MagicPaint: Operate Anything for Image Inpainting with Diffusion Model

Qinhong Yang[1,2], Dongdong Chen[3], Qi Chu[1,2], Tao Gong[1,2*], Qiankun Liu[5], Zhentao Tan[6], Xulin Li[1,2], Huamin Feng[4], Nenghai Yu[1,2]

[1] University of Science and Technology of China
[2] Anhui Province Key Laboratory of Digital Security
[3] Microsoft CoreAI
[4] Beijing Electronic Science and Technology Institute
[5] University of Science and Technology Beijing
[6] Independent Researcher
qhyang233@mail.ustc.edu.cn, cddlyf@gmail.com, qchu@ustc.edu.cn, tgong@ustc.edu.cn, liuqk3@ustb.edu.cn, zhentaotan5@gmail.com, fenghm@besti.edu.cn, ynh@ustc.edu.cn

Abstract

Recent diffusion-based models have significantly improved inpainting quality. However, existing methods struggle with multi-task inpainting due to conflicting optimization objectives, and current datasets are typically limited to task-specific scenarios, hindering joint training. To address these challenges, we propose MagicPaint, a unified diffusion-based inpainting model that supports object addition, removal, and unconditional inpainting across both text and image modalities. MagicPaint semantically decouples operation types and target content by learnable tokens in our proposed MMToken Module, effectively reconciling conflicting optimization objectives and enabling robust multi-task, multi-modal inpainting. Besides, we use a novel module named MagicMask, encodes operating intent directly into the mask and applies a mask loss for spatially precise supervision. In addition, existing inpainting datasets are insufficient for multi-task and multi-modal scenarios, limiting the capability of inpainting models. Thus, we further introduce a new dataset comprising 2.1M image tuples. It is dedicatedly designed to support diverse inpainting scenarios and significantly improves upon existing datasets, particularly in object removal. Through efforts from both model and data perspectives, MagicPaint enables users to operate anything—add, remove or inpaint content which is specified through either text or image modalities in a seamless and unified manner. Extensive experiments demonstrate that MagicPaint achieves state-of-the-art performance across three key tasks (i.e., text-guided addition, image-guided addition, and object removal) and produces outputs with superior visual consistency and contextual fidelity compared to existing methods.

Code — https://github.com/littleYaang/MagicPaint

Introduction

Image inpainting aims to restore missing or masked regions in an image by reconstructing backgrounds, removing undesired objects, and synthesizing new content based on user-provided conditions while ensuring seamless visual and contextual coherence. Recent advancements (Lugmayr et al. 2022) have leveraged pre-trained priors from text-to-image generative models (Rombach et al. 2022), significantly enhancing inpainting quality and expanding its applications (Nichol et al. 2021; Avrahami, Lischinski, and Fried 2022; Xie et al. 2023). In addition to background reconstruction, diffusion-based inpainting models are capable of generating intricate, contextually coherent content and synthesizing specific objects conditioned on various conditional inputs (Chen et al. 2024; Yang et al. 2023b).

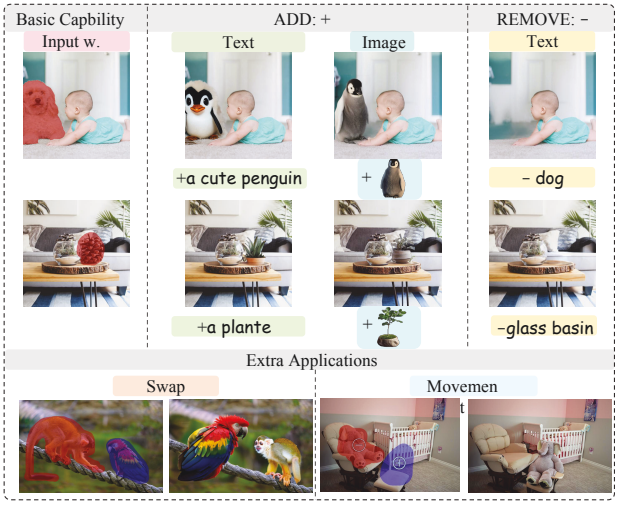

Figure 1: Our unified multi-task inpainting model can support different types of inpainting tasks harmoniously, including but not limited to text-guided/ reference image-guided object addition, text-guided object removal, object swap, and movement.

Despite their strong generative capabilities, existing methods designed for specific tasks (object addition, object removal or inpainting) are only able to achieve success in one of them and incapable of multitasking simultaneously. In particular, for object removal tasks, generative priors often lead to undesired outcomes: instead of accurately restoring the original background, these models tend to introduce unnatural artifacts (Rombach et al. 2022). We analyze the

*Corresponding author.

underlying limitations and categorize them into two main aspects: 1) **Conflicting Optimization Objectives** of object addition, removal and unconditional inpainting, hindering the model's ability to distinguish and handle these tasks effectively. Existing methods attempt to mitigate this through text prompt tuning; however, we contend that such approaches lack finer-grained guidance in both semantic (e.g., coarse-grained text prompt tuning) and spatial dimensions. 2) **Limitations of Training Datasets.** There is a lack of datasets specifically designed for multi-task and multi-modal scenarios. In particular, acquiring paired training data for object removal remains challenging.

