.; Gao, R.; Stathopoulos, A.; He, X.; Chen, Y.; Liu, D.; Zhangli, Q.; Jiang, J.; Xia, Z.; Srivastava, A.; and Metaxas, D. 2023. Improving Tuning-Free Real Image Editing with Proximal Guidance. arXiv:2306.05414.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv:2208.01626.

Hertz, A.; Voynov, A.; Fruchter, S.; and Cohen-Or, D. 2024. Style Aligned Image Generation via Shared Attention. arXiv:2312.02133.

Huang, Y.; Huang, J.; Liu, Y.; Yan, M.; Lv, J.; Liu, J.; Xiong, W.; Zhang, H.; Chen, S.; and Cao, L. 2024. Diffusion Model-Based Image Editing: A Survey. arXiv:2402.17525.

Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion. arXiv:2403.06976.

Levin, E.; and Fried, O. 2024. Differential Diffusion: Giving Each Pixel Its Strength. arXiv:2306.00950.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Gool, L. V. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv:2201.09865.

Mirzaei, A.; Aumentado-Armstrong, T.; Brubaker, M. A.; Kelly, J.; Levinshtein, A.; Derpanis, K. G.; and Gilitschenski, I. 2023. Watch Your Steps: Local Image and Scene Editing by Text Instructions. arXiv:2308.08947.

Miyake, D.; Iohara, A.; Saito, Y.; and Tanaka, T. 2024. Negative-prompt Inversion: Fast Image Inversion for Editing with Text-guided Diffusion Models. arXiv:2305.16807.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Mou, C.; Wang, X.; Song, J.; Shan, Y.; and Zhang, J. 2024. DiffEditor: Boosting Accuracy and Flexibility on Diffusion-based Image Editing. arXiv:2402.02583.

OpenAI. 2023. DALL·E 3 System Card. Online. Accessed: 2024-01-03.

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:2307.01952.

Ren, J.; Xu, M.; Wu, J.-C.; Liu, Z.; Xiang, T.; and Toisoul, A. 2024. Move Anything with Layered Scene Diffusion. arXiv:2404.07178.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487.

Shi, Y.; Xue, C.; Liew, J. H.; Pan, J.; Yan, H.; Zhang, W.; Tan, V. Y. F.; and Bai, S. 2024. DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8839–8849.

Singh, J.; Zhang, J.; Liu, Q.; Smith, C.; Lin, Z.; and Zheng, L. 2023. SmartMask: Context Aware High-Fidelity Mask Generation for Fine-grained Object Insertion and Layout Control. arXiv:2312.05039.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. arXiv:2212.11565.

Yang, S.; Chen, X.; and Liao, J. 2023. Uni-paint: A Unified Framework for Multimodal Image Inpainting with Pretrained Diffusion Model. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23. ACM.

Yu, J.; Wang, Y.; Zhao, C.; Ghanem, B.; and Zhang, J. 2023. FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model. arXiv:2303.09833.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. arXiv:2306.10012.

Zhang, L.; Rao, A.; and Agrawala, M. 2023a. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.

Zhang, L.; Rao, A.; and Agrawala, M. 2023b. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, X.; Li, Y.; Li, F.; Jiang, H.; Wang, Y.; Zhang, L.; Zheng, L.; and Ding, Z. 2024. Ship-Go: SAR ship images inpainting via instance-to-image generative diffusion models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 207: 203–217.

Zhao, M.; Bao, F.; Li, C.; and Zhu, J. 2022. EGSDE: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations. arXiv:2207.06635.