SIMON: SImultaneous Multi-Object Navigation

Yifeng Zhu[1,†], Siyuan Huang[2,†], Jun Bao[3], Jun Yu[1], and Buyu Liu[1,*]

[1]School of Intelligence Science and Engineering [2]School of Computer Science and Technology [3]School of Biomedical Engineering
Harbin Institute of Technology (Shenzhen)

Abstract. This paper tackles the challenge of simultaneously manipulating multiple objects in image editing—a scenario that extends beyond the capabilities of existing methods focused on single-object manipulation. Our approach SIMON efficiently handles complex rearrangements such as object shuffling, achieving higher visual quality at both image and region levels, while reducing inference time. To this end, we propose three key modules: a content-aware attention mechanism for region-adaptive focus, a multi-object energy guidance strategy for subtask-specific consistency, and a latent initialization technique for artifact suppression. Together, these components enhance visual coherence across inpainting, object relocation, and background preservation subtasks. To evaluate our framework, we curate 1,000 images from the 3D-FUTURE dataset, each containing 2 to 5 target objects. For every object, we annotate a new target location, ensuring that the resulting layouts are physically plausible and semantically meaningful. We conduct comparative experiments against three state-of-the-art methods, using both image-level, region-level, and instance-level metrics. Both quantitative results and human analysis confirm that our framework delivers significant improvements in both visual quality and computational efficiency. We further validate our design through extensive ablation studies with diverse architectures and applications. The code and annotated dataset are publicly available at https://github.com/yijichar/SIMON-SImultaneous-Multi-Object-Navigation.

Keywords: Text To Image Generation · Controllable Diffusion Models · Multi-Instance Editing

1 Introduction

Image editing [6, 29] has long been a central topic in computer vision. Depending on the nature of the task and constraints, a variety of approaches have been explored, including GAN-based methods [7, 15], NeRF-based techniques [2, 23], and diffusion models [11, 32]. In particular, fine-grained instance-level manipulation [9, 21, 40] has attracted growing interest due to its broader applicability compared to traditional image-level editing.

† Equal contribution.
* Corresponding author.

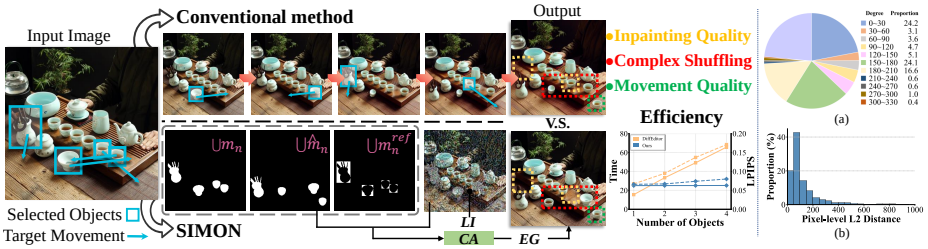

Fig. 1: SIMON delivers greater efficiency (see chart) and superior visual quality over existing methods. Visual comparisons are highlighted by the dashed bounding box. Additionally, subfigures (a) and (b) illustrate the distribution of object movements in our annotated dataset in terms of angles and L2 distances, respectively.

Though several methods—such as DragGAN [28] and StyleGAN-XL [35]—have been proposed to address the aforementioned challenges, they are often constrained by the limited generative capacity of GANs [7]. Recently, diffusion models—particularly DragonDiff [25] and DiffEditor [24]—have gained more attention due to their impressive generalization capabilities, training-free manner, and state-of-the-art performance on facial datasets [14]. Nevertheless, we find that these methods struggle with multi-object manipulation, especially in scenes featuring complex backgrounds or intricate spatial layouts (see Fig. 1).

In this paper, we aim to address the challenge of simultaneous multi-object manipulation in image editing—a task that goes beyond the scope of existing methods primarily designed for single-object scenarios. We propose *SIMON*, a framework capable of handling complex rearrangements such as object shuffling within diverse and cluttered backgrounds (see Fig. 1). SIMON delivers superior visual quality at both the image and region levels, all while maintaining inference times comparable to those of single-object approaches. To this end, we decompose the overall task into three subtasks: inpainting, object relocation, and background preservation. Based on this decomposition, we propose three novel modules—content-aware attention, multi-object energy guidance, and latent initialization—each grounded in pre-defined task-specific masks. The content-aware attention module selectively boosts attention weights within instance-relevant regions using independent multiplicative scaling and additive biasing to solve attention diffusion, thereby enhancing fine-grained visual consistency and ensuring precise alignment between editing constraints and outputs. The multi-object energy guidance module introduces a composite energy function comprising three subtask-specific terms, each enforcing similarity between generated and original features within designated regions. This design enables synchronized multi-object manipulation, leading to better practical efficiency, and significantly improves overall editing quality. Lastly, the latent initialization module reduces residual artifacts in inpainting regions by replacing the cached latents with Gaussian noise sampled from a standard normal distribution, while keeping the background latents unchanged during the early steps. This effectively

removes inherited distortions and allows the diffusion model to fully leverage its generative prior, yielding more coherent and visually natural reconstructions.