To overcome these limitations, we conduct a comprehensive investigation from both model and data perspectives. For the model, we propose MagicPaint, a unified diffusion based inpainting model designed to handle diverse tasks under multimodal user conditions. MagicPaint decouples operations (add, remove and inpaint) and target modalities (text and image) to enhance the semantic distinction between different inpainting operations in **MMToken**. To provide spatially precise supervision to the model, a novel module named, **MagicMask**, which encodes operate intent directly into the input mask with a dedicated mask loss. For the data, we construct a new large-scale multi-task, multi-modal dataset **MMData** based on SA-1B (Kirillov et al. 2023), which covers various inpainting tasks, supporting four key scenarios: context-driven unconditional inpainting, text-guided object addition, reference image-guided object addition, and text-guided object removal.

Specifically, our approach incorporates three key designs: 1) **MMToken Fusion Module**: It decouples the input prompts along two dimensions: operation and modality before fusing multiple tokens to reconcile multi-task conflicts. This decoupling mitigate the verb under-representation and semantic entanglement effectively reconciles conflicting optimization objectives between different operations, while extending support beyond text to include visual conditions. 2) **MagicMask**: Unlike existing inpainting paradigms that use input masks solely to distinguish editable from non-editable regions without explicit task differentiation, the proposed MagicMask encodes operating intent directly within the mask. By predicting a set of output masks and applying pixel-level constraints through a dedicated Mask Loss, MagicMask provides granular, precise guidance to resolve task ambiguities. 3) **MMData**: Beyond model innovations, we revisit data synthesis pipelines and introduce MMData, a large-scale region-targeted dataset tailored for the four key tasks. It supports both text and image conditions, featuring a specialized pipeline for high-quality object removal. With 2.1M high-quality training samples, MMData significantly enhances training efficiency and generalization across scenarios.

In conclusion, our main contributions are as follows:

• We propose MagicPaint, a unified diffusion-based model that effectively handles various inpainting tasks under a single framework. It supports complex multimodal inpainting operations, seamlessly accommodating different user inputs for both background reconstruction and object synthesis.

• We introduce MMData, a large-scale, high-quality inpainting dataset containing 2.1M high-quality editing samples. It captures diverse inpainting scenarios and supports both text and image input modalities.

• We comprehensively evaluate our unified model on multiple inpainting tasks, demonstrating its superiority over existing inpainting methods.

Related Work

Mask-Guided Image Inpainting

Mask-driven image inpainting has advanced significantly, aiming to restore missing regions with contextual coherence. CNN and Transformer-based approaches (Wan et al. 2021; Ko and Kim 2023; Suvorov et al. 2022) introduce dynamic mask updates for more adaptive inpainting but struggle with convergence speed and boundary consistency in extensive missing regions. With the rapid progress in text-to-image generation, diffusion-based models like StableDiffusion (Rombach et al. 2022) and DiffEdit (Couairon et al. 2022) leverage powerful generative capabilities to produce high-quality objects or regions guided by user text prompts. More recent works, such as UniPaint (Yang et al. 2023b), PowerPaint (Zhuang et al. 2023) Paint by Example (Yang et al. 2023a) and AnyDoor (Chen et al. 2024), integrate prompt interpolation and exemplar guidance to enhance object placement precision. While these approaches improve local fidelity, they sometimes compromise global realism and diversity. In this paper, we address this conflict between object removal and addition through both model design and data synthesis strategies.

Multi-task Image Editing

In developing a unified model for diverse image editing operations, existing research mainly explores two directions: training-free approaches and fine-tuning-based methods. Training-free approaches (Dong et al. 2023), often built upon the DDIM Inversion framework (Mokady et al. 2023; Cao et al. 2023; Zhang, Chen, and Liao 2024; Jia et al. 2024). These methods often suffer from inherent DDIM limitations, such as error accumulation and singularity, which lead to inferior editing quality, particularly in object removal, where failure rates are high and artifacts are common. Another notable fine-tuning-based framework for image editing is instruction tuning(Zhang et al. 2023; Hui et al. 2024; Zhao et al. 2024a; Sheynin et al. 2024; Wang et al. 2023), exemplified by InstructPix2Pix(Brooks, Holynski, and Efros 2023). This approach encodes both the operation (e.g., addition or removal) and the desired content into a single text prompt and fine-tunes the model on instructional datasets. However, it struggles with object insertion and removal due to 1) limited training data for removal, 2) conflicts in instruction prompt semantics, and 3) insufficient spatial precision.

Method

MagicPaint Training Paradigms

In Stable Diffusion (SD)-based image inpainting, the training framework is built upon a pretrained text-to-image diffu-

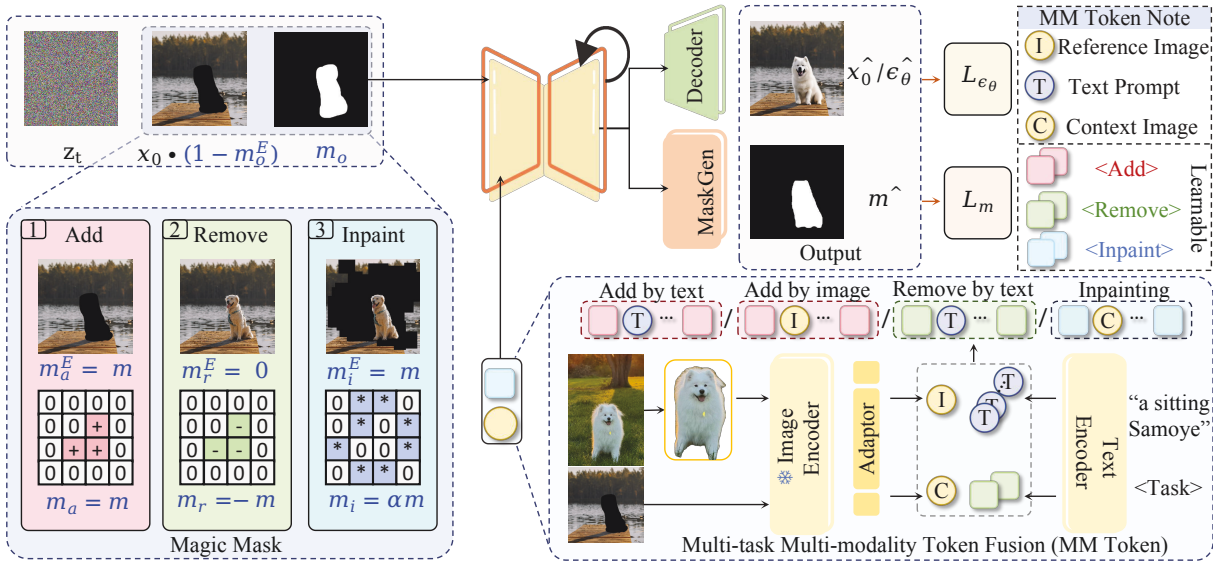

Figure 2: The architecture of MagicPaint comprises two key modules: (1) the MM Token Fusion Module, which aligns diverse input tokens into a unified representation space, thereby enabling robust multi-task and multi-modal inpainting, and (2) the MagicMask Module, which provides precise spatial guidance for multi-operation inpainting tasks.

sion model (Stable Diffusion), which involves both the forward and reverse diffusion processes (Ho, Jain, and Abbeel 2020; Wan et al. 2021).

Forward. Gaussian noise is progressively added to the latent encode x_0 of a clean image I through a Markov chain:

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0,1), \tag{1}$$

where x_t denotes the noised image at timestep t, $\overline{\alpha_t}$ represents the corresponding noise schedule, and ϵ denotes the Gaussian noise.

Reverse. A parameterized neural network θ is trained to estimate the added noise ϵ_t, enabling iterative denoising from pure Gaussian noise to generate coherent images.

Conventional Setup. In the image inpainting task, the input comprises an image I, a binary mask m, and a user-provided instruction c. The latent embedding x_0 is obtained from I via the VAE encoder: $x_0 = E(I)$. The mask m uses values of 0 (for background regions to remain unchanged) and 1 (for areas to be edited). The instruction c specifies both the operation (e.g., removal or insertion) and the desired content for editing. The model's objective is to produce an output image I' that satisfies the operation in c, aligns with the described content, and seamlessly integrates contextually consistent background elements to preserve visual coherence with the unmasked regions. The training optimizes ϵ_θ to predict the noise at timestep t, with the loss formulated as:

$$\mathcal{L}_\epsilon = \mathbb{E}_{x_0,m,t,c,\epsilon_t}\left[\left|\epsilon_t - \epsilon_\theta\left(x_t, \tau_\theta(c), t\right)\right|_2^2\right] \tag{2}$$

where the input x_t to the network is the concatenation $[z_t, m, (1-m)\odot x_0]$, combining the noisy latent z_t, the mask m, and the masked latent embedding of the input image I, the τ_θ is the text encoder.

However, such a paradigm faces significant challenges when applied to unified inpainting, primarily due to **conflicting optimization directions** across tasks, which make joint learning difficult. We contend that existing approaches lack fine-grained guidance at both the 1) *semantic level*: where operation and content are entangled during training, hindering operation-specific optimization. 2)*spatial level*, due to the absence of pixel-level cues necessary for precise control.

To address the above issues, MagicPaint introduces the following key innovations as shown in Figure 2:

1) MMToken: Decoupling across two dimensions, operations and modalities. We explicitly separate operations from content by defining three distinct operation tokens (<Add>, <Remove>, <Inpaint>). Specifically, <Add> denotes object addition, <Remove> denotes object removal, and <Inpaint> refers to the filling of masked regions excluding the target objects (i.e., contextual inpainting). The target content is further classified into two modalities: textual (e.g., descriptive prompts) and visual (e.g., reference images). A detailed explanation of how these tokens are aligned and fused is provided in Section 3.2 (MMToken).

2) MagicMask: Operation-aware spatial guidance. MagicMask provides operation-specific spatial guidance, enabling the model to effectively distinguish between *addition*, *removal*, or *inpainting* operations. To further enhance spatial precision, we introduce *MaskGen*, a dedicated submodule that predicts the fine-grained spatial extent of the target object by assigning distinct values to each operation. This is detailed in Section 3.3 (Magic Mask).