Rather than working on face datasets as existing work did [28], we turn to 3D-FUTURE [5], which is more layout-aware and human interpretable. We select 1000 images from its test set based on object count inside. And we annotate the new location of each object in one image, ensuring that the new layout is physically meaningful and reasonable. More details of our annotated images can be found in Fig. 1. We evaluate our method on the annotated image dataset, comparing it against three SOTA approaches: DragonDiff [25], DiffEditor [24], and GeoDiffuser [34]. These methods were selected for their public availability, strong performance, and to ensure a fair comparison. Our evaluation includes both image-level, region-level, and instance-level comparisons using standard metrics, including 4 for inpainting quality, 2 for object relocation accuracy, 2 for background preservation, 1 for identity preservation, and 1 for overall quality. Results clearly demonstrate the effectiveness and efficiency of SIMON for multi-object manipulation. For example, SIMON outperforms the SOTA training-free DiffEditor by 50.34%, 20.64%, and 26.87% across MSE, LPIPS, and FID, using only 30.3% of its runtime. Notably, SIMON surpasses MagicFixup [1] and Insert [38]—trained on 5M images and 159K pairs, respectively—in both perceptual quality and inference speed, without any training cost. In addition, we conduct a human analysis involving 5 participants, each anonymously voting for the image they find more natural and realistic. Again, SIMON is consistently preferred in our user study. In all, our contribution can be summarized as follows:

- A novel framework designed for simultaneous fine-grained manipulation of multiple objects in images with diverse backgrounds and complex layouts.
- Three novel modules enabling region-adaptive attention, subtask-specific consistency enforcement, and effective artifact suppression.
- SOTA performances with reduced inference time across three subtasks, validated by image-level, region-level, and instance-level metrics, along with consistent preference in human evaluations.

2 Related Work

Instance-based image editing Instance-level image editing involves precise control over individual objects in an image, supporting a range of manipulations such as addition, removal, relocation, replacement, and resizing. One of the key challenges in this area is accurately manipulating specific instances, particularly in complex scenes or when dealing with large spatial transformations. Early approaches, such as [47], leveraged GAN-based models [7] to generate face images with varying attributes (e.g., adding glasses), pushing beyond traditional image-to-image style transfer [12]. Recent works like DragGAN [28] introduced interactive, point-based editing in GAN latent space [7], allowing for smooth object deformation. However, GANs often struggle with generalization in complex or diverse scenes due to their limited expressiveness. With the emergence

of more advanced generative models, diffusion models [11] have become a compelling alternative. These models have shown strong capabilities in instance-level editing by integrating multimodal guidance, such as textual prompts [26], bounding boxes [46], or both [41], enabling more flexible and semantically aware manipulations. Most existing works [4, 24, 25, 34, 42] have focused on editing a single object. DragonDiffusion [25] and its successor DiffEditor [24] have successively achieved state-of-the-art performance in terms of visual fidelity and control accuracy on face benchmarks [14]. More recent work [34] further advances the paradigm by imposing constraints on attention maps to support both 2D and 3D editing in a training-free manner. However, their generalization and efficiency in handling multi-object manipulations, especially under complex rearrangements and in diverse scene contexts, remain largely unexplored. Though MagicFixup [1] attempts to tackle multi-instance editing, it relies on extensive training on large-scale datasets, incurring significant computational overhead and limiting its practical flexibility. SceneDiffusion [31] aims to address complex layouts by introducing a layered representation for compositional layout reasoning, it remains evaluated under single-object settings and incurs substantial memory and computation overhead due to the need to denoise multiple layouts simultaneously. Similar limitation is observed in [42], as the sequence-to-sequence video diffusion model faces difficulties in handling non-coplanar and multi-object interactions, and incurs significant computational overhead. In this work, we introduce a training-free framework for multi-instance image editing that enables manipulation of multiple objects with strong consistency and coherence. Experiments on a complex indoor dataset demonstrate superior performances over state-of-the-art approaches in both editing quality and runtime efficiency.

Diffusion models and external conditions In diffusion models [11], data generation is formulated as a reverse process that reconstructs real data by learning to reverse a predefined forward process, which gradually corrupts data into pure Gaussian noise. This process typically unfolds over a sequence of T steps and is often formulated as a Markov chain, where each step $t \in \{0, \ldots, T\}$ produces a noisy sample $\mathbf{x}_t$, with $\mathbf{x}_0$ being the original image and $\mathbf{x}_T$ approaching standard Gaussian noise. The key objective during sampling is to estimate the score function $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$—the gradient of the log-probability density—which is used to generate $\mathbf{x}_{t-1}$ by guiding $\mathbf{x}_t$ toward regions of higher data density. Typically, U-Net [11] is involved in this sampling process by taking a noisy latent vector and predicting the noise that needs to be removed. When incorporating external conditions $\mathbf{c}$, methods like DreamBooth [33], GLIGEN [19], and ControlNet [43] offer diverse solutions for personalization, task-specific control, and general-purpose conditioning. In contrast, conditional diffusion models [3] provide a training-free way to steer the sampling trajectory based on $\mathbf{c}$. This results in a modified gradient of the form $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{c}) \propto \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c} \mid \mathbf{x}_t)$, where the second term encodes guidance from $\mathbf{c}$ to better align the generation process with the desired context. To effectively model the second term, energy function $E(\mathbf{x}_t, \mathbf{c})$ is thereby introduced [3]. A suitable energy function should satisfy: $\nabla_{\mathbf{x}_t} \log p(\mathbf{c} \mid \mathbf{x}_t) \propto -\nabla_{\mathbf{x}_t} E(\mathbf{x}_t, \mathbf{c})$. Therefore, conditioning on $\mathbf{c}$ during

sampling is equivalent to adding the negative gradient of the energy function to the original score estimate. The sampling update under energy-based guidance is formulated as $\mathbf{x}_{t-1} = \hat{\mathbf{x}}_t - \eta \nabla_{\mathbf{x}_t} E(\mathbf{x}_t, \mathbf{c})$, where $\hat{\mathbf{x}}_t$ is the standard denoised estimate at timestep t, and η is a guidance strength hyperparameter. This framework offers a unified and modular approach for conditioning diffusion models, enabling a variety of downstream applications—including classifier guidance [8], free-form inpainting [22], feature-level similarity constraints [13], and complex attribute manipulation [25]—without requiring retraining of the underlying diffusion model. In our work, we propose a multi-object energy guidance mechanism that integrates three subtask-specific terms to enforce feature-level similarity between the generated and original content within designated regions—all without requiring any training.