In our MagicPaint training paradigm, the final loss function can be formulated as :

$$\mathcal{L}_\epsilon = \mathbb{E}_{x_0,m,t,c,\epsilon_t}||\epsilon_t - \epsilon_\theta(x_t', \tau_\theta(c'), t)||_2^2 \tag{3}$$

$$\mathcal{L} = \mathcal{L}_\epsilon + \lambda_1(\mathcal{L}_{\text{bce}}(m_{gt}, m') + \lambda_2\mathcal{L}_{\text{dice}}(m_{gt}, m')) \tag{4}$$

where x_t' is obtained from the *Magic Mask*, c' is derived from the *MM Token*, and m' is predicted by the *MaskGen* network. m_{gt} represents the ground-truth fine-grained mask with spatial extent of the operations (detailed in Section

3.3). The loss terms $\mathcal{L}_{\text{bce}}$ and $\mathcal{L}_{\text{dice}}$ denote the binary cross-entropy loss and Dice loss, respectively.

MM Token

Recent studies (Momeni et al. 2023) has shown that CLIP-based text encoders (Radford et al. 2021) are biased toward static concept words—such as objects and backgrounds—while under-representing verbs (add, remove .etc). As a result, in multi-task Stable Diffusion based inpainting, simply adopting an instruction-based prompt format, where operation and target share the same encoding strategy, provides insufficient operation-specific discrimination. Due to insufficient representation of verb semantics, the model often fails to differentiate between operations such as addition and removal.

To address this semantic entanglement between operations and targets, we introduce the **MMToken Module**, which explicitly decouple the multi-task inpainting into two dimensions: *operation* and *target modality*. As shown in the top-right corner of Fig. 2, the *operation tokens* are categorized into three types: {add, remove, inpaint} (depicted as rectangles)—each implemented as an independent learnable token. The *target* token depicted as circles in the figure, depending on its modality: {text, image} are extracted via a corresponding text encoder or image encoder.

For each task, we unify the output representation by defining the MM Token in a consistent form: $c' = [c_{\text{ope}}, c_{\text{text}}]$. where c_{ope} denotes the operation token and c_{text} represents the target feature. In particular, for the *Context-Driven Traditional Inpainting*, the model must ensure seamless contextual consistency within the masked region relative to the surrounding content, an aspect that is difficult to capture using text descriptions. To address this, we extract background image features as a *context token* (Context Image in Fig 2).

After performing the above decoupling, we make the text encoder trainable and jointly optimize it along with the learnable operation tokens to enhance semantic understanding of operate intent. On the other hand, to align the condition content from the image modality with the feature space, we introduce an additional MLP layer attached to the image encoder as a feature projection module. This alignment ensures that both text and image conditions are mapped into a unified embedding space.

Finally, the MM Token is fed into the diffusion model as c' in Eq. (3). This MM Token effectively reconciles the conflicting optimization objectives between object removal and addition, while also extending support from purely text-based conditions to visual inputs. By aligning diverse tokens into a unified representation space, the module enables robust multi-task and multi-modal inpainting.

MagicMask

We further introduce the **MagicMask** module, which incorporates *operation-aware spatial guidance* to help the model explicitly distinguish between different editing operations, thereby enabling more effective multi-task optimization.

For each operation $o \in \{\text{add}, \text{remove}, \text{inpainting}\}$ defined by MMToken, we apply distinct processing strategies to the network input:

$$x'_t = [z_t, m_o, x_0 \odot (1 - m_o^E)], \tag{5}$$

as illustrated in the lower-left corner of Fig. 2. Here, m_o corresponds to the mask for operation and serves a dual role: it specifies the editing region and simultaneously encodes the operation type. The term m_o^E denotes the mask applied to the input image latent vectors. In conventional inpainting pipelines, m_o^E is uniformly set to m across all operations, whereas in our design. it is operation-specific.

Given the condition of the target object, we conceptualize *add* and *remove* as two opposing operations within a unified inpainting framework. Notably, since the **MMToken** incorporates a *context token* to provide background information for inpainting, so in m_o we adopt the same mask strategy for *inpaint* as for *add*. Formally, m_o is defined as follows: it equals positive m when the operation o is *add* or *inpaint*; it equals negative m when o is *remove*.

For the input $x'_o = x_0 \odot (1 - m_o^E)$, which provides the model with information about the visible regions of the image, we also adopt operation-specific strategies to deliver tailored guidance for different tasks. For the *remove* task, following the Masked-Region Guidance paradigm introduced in SmartEraser (Jiang et al. 2025), we set $m_r^E = 0$ which means in removal scenarios, the model is provides with a full image: $x'_r = x_0 \odot (1 - 0) = x_0$. The model then identifies and eliminates the target object based on the provided prompt. In contrast, for *inpainting* tasks, the input image is inherently incomplete (i.e., masked), and the model must synthesize missing content. Therefore, we set the input to $x'_i = x_0 \odot (1 - m^E)$. Similarly, for object addition tasks, we apply the same masking strategy: $x'_a = x_0 \odot (1 - m_o^E)$, to prevent the model from simply copying visible content into the masked region. This design aligns with strategies employed in prior diffusion-based editing frameworks to encourage meaningful content generation. This process can be formalized as: m_o^E equals m if the operation o is in the set add, inpaint, and equals 0 if o is remove.

We input $x'_t = [z_t, m_o, x_0 \odot (1 - m_o^E)]$ into the diffusion model. In addition to obtaining the network-predicted noise $\hat{\epsilon_\theta}$, we also utilize MaskGen to predict m' by first deriving the predicted $\hat{x_0}$ from $\hat{\epsilon_\theta}$ (via the standard DDPM sampling equation $\hat{x_0} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon_\theta}}{\sqrt{\bar{\alpha}_t}}$), and then combining $\hat{x_0}$ with the timestep t and the MM token to generate the predicted mask:

$$m' = \text{MaskGen}(\hat{x_0}, t, \tau_\theta(c')), \tag{6}$$

where MaskGen is a network and c' is the MM Token.

To provide stronger supervision, we assign distinct ground-truth labels to masks corresponding to different operations and apply data augmentation (e.g., erosion, dilation) to the input masks. This ensures that the predicted masks are constrained by accurate multi-label supervision. Specifically, we employ operation-specific masks by assigning distinct values to each operation: *add* is defined as the first class, *remove* as the second class, and *inpainting* as the third class. To further guide learning, we introduce a mask loss $L_m(m_{gt}, m')$ (Eq. 4), which encourages the predicted mask m' to closely match the ground-truth mask m_{gt}.

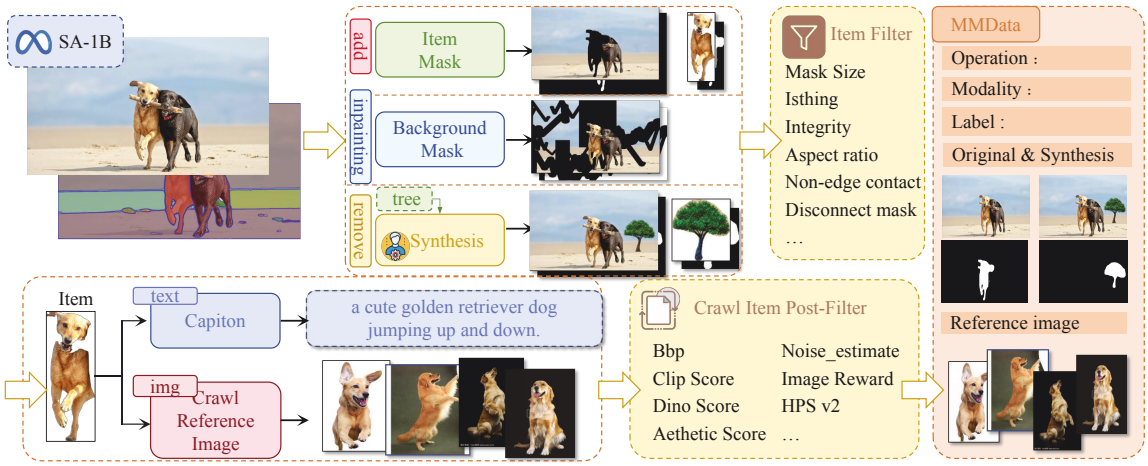

Figure 3: The construction pipeline of MMData. We developed four distinct data construction processes tailored to each task, employing substantial volumes of meticulously annotated data from (Rasheed et al. 2024; Schuhmann et al. 2022) to generate paired annotation sets through systematic quality assurance protocols.

MMData

We discuss the limitations of prior datasets, highlight the advantages of MMData, and outline MMData's construction process.

Limitations of Existing Datasets

• **Lack of multi-task and multi-modal datasets.** Prior works (Yu et al. 2025; Winter et al. 2024) demonstrate that joint training on object insertion and removal enhances a model's grasp of physical consistency, as these tasks are counterfactual and interdependent. However, available datasets for such training are either small-scale or confined to single tasks, lacking support for large-scale multi-task and multi-modal inpainting. A comprehensive comparison is available in the supplementary material.

• **Difficulty in obtaining paired data for object removal.** Securing high-quality paired data for real-world object removal remains challenging. Among comparable datasets: Counterfactuals in Object Drop (Winter et al. 2024) and OmniPaint (Yu et al. 2025) use real paired data from identical scenes, preserving physical effects like shadows and occlusions, but high collection costs limit scalability. Inst-Inpaint (Yildirim et al. 2023) employs traditional inpainting models (e.g., LAMA (Suvorov et al. 2022)) for synthetic pairs, yet suboptimal inpainting quality caps removal performance. SmartErase (Jiang et al. 2025) pastes objects synthetically onto backgrounds for scalability, but it inadequately captures complex interactions such as realistic placement and occlusions.

To surmount these issues, MMData facilitates unified training with 2.1M edits, will be visualized in supplementary.

Advantages of Our Dataset

• **Multi-task image editing.** Unlike existing instruction-based datasets, MMData offers multi-modal prompts (textual and visual), high-resolution images, and explicit region-specific edits. With 2.1M samples, it bridges the scale gap in object inpainting datasets (see supplementary.).

• **Image object removal.** Leveraging Diffree (Zhao et al. 2024b), trained on real scenes (e.g., COCO), MMData captures physical relationships like occlusions and shadows across 1.8M removal edits. It also uniquely enables multi-turn removal in the same scene, absent in prior datasets (in supplementary).

Dataset Construction

Building on the SA-1B segmentation dataset (Kirillov et al. 2023), we create four task-specific subsets via tailored processing: (1) context-driven unconditional inpainting, (2) text-guided object addition, (3) reference image-guided addition, and (4) text-guided object removal. The pipeline (Fig. 3) involves:

•**Step 1**: Extracting cropped objects and masks from SA-1B annotations for addition; for removal, inserting objects into scenes alongside random background masks.

•**Step 2**: Filtering low-quality samples by mask size and aspect ratio.

•**Step 3**: Generating text prompts via captioning models and retrieving similar images for multi-modal support.

•**Step 4**: Applying quality filters using CLIP, DINO, and semantic metrics.

This yields the high-quality MMData; detailed pipeline specifics are in the supplementary material.

Experiments

Experimental Setup

Implementation Details. To ensure a fair comparison, we follow standard experimental protocols and fine-tune MagicPaint using the Stable Diffusion v1.5 Inpainting backbone. The joint training process updates the UNet, text encoder, an MLP for visual–text feature alignment, learnable operation tokens, and a 34M-parameter MaskGen network. Our model is trained on the proposed MMData dataset with a batch size of 32 for 800K iterations. Optimization is performed using

Method	40-50% masked		All samples	
	FID ↓	LPIPS ↓	FID↓	LPIPS ↓
DeepFillv2 (Yu et al. 2019)	29.80	0.2469	7.55	0.1508
LaMa (Suvorov et al. 2022)	21.07	0.2133	3.48	0.1193
SD-Inpainting(Rombach et al. 2022)	19.73	0.2322	3.03	0.1312
PowerPaint (Zhuang et al. 2023)	17.91	0.2225	2.59	0.1263
MagicPaint†	**12.32**	**0.1921**	**2.19**	**0.0967**

Table 1: Quantitative comparisons with existing methods for context-driven unconditional inpainting on Places2 (Zhou et al. 2017). The input images are maksed with the mask provided by LaMa (Suvorov et al. 2022).

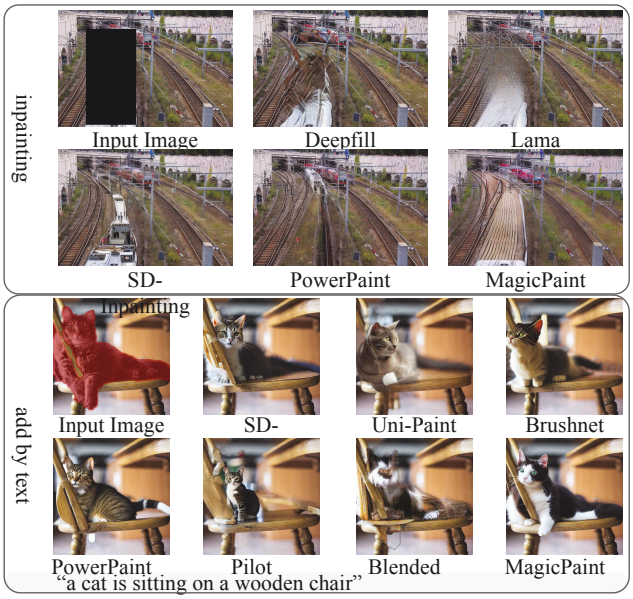

Figure 4: Compared with existing methods for context-driven unconditional inpainting and text-guided object addition task.

Method	MSCOCO (Lin et al. 2017)		
	FID ↓	L-FID ↓	CLIP↑
BLD (Avrahami et al. 2023)	35.11	8.66	24.91
SD-Inpainting (Rombach et al. 2022)	15.77	5.65	24.81
UniPaint (Yang et al. 2023b)	16.68	5.64	26.17
BrushNet (Ju et al. 2024)	13.40	6.08	25.05
PowerPaint(Zhuang et al. 2023)	11.61	**5.12**	25.95
PILOT (Pan et al. 2024)	13.80	5.84	25.24
MagicPaint†	**11.29**	5.53	**27.73**

Table 2: Quantitative comparison with SOTA models for text-guided addition inpainting with object layout masks on MSCOCO (Lin et al. 2017) dataset. 10K images are randomly sampled for evaluation and the instance mask with the largest area is used to produce the masked-input image.

AdamW with a uniform learning rate of 2×10^{-5} applied to all trainable modules.

Task Settings. As a unified multi-task framework, the proposed MagicPaint is evaluated across diverse inpainting scenarios. We categorize the tasks as four tasks that is, (1) Context-Driven Unconditional Inpainting, (2) Text-Guided Object Addition Inpainting, (3) Reference Image-Guided Addition Inpainting, (4) Text-Guided Object Removal Inpainting.

Method	COCOEE (Yang et al. 2023a)		
	DINO ↑	LPIPS ↓	FID ↓
PBE (Yang et al. 2023a)	11.99	0.4034	**12.17**
Anydoor (Chen et al. 2024)	14.26	0.5375	24.42
MagicPaint†	**14.52**	**0.1691**	14.98

Table 3: Quantitative comparisons with existing methods for reference image-guided addition inpainting on COCOEE (Yang et al. 2023a), which provided a reference image for a masked input image.

Method	MagicBrush (Zhang et al. 2023)	
	REMOVE ↑	FID ↓
SD-Inpainting(Rombach et al. 2022)	0.768	86.0
PowerPaint (Zhuang et al. 2023)	0.763	83.7
DesignEdit (Jia et al. 2024)	0.811	67.3
SmartEarse (Jiang et al. 2025)	0.787	71.5
MagicPaint†	**0.817**	**65.8**

Table 4: Quantitative comparison with existing methods for text-guided object removal inpainting on MagicBrush (Zhang et al. 2023). with "remove" instruction.

Evaluation Benchmarks. To thoroughly validate our method's effectiveness, we compare MagicPaint against state-of-the-art (SOTA) methods on the following public benchmarks: MSCOCO (Lin et al. 2017) for Text-Guided Object Addition Inpainting and MagicBrush (Zhang et al. 2023) for Text-Guided Object Removal Inpainting, COCOEE (Yang et al. 2023a) for Reference Image-Guided Addition Inpainting, and Places2 (Zhou et al. 2017) for Context-Driven Unconditional Inpainting.

Comparisons with State-of-the-Art

Context-Driven Unconditional Inpainting. As shown in Table 1 and Figure 4, in the task of context-driven unconditional inpainting task, MagicPaint outperforms SOTA inpainting methods, including traditional approaches (LaMa, DeepFillv2), which often yield blurry and poorly integrated details, and diffusion-based methods such as SD-Inpainting and the recent PowerPaint. Our method achieves the lowest FID and LPIPS, while producing clear, continuous structures with accurate textures.

Text-Guided Object Addition Inpainting. As shown in Table 2 and Figure 4, MagicPaint outperforms existing methods on text-guided object addition inpainting, achieving the best performance in FID Score and the second best L-FID, which suggests high realism of generated content. Besides, our MagicPaint gets a better CLIP score than PowerPaint (Zhuang et al. 2023), indicating superior alignment between the generated objects and text descriptions.

The qualitative results are shown in Figure 4. MagicPaint achieves superior performance in generating objects that accurately align with text descriptions while maintaining con-

Module	Task	Add by Text		Add by Image		Remove by text		Inpaint	
		L-FID↓	CLIP↑	FID ↓	LPIPS ↓	REMOVE↑	FID↓	FID ↓	LPIPS ↓
SD-I	ori	5.65	24.81	-	-	0.768	86.0	3.03	0.1312
SD-I	remove text	9.13	18.77	-	-	0.787	71.7	3.00	0.1307
SD-I (+VE)	all	7.29	22.23	24.98	0.2761	0.743	112.3	2.71	0.1339
+ MM Token		7.07	25.86	19.43	0.1898	0.783	74.9	2.67	0.1134
+ MagicMask		6.84	25.25	19.98	0.2375	0.796	72.1	2.46	0.1227
MagicPaint†		**5.53**	**27.73**	**14.98**	**0.1691**	**0.817**	**65.8**	**2.19**	**0.0967**

Table 5: Ablation study on proposed module. Here, SD-I(+VE) refers to the variant of SD-I that concatenates features extracted by the image encoder with text embeddings as the condition. *ori* represents the pretrained model without fine-tuning, *remove text* indicates single-task training on the removal task only, and *all* denotes training on all tasks jointly.

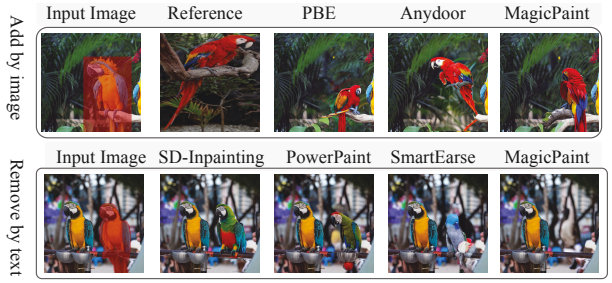

Figure 5: Compared with existing methods for reference image-guided object addition and text-guided object removal inpainting task.

textual consistency. Compared to SD-Inpainting (Rombach et al. 2022), which often introduces artifacts (the *e.g.*, case in line 2), and Uni-Paint BrushNet, and PowerPaint, which struggle with fine details, MagicPaint produces more realistic objects with better integration into the background.

Reference Image-Guided Addition Inpainting. The results of reference image-guided addition inpainiting is shown in Table 3. Compared with the methods that are dedicated to the reference image-guided addition inpainting task, our MagicPaint achieves the best DINO and LPIPS scores and a comparable FID score, indicating our capabilities.

Shown in Figure 5, MagicPaint accurately reproduces the color, texture, and fine details of reference objects, outperforming PBE (Yang et al. 2023a) in all cases. Compared with Anydoor (Chen et al. 2024), it also produces more plausible object placement (e.g., avoiding unrealistic positions such as floating in mid-air) by leveraging background context for more coherent integration.

Text-Guided Object Removal Inpainting. Lastly, the results of the text-guided object removal inpainting task are presented in Table 4. As we can see, MagicPaint achieves the best REMOVE score and FID score, which are superior to that of PowerPaint and SD-Inpainting. This demonstrates MagicPaint's effectiveness in removing objects while keeping consistency with the background.

As shown in Figure 5, MagicPaint produces the most realistic results, fully removing target objects and restoring coherent backgrounds. In contrast, SD-Inpainting leaves residual traces, and PowerPaint (Zhuang et al. 2023) often generates incomplete or unrealistic background textures.

Ablation Studies

Conflicts Between Operations. As shown in Table 5, we first examine the conflicts between different operations. The 1st row reports the original SD-Inpainting model, which performs well on *addition* tasks due to strong text priors, but performs poorly on *remove*. Fine-tuning on a single-task removal setting (2nd row) improves removal performance but causes a clear drop in addition performance, indicating conflicting optimization directions. Mixed multi-task training (3rd row, SD-I+VE) further exacerbates this issue, significantly degrading performance of addition, which shows that jointly optimizing multiple operation types is challenging.

Effectiveness of Modules. To address above issues, we introduce two key modules: MMToken and MagicMask. When used individually, MMToken yields significant gains for the image-guided setting and improves all tasks over the baseline, not only the benefit of disentangling operations from content for better task discrimination but also its advantage in dealing alignment across modalities. MagicMask, on the other hand, brings notable improvements on the removal task, confirming the benefit of operation-aware spatial guidance in improving removal accuracy.

When combined, MMToken and MagicMask deliver the best performance across all four tasks. We attribute this to the synergy between semantic disentanglement and operation-specific spatial supervision, where the operation-conditioned mask prediction in MagicMask aligns well with the operation-content separation in MMToken, leading to more accurate mask predictions and consistent gains across all tasks.

Conclusion

This paper presents **MagicPaint**, a unified diffusion-based model for diverse inpainting tasks. By semantically disentangling operation types from target content and encoding operation intent directly into the mask to provide operation-aware spatial supervision, MagicPaint alleviates the conflicts inherent in multi-task inpainting. To further enable effective training, we introduce **MMData**, a large-scale dataset with 2.1M samples, dedicatedly designed to support diverse inpainting scenarios and significantly improves upon existing datasets, particularly in object removal. Trained on MMData, MagicPaint achieves high-quality, contextually consistent results across various scenarios, laying a strong foundation for future multimodal inpainting and image synthesis research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62472396, 62121002), Anhui Provincial Natural Science Foundation (2508085QF212) and the advanced computing resources provided by the Supercomputing Center of the USTC.

References

Avrahami; et al. 2023. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4): 1–11.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218.

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.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6593–6602.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Dong, W.; Xue, S.; Duan, X.; and Han, S. 2023. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7430–7440.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hui, M.; Yang, S.; Zhao, B.; Shi, Y.; Wang, H.; Wang, P.; Zhou, Y.; and Xie, C. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*.

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*.

Jiang, L.; Wang, Z.; Bao, J.; Zhou, W.; Chen, D.; Shi, L.; Chen, D.; and Li, H. 2025. Smarteraser: Remove anything from images using masked-region guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24452–24462.

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 preprint arXiv:2403.06976*.

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.

Ko, K.; and Kim, C.-S. 2023. Continuously masked transformer for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13169–13178.

Lin, T.-Y.; Patterson, G.; Ronchi, M. R.; Cui, Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, L.; and Dollár, P. 2017. COCO 2017: Common Objects in Context 2017.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

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.

Momeni, L.; Caron, M.; Nagrani, A.; Zisserman, A.; and Schmid, C. 2023. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15579–15591.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Pan, L.; Zhang, T.; Chen, B.; Zhou, Q.; Ke, W.; Süsstrunk, S.; and Salzmann, M. 2024. Coherent and Multi-modality Image Inpainting via Latent Space Optimization. *arXiv preprint arXiv:2407.08019*.

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.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13009–13018.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sheynin, S.; Polyak, A.; Singer, U.; Kirstain, Y.; Zohar, A.; Ashual, O.; Parikh, D.; and Taigman, Y. 2024. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8871–8879.

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.

Wan, Z.; Zhang, J.; Chen, D.; and Liao, J. 2021. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4692–4701.

Wang, S.; Saharia, C.; Montgomery, C.; Pont-Tuset, J.; Noy, S.; Pellegrini, S.; Onoe, Y.; Laszlo, S.; Fleet, D. J.; Soricut, R.; et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18359–18369.

Winter, D.; Cohen, M.; Fruchter, S.; Pritch, Y.; Rav-Acha, A.; and Hoshen, Y. 2024. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *European Conference on Computer Vision*, 112–129. Springer.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22428–22437.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023a. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Yang, S.; Chen, X.; Liao, J.; et al. 2023b. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3190–3199.

Yildirim, A. B.; Baday, V.; Erdem, E.; Erdem, A.; and Dundar, A. 2023. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4471–4480.

Yu, Y.; Zeng, Z.; Zheng, H.; and Luo, J. 2025. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. *arXiv preprint arXiv:2503.08677*.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36: 31428–31449.

Zhang, Z.; Chen, D.; and Liao, J. 2024. SGEdit: Bridging LLM with Text2Image Generative Model for Scene Graph-based Image Editing. *arXiv preprint arXiv:2410.11815*.

Zhao, H.; Ma, X. S.; Chen, L.; Si, S.; Wu, R.; An, K.; Yu, P.; Zhang, M.; Li, Q.; and Chang, B. 2024a. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37: 3058–3093.

Zhao, L.; Yang, T.; Shao, W.; Zhang, Y.; Qiao, Y.; Luo, P.; Zhang, K.; and Ji, R. 2024b. Diffree: Text-guided shape free object inpainting with diffusion model. *arXiv preprint arXiv:2407.16982*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.

Zhuang, J.; Zeng, Y.; Liu, W.; Yuan, C.; and Chen, K. 2023. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint arXiv:2312.03594*.