3 Our Framework: SIMON

When manipulating multiple objects in an image, three key subtasks must be addressed. First, the vacant regions left by the relocated target objects should be naturally inpainted to match the surrounding context. Second, the appearance of the target objects should be faithfully preserved at their new locations. Lastly, the rest of the image should remain unchanged to ensure overall visual consistency. As foundational components, we leverage DDIM inversion [37] and the key-value (KV) replacement strategy [36] to cache intermediate latents and features, enabling fine-grained manipulation during subsequent guidance. To address the aforementioned subtasks, we propose SIMON, a unified framework that incorporates a content-aware attention mechanism into the self-attention layers. This design simultaneously enhances background inpainting, object relocation, and context preservation (Sec. 4.3). In addition, SIMON exploits the cached features to enable gradient-based multi-object guidance, steering the generation process toward the desired editing objectives (Sec. 3.1). To further ensure that no latent information from the original image contaminates the inpainting process, SIMON integrates a latent initialization technique (Sec. 3.2). Importantly, all components in SIMON are plug-and-play and can be seamlessly integrated into existing diffusion-based generative models without the need for fine-tuning or retraining. See Fig. 2 for an overview[1].

Let the original image be denoted as $\mathbf{x_0}$. We define the set of target objects as $\{\mathbf{o}_n\}_{n=1}^{N}$, where each object $\mathbf{o}_n$ is annotated by a bounding box, and N is the total number of objects. These bounding boxes can be converted into binary masks using off-the-shelf tools [17]; we denote the binary mask for the n-th object as $\mathbf{m}_n$. For each object, we also specify a desired new location, defined by its 2D coordinates $c_n = [c_n^x, c_n^y]$. Assuming the object scale remains unchanged[2],

[1] The pipeline is shown with a U-Net backbone for clarity, but SIMON is applicable to architectures whose latent/token grids preserve approximate spatial correspondence. Examples are shown in Appendix.

[2] SIMON also supports scale variation if such information is provided during relocation. Examples are shown in our experiments.

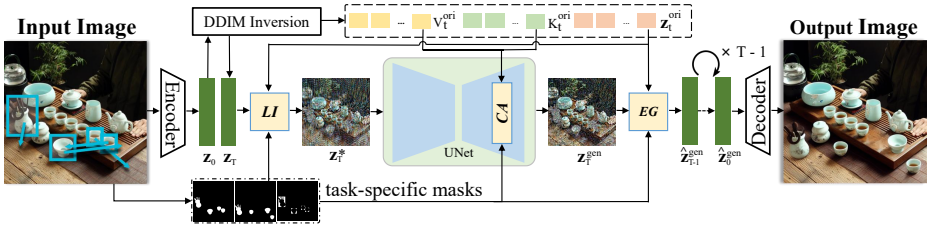

Fig. 2: Given an input image $\mathbf{x}_0$, along with the targeted objects and their desired locations annotated using bounding boxes and arrows, DDIM inversion is first applied to obtain the latent $\mathbf{z}_T$ and corresponding step-wise cached features. Our SIMON framework then employs Latent Initialization (LI) to mitigate the negative influence of $\mathbf{z}_T$ in the inpainting regions, while leveraging contextual information from the surrounding areas during early denoising steps to achieve natural and effective inpainting. Content-aware Attention (CA), along with the cached features, is integrated at each denoising step to selectively enhance task-relevant regions without disturbing the surrounding context. Combined with our Multi-Object Energy Guidance (EG), SIMON produces seamless and coherent results with significantly improved runtime efficiency compared to existing methods.

we compute the translated mask at the desired location as $\hat{\mathbf{m}}_n$ (See Fig. 3). Our goal is to generate a new image where multiple objects are relocated to their designated positions, while ensuring visually plausible inpainting of vacated regions, faithful appearance preservation of the objects, and no alteration to the untouched areas.

Preliminary. As demonstrated in [3, 16], *DDIM inversion* [37] offers a strong generative prior that helps preserve consistency with the original image. Building on this, we utilize DDIM inversion to map the original image x_0 into a Gaussian latent $\mathbf{z}_T$ within the latent space of the pre-trained diffusion model [18,32]. During this inversion process, we also cache intermediate features at each timestep, including the key and value tensors from the attention layers—denoted as K_t^{ori} and V_t^{ori}—as well as the latent variables $\mathbf{z}_t^{\text{ori}}$. These are stored in a memory bank and later used to guide various stages of the editing process [3]. Specifically, these saved keys and values are reused during the attention computation in the generation process, as typical *KV replacement* [36] suggested:

$$A_t(Q_t^{\text{gen}}, K_t^{\text{ori}}) = \frac{Q_t^{\text{gen}}(K_t^{\text{ori}})^\top}{\sqrt{d}} \tag{1}$$

$$\text{Att}(Q_t^{\text{gen}}, K_t^{\text{ori}}, V_t^{\text{ori}}) = \text{softmax}(A_t(Q_t^{\text{gen}}, K_t^{\text{ori}}))V_t^{\text{ori}} \tag{2}$$

where Q_t^{gen} is the query produced by diffusion layer.

[3] Note that SIMON supports the use of external reference images for appearance replacement by incorporating them into the DDIM inversion procedure. Examples are shown in Appendix.

Fig. 3: Visualization of masks used in editing.

Although the techniques introduced in the preliminary section provide a strong starting point, they often suffer from fine-grained inconsistencies, primarily due to their inability to selectively focus on task-relevant regions during different subtasks. This attention dispersion issue intensifies in complex multi-object scenarios where multiple visually similar objects co-occur. We refer the readers to our ablations in Sec. 4.3 for more vivid examples.

Motivated by this observation, we propose a *content-aware attention mechanism* integrated into the attention layers of our diffusion backbone. This mechanism selectively enhances attention weights toward instance-level relevant regions, thereby improving the visual consistency of the generated output—particularly in scenarios where only a subset of the image is edited.

Specifically, we first introduce the revised version of Eq. 1 where multiplicative scaling and additive biasing are included. Mathematically, we define:

$$f(\gamma, \delta, A_t(Q_t^{\text{gen}}, K_t^{\text{ori}})) = \gamma \cdot A_t(Q_t^{\text{gen}}, K_t^{\text{ori}}) + \delta \tag{3}$$

where $\gamma > 1$ amplifies the relative importance of selected query-key pairs, and $\delta > 0$ introduces a bias term to directly increase the attention weight for target regions. These two designs are complementary: multiplicative scaling maintains the original attention structure and is well-suited for scenarios with strong attention contrast, whereas additive biasing provides a more direct and linear enhancement, making it particularly effective when the initial attention weights are weak or diffuse.

We then extend Eq. 3 to a context-aware formulation. Specifically, we identify three distinct contextual regions based on the aforementioned subtasks (See Fig. 3). First, we define the mask of untouched background region as $\mathbf{m}^{\text{bg}} = \complement\left(\bigcup_{n=1}^{N} \mathbf{m}_n \cup \hat{\mathbf{m}}_n\right)$, where $\complement(\cdot)$ is the complement with respect to the full image domain. Next, we define the inpainting region for the n-th object as $\mathbf{m}_n^{\text{ipt}} = \mathbf{m}_n \setminus \bigcup_{n=1}^{N} \hat{\mathbf{m}}_n$ representing the part of the original object location that does not overlap with any of the relocated targets. Lastly, we denote $\mathbf{m}_n^{\text{ref}}$ as the reference region that provides contextual guidance for inpainting following the relocation of n-th object. In practice, this region can be obtained from background segments identified by an off-the-shelf segmentation tool. Our final formulation is then defined as:

$$A_t^*(Q_t^{\text{gen}}, K_t^{\text{ori}}) = \begin{cases} f(\gamma_1, \delta_1, A_t(Q_t^{\text{gen}}, K_t^{\text{ori}})), & \text{if } Q_t^{\text{gen}} \in \hat{\mathbf{m}}_n \wedge K_t^{\text{ori}} \in \mathbf{m}_n, \\ f(\gamma_2, \delta_2, A_t(Q_t^{\text{gen}}, K_t^{\text{ori}})), & \text{if } Q_t^{\text{gen}} \in \mathbf{m}_n^{\text{ipt}} \wedge K_t^{\text{ori}} \in \mathbf{m}_n^{\text{ref}}, \\ f(\gamma_3, \delta_3, A_t(Q_t^{\text{gen}}, K_t^{\text{ori}})), & \text{if } Q_t^{\text{gen}} \in \mathbf{m}^{\text{bg}} \wedge K_t^{\text{ori}} \in \mathbf{m}^{\text{bg}}, \\ A_t(Q_t^{\text{gen}}, K_t^{\text{ori}}), & \text{otherwise.} \end{cases} \tag{4}$$

where the conditions hold for all $n \in \{1, \ldots, N\}$.

After applying the proposed modifications, we update the attention output in Eq. 2 by replacing $A_t(Q_t^{\text{gen}}, K_t^{\text{ori}})$ with the context-enhanced attention $A_t^*(Q_t^{\text{gen}}, K_t^{\text{ori}})$, followed by the KV replacement step. This context-aware attention mechanism explicitly guides SIMON to concentrate on important regions, thereby enhancing the fidelity and consistency of the generated results—particularly in scenarios that require precise alignment between the editing constraints and the visual output.

3.1 Multi-object Energy Guidance

Starting from $\mathbf{z}_T$, which is the Gaussian latent obtained via DDIM inversion, the standard denoising processes in diffusion models will approximately reconstruct the original image $\mathbf{x}_0$ if no external modifications are introduced. Incorporating gradient guidance derived from an energy function aligned with an editing objective has proven effective for steering the generation path toward its desired outcome [3]. However, existing energy functions are predominantly designed for single-object manipulation [24, 25], leaving the challenge of handling multiple objects largely unaddressed. A straightforward extension—sequentially applying single-object energy functions to each target object—may seem plausible, but it suffers from significant drawbacks such as error accumulation and limited capacity to handle complex edits, e.g., object shuffling (See Fig. 1).

To address this, we introduce a composite energy function $\mathbf{E}$ in SIMON for simultaneous multi-object navigation. Specifically, it comprises of three terms—$\mathbf{E}_{\text{ipt}}$, $\mathbf{E}_{\text{relo}}$, and $\mathbf{E}_{\text{bg}}$—which correspond to the subtasks of inpainting, object relocalization, and background preservation. Each term is designed to enforce visual consistency within its designated region. Inspired by [25], we define a similarity function $g(a, b)$ to quantify the visual consistency between inputs a and b. This function serves as the foundation for constructing all energy terms, each of which is formulated based on appropriate pairs of inputs evaluated through g.

The objective of inpainting is to reconstruct the regions vacated by object relocation using reference appearance cues. Motivated by this, we define the inpainting energy term $\mathbf{E}_{\text{ipt}}$ as:

$$\mathbf{E}_{\text{ipt}} = \sum_{n=1}^{N} \frac{1}{\alpha + \beta \cdot g(F_t^{\text{gen}}[\mathbf{m}_n^{\text{ipt}}], F_t^{\text{ori}}[\mathbf{m}_n^{\text{ref}}])} \tag{5}$$

where α and β are hyper-parameters. F_t^{gen} denotes the feature extracted from the activations in generated latent $\mathbf{z}_t^{\text{gen}}$, which is produced by the UNet denoiser at

time step t. Similarly, F_t^{ori} is extracted from the cached latent $\mathbf{z}_t^{\text{ori}}$. The notation $F_t^{\text{gen}}[\mathbf{m}_n^{\text{ipt}}]$ indicates the subset of F_t^{gen} features within the inpainting mask $\mathbf{m}_n^{\text{ipt}}$.

Similarly, we define the remaining two terms as:

$$\mathbf{E}_{\text{relo}} = \sum_{n=1}^{N} \frac{1}{\alpha + \beta \cdot g(F_t^{\text{gen}}[\hat{\mathbf{m}}_n], F_t^{\text{ori}}[\mathbf{m}_{\text{n}}])} \tag{6}$$

$$\mathbf{E}_{\text{bg}} = \frac{1}{\alpha + \beta \cdot g(F_t^{\text{gen}}[\mathbf{m}^{\text{bg}}], F_t^{\text{ori}}[\mathbf{m}^{\text{bg}}])} \tag{7}$$

Clearly, both terms are designed to reflect the intuition that the generated features should closely resemble the original features within their corresponding regions—whether transferring from individual $\hat{\mathbf{m}}_n$ to $\mathbf{m}_n$, or remaining within the background region $\mathbf{m}^{\text{bg}}$.

To further enhance spatial disentanglement, we apply a spatially-aware weighted combination of the three energy terms, where each term is selectively enforced within its designated region:

$$\mathbf{E} = \mathbf{m}^{\text{bg}} \cdot w_b \mathbf{E}_{\text{bg}} + (1 - \mathbf{m}^{\text{bg}}) \cdot (w_r \mathbf{E}_{\text{relo}} + \mathbf{E}_{\text{ipt}}) \tag{8}$$

where w_b and w_r are hyperparameters that control the relative importance of each energy term. This region-specific application prevents interference between unrelated areas and encourages localized consistency in the generation process. By updating according to $\mathbf{z}_{\text{t-1}} = \mathbf{z}_t^{\text{gen}} - \eta \nabla_{\mathbf{z}_t^{\text{gen}}} \mathbf{E}$, we effectively steer the generation process toward achieving our ultimate goal of complex multi-object manipulation. Again, our design simultaneously constrains the positional relationships among all objects, enabling synchronized multi-object editing and substantially enhancing editing quality.

3.2 Latents Initialization

Although $\mathbf{E}_{\text{ipt}}$ is designed to enforce visual consistency with the reference mask $\mathbf{m}_n^{\text{ref}}$, we still observe residual artifacts from the original image in the inpainting regions, as in Fig. 1. These issues become more pronounced in the presence of complex backgrounds, subtle object boundaries, or imprecise reference masks, resulting in visible artifacts or incomplete replacements in the generated image. To address this, we propose a simple yet effective strategy: explicitly eliminating the influence of cached latents within the masked inpainting regions. Specifically, after DDIM inversion to obtain $\mathbf{z}_T$, we overwrite the region specified by the inpainting mask $\mathbf{m}_n^{\text{ipt}}$ with Gaussian noise sampled from a standard normal distribution, or $\mathbf{z}_T[\mathbf{m}_n^{\text{ipt}}] \sim \mathcal{N}(0, 1)$. Then we reuse the cached latent $\mathbf{z}_t^{\text{ori}}$ from the DDIM inversion trajectory to replace the background region $\mathbf{m}^{\text{bg}}$ during the first k denoising steps. Mathematicall, we have $\mathbf{z}_t^{\text{gen}}[\mathbf{m}^{\text{bg}}] = \mathbf{z}_t^{\text{ori}}[\mathbf{m}^{\text{bg}}], \text{if } t < k$.

This design ensures that no latent traces from the original image remain in the masked regions, effectively eliminating structural remnants and harnessing the generative prior for content reconstruction. This prior is further reinforced

Table 1: Quantitative results on multi-object navigation. Noticeably, SIMON consistently outperforms training-free baselines and shows superior perceptual quality and identity fidelity even compared to the training-based counterparts such as MagicFixup, while maintaining much higher inference efficiency.

Method	Background Consistency		Navigation Consistency		Identity Fidelity	Image Quality
	MSE $\times 10^3\downarrow$	LPIPS $\times 10^{-1}\downarrow$	MSE $\times 10^3\downarrow$	LPIPS $\times 10^{-1}\downarrow$	DINO↑	FID ↓
Training-based methods						
MagicFixup	0.340	<u>1.422</u>	**1.337**	<u>4.925</u>	0.772	**21.964**
Insert+LaMA	<u>0.216</u>	1.523	2.723	7.661	<u>0.867</u>	30.238
Training-free methods						
DragonDiff	0.539	2.163	1.906	5.543	0.836	30.977
DiffEditor	0.384	1.856	1.721	5.305	0.849	30.193
GeoDiffuser	0.649	2.353	2.237	7.120	0.797	30.633
Ours	**0.209**	**0.969**	<u>1.649</u>	**4.531**	**0.869**	<u>22.826</u>

by the background replacement strategy: by preserving structural background information during the early denoising steps, SIMON gains stronger guidance for reconstructing noisy regions, yielding more visually coherent outputs.

4 Experiments

In this section, we present a comprehensive overview of our experimental setup. We then detail main results on multi-object manipulation, providing quantitative metrics for each subtask. Finally, we include extensive ablation studies and qualitative examples showcasing broader applications of ours. The limitations of our method are discussed in the Appendix.

4.1 Experimental Setup

Datasets curation. We prepare a new benchmark from the 3D-FUTURE dataset [5] by selecting test scenes containing between two and five furniture items and manually annotating target positions for each object, resulting in 1000 images and a total of 3405 objects (see Fig. 1). Compared to datasets like COCO [20], 3D-FUTURE offers more consistent object counts per image and spatially coherent layouts that reflect realistic indoor arrangements. These properties make it particularly suitable for multi-object manipulation, reducing the risk of generating physically implausible outcomes such as unnatural overlaps or misaligned placements. All images in our benchmark are used exclusively for testing. Hyperparameters are calibrated on a small disjoint calibration set, as detailed in Appendix. Notably, our annotations ensure that the resulting layouts are physically plausible and semantically meaningful. Additionally, we generate text prompts based on object counts in each image using the format: *"A photo of a room with* cnt_1 *class_*$name_1$*,* cnt_2 *class_*$name_2$*, ..."*, where class names are taken from the original 3D-FUTURE annotations. Further details are in the Appendix.

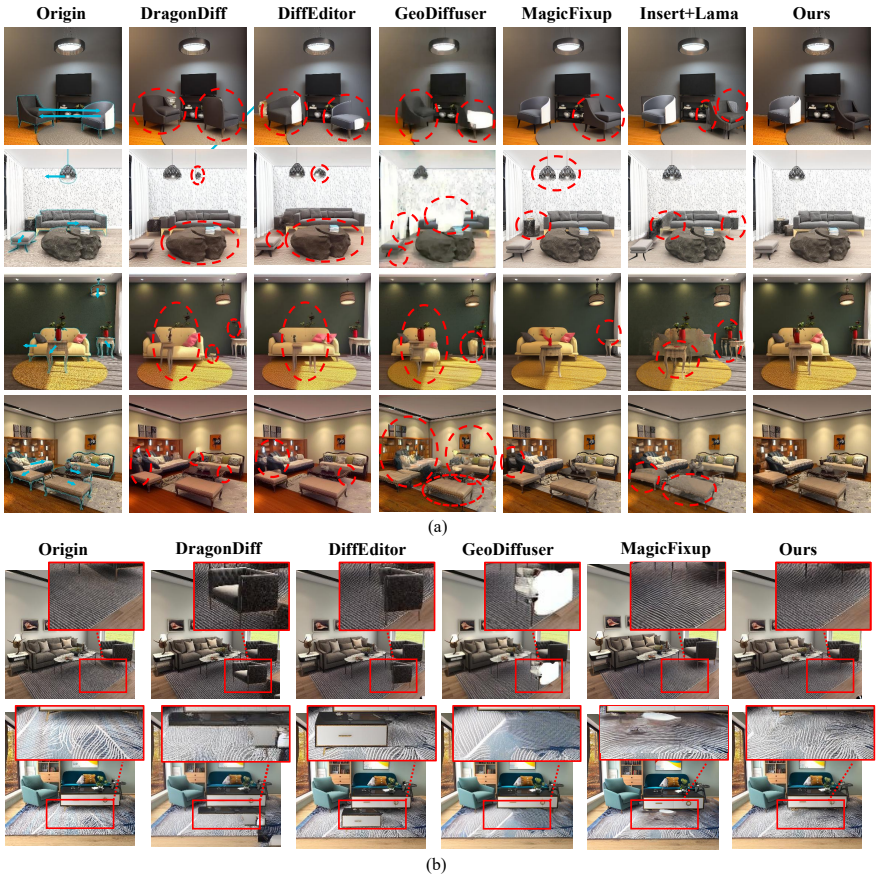

Fig. 4: Visual examples of both multi-object navigation (a) and single-object inpainting (b). Target objects and regions of interest are highlighted in blue and red, respectively. More examples in outdoor scenarios are provided in Appendix.

Existing methods. We evaluate SIMON against five SOTA baselines, categorized by their training requirements. **Training-free methods** include DragonDiff [25], DiffEditor [24], and GeoDiffuser [34]; although designed for single-object drag, we extend them to multi-object tasks via sequential editing. **Training-based methods** comprise MagicFixup [1], which natively supports multi-instance editing but necessitates massive pre-training on 5 million images. Specifically for the multi-object navigation task, we introduce a composite two-stage pipeline which simulates the movement effect by combining LaMa [39] for hole-filling and Insert [38] for instance insertion. These baselines are selected to ensure a comprehensive evaluation of editing quality and resource efficiency.

Evaluation metrics. We design two primary sets of experiments on our benchmark to quantitatively evaluate the three subtasks—inpainting, relocation, and

Table 2: Results on inpainting subtask. A high-quality inpainting result should simultaneously exhibit high residual suppression (effective object removal) and low reconstruction fidelity (high scene coherence). SIMON achieves the best balance between these two competing objectives. "w/o LI" is the variant without latent initialization.

Method	**Residual Suppression**		**Reconstruction Fidelity**	
	MSE $_{\times 10^3}\uparrow$	LPIPS↑	MSE $_{\times 10^3}\downarrow$	LPIPS↓
Training-based methods				
MagicFixup	4.737	1.182	2.940	1.231
Training-free methods				
DragonDiff	3.431	1.173	<u>2.668</u>	<u>0.761</u>
DiffEditor	3.774	1.129	2.881	0.801
GeoDiffuser	**9.488**	**1.847**	6.524	1.514
Ours	<u>6.106</u>	<u>1.435</u>	**1.964**	**0.670**
w/o LI	3.319	0.996	3.155	0.991

background preservation—at instance, regional and image levels. The first set focuses on multi-object navigation, where we evaluate performance using four metrics groups: identity fidelity, navigation consistency, background consistency, and image quality. Specifically, identity fidelity measures the preservation of instance-level identities using DINOv2 [27]. Meanwhile, background and navigation consistency measure regional fidelity by assessing appearance similarity with MSE and LPIPS [45]. Finally, we assess image quality using standard FID [10], which measures visual realism of generated images[4]. Our second set of experiments focuses on inpainting. For each original image in our benchmark, we select suitable annotated objects and paste it to a new annotated location, forming a modified image. Our goal of this task is to remove the inserted object and restore the original. Here we introduce two complementary objectives for evaluation based on the appearance similarity within the inpainting region: residual suppression, which compares the generated image to the modified image, and reconstruction fidelity, which compares it to the original image. Again, we use MSE and LPIPS for both. More details on selection criteria, metric design, and human analysis are in the Appendix.

Implementation details. To perform fair comparisons with existing methods, we use SD v1.5 [32] as our generative backbone and perform 50-step denoising using the DDIM sampler. In practice, we further extend SIMON to transformer-based architectures, such as Flux [18] and we refer the readers to the Appendix for visual examples.

[4] CLIPScore [30] is not included in our main paper as the original unedited images—arguably the best references—receive surprisingly low values, but we include it in the Appendix for completeness

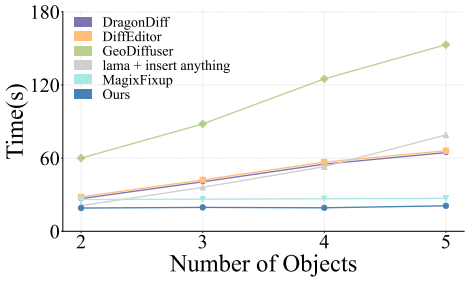

Fig. 5: Runtime efficiency w.r.t. various number of target objects.

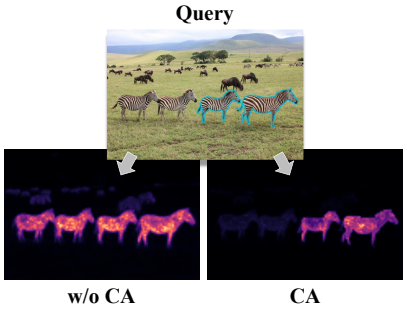

Fig. 6: With our CA mechanism, attention distribution becomes more concentrated on given targets (outlined in blue) rather than spreading out over all subjects that are visually similar.

4.2 Main Result

Overall performance. SIMON consistently outperforms all training-free counterparts and the optimized two-stage pipeline (159K pairs) across nearly all metrics. Notably, SIMON achieves the highest DINOv2 score and superior background consistency, demonstrating its robust structural preservation. Remarkably, even compared to the computationally-heavy MagicFixup (pre-trained on 5M images), our training-free framework yields superior background consistency and identity fidelity, bridging the gap between inference efficiency and high-end generative quality, achieving a state-of-the-art balance without the need for any costly optimization or large-scale data dependencies. For the inpainting task in Tab. 2, SIMON again demonstrates its advantage by achieving an optimal balance between residual suppression and reconstruction fidelity. While GeoDiffuser shows high residual suppression at the cost of severe distortion, SIMON maintains the lowest reconstruction error while delivering nearly complete object removal, significantly surpassing MagicFixup and others in overall coherence.

4.3 Content-aware Attention

We further provide *qualitative comparisons* in Fig. 4, where multi-object navigation and single-object inpainting are shown on the left and right, respectively. For the multi-object examples, we present four cases with an growing number of target objects from left to right. As expected, sequential methods such as DragonDiff and DiffEditor often struggle with complex layout rearrangements—particularly object shuffling—where early edits lead to overlaps that disrupt subsequent steps (see first row in (a)). While Geodiffuser sometimes introduces substantial inconsistencies and structural distortion. As the number of manipulated objects increases, these methods exhibit more severe distortions

Table 3: Ablation study on the effectiveness of two modules. "EG" and "CA" refer to multi-object energy guidance and content-aware attention respectively.

Method	Background Consistency		Navigation Consistency		Image Quality
	MSE $\times 10^3\downarrow$	LPIPS $\times 10^{-1}\downarrow$	MSE $\times 10^3\downarrow$	LPIPS $\times 10^{-1}\downarrow$	FID$\downarrow$
Baseline	0.387	1.757	1.768	5.229	28.403
+EG	0.235	1.017	1.693	4.752	23.212
+EG+CA	0.246	1.008	1.658	**4.525**	**22.786**
Ours (Full)	**0.209**	**0.969**	**1.649**	4.531	22.826

and reduced global consistency (see fourth row in (a)). In contrast, our method performs simultaneous navigation, avoiding sequential interference and maintaining stronger structural integrity and visual coherence, even in challenging scenes. For the inpainting task, existing methods often leave visible residuals or introduce texture inconsistencies, especially when dealing with objects of complex shapes or intricate backgrounds (see second row in (b)). In contrast, our method effectively suppresses such artifacts and generates more natural, semantically coherent results. Additional visual examples, such as single-object editing scenarios, are provided in the Appendix.

Efficiency and human analysis. We further report the runtime efficiency of all methods under fixed image resolution of 768×768 as the number of moved objects increases in Fig. 5. Unsurprisingly, existing methods exhibit linear growth in runtime due to stepwise object processing. In contrast, ours performs multi-object navigation in a single forward process, leading to substantially lower and more stable inference time. We refer the readers to the Appendix for all necessary details about human analysis.

Ablation studies. To assess the effectiveness of each module, we conduct extensive ablation studies and present the results in Tab.3 and Tab.2. The results in the first table show that integrating Energy Guidance (EG) and Content-aware Attention (CA) boosts all metrics in the multi-object navigation task, respectively—highlighting the benefits of simultaneous editing and spatially adaptive attention modulation. Interestingly, adding Latent Initialization (LI) ("Ours") yields improvements in background and navigation consistency but causes slight drops in image quality, likely due to the injected noise slightly disrupting editing precision. Nonetheless, the positive effect of LI is evident in Tab. 2, where removing it sharply degrades performance versus the full model ("Ours"). Details of baselines and examples of each module are provided in the Appendix. Last, but by no means least, as illustrated in Fig. 6, a naive model tends to distribute attention uniformly across all zebra instances, thereby weakening instance-specific consistency during object manipulation.

More applications. Although SIMON is primarily designed for multi-object navigation, its versatility extends to a wide range of image editing tasks. We showcase its applicability on multi-object appearance replacement, resizing, and

Fig. 7: Visualization of different applications. From left to right: multi-object appearance replacement, object resizing, and spatial shuffling. Our method demonstrates high visual consistency and structural integrity even under complex editing requests.

shuffling, with visual results presented in Fig. 7. Implementation details and more examples are available in the Appendix.

Limitations and future work While our method produces geometrically and semantically plausible edits in a wide range of cases, it still struggles with faithfully modeling illumination and occlusion handling.

As for the former, since the editable region is localized by a binary mask, illumination and cast shadows, which typically extend beyond the mask and depend on global scene context, may appear locally plausible but globally inconsistent. As shown in the Appendix, the water surface fails to fully capture the reflection of the target duck, revealing a mismatch with the surrounding environment. We observe similar failure modes in other training-free methods such as DragonDiffusion, DiffEditor, and GeoDiffuser, indicating that maintaining illumination consistency remains an open challenge. Inspired by [44], we believe incorporating physically-motivated lighting constraints into the energy term to guide latent updates is a promising direction for future work.

5 Conclusion

We present a novel framework, SIMON, designed for simultaneous multi-object navigation. Our approach integrates three key components: content-aware attention, multi-object energy guidance, and latent initialization. Extensive experiments on the 3D-FUTURE benchmark demonstrate that SIMON consistently outperforms SOTA methods across inpainting, relocation, and background preservation tasks, achieving superior performance on both regional and global metrics, while offering significantly improved runtime efficiency and human favorable results.

6 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62406281, U24B20174), Guangdong Provincial Natural Science Foundation (No. 2026A1515011580), Shenzhen Key Laboratory of General-purpose Embodied Intelligent Robotics (SYSRD20250529114406009), and Shenzhen High Level Talent (No. ZK1140900125, BA1140901225).

References

1. Alzayer, H., Xia, Z., Zhang, X., Shechtman, E., Huang, J.B., Gharbi, M.: Magic fixup: Streamlining photo editing by watching dynamic videos. ACM Transactions on Graphics **44**(5), 1–25 (2025) 3, 4, 11
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5855–5864 (2021) 1
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 4, 6, 8
4. Duan, Z.P., Zhang, J., Liu, S., Lin, Z., Guo, C.L., Zou, D., Ren, J., Li, C.: A diffusion-based framework for occluded object movement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2816–2824 (2025) 4
5. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. arXiv preprint arXiv:2009.09633 (2020) 3, 10
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2414–2423 (2016) 1
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014) 1, 2, 3
8. Han, L., Wen, S., Chen, Q., Zhang, Z., Song, K., Ren, M., Gao, R., Stathopoulos, A., He, X., Chen, Y., et al.: Proxedit: Improving tuning-free real image editing with proximal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4291–4301 (2024) 5
9. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. In: International Conference on Learning Representations (2022) 1
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. vol. 30 (2017) 12
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 1, 4
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 3
13. Jiang, R., Fu, X., Zheng, G., Li, T., Yao, T., Li, X.: Energy-guided optimization for personalized image editing with pretrained text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4048–4056 (2025) 5
14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018) 2, 4
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 1
16. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2426–2435 (June 2022) 6

17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023) 5
18. Labs, B.F.: Flux (2024) 6, 12
19. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023) 4
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014) 10
21. Liu, C., Li, X., Ding, H.: Referring image editing: Object-level image editing via referring expressions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13128–13138 (2024) 1
22. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022) 5
23. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 1
24. Mou, C., Wang, X., Song, J., Shan, Y., Zhang, J.: Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8488–8497 (2024) 2, 3, 4, 8, 11
25. Mou, C., Wang, X., Song, J., Shan, Y., Zhang, J.: Dragondiffusion: Enabling drag-style manipulation on diffusion models. In: Proceedings of the International Conference on Learning Representations (2024) 2, 3, 4, 5, 8, 11
26. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. pp. 16784–16804. PMLR (2022) 4
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 12
28. Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., Theobalt, C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023) 2, 3
29. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016) 1
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 12
31. Ren, J., Xu, M., Wu, J.C., Liu, Z., Xiang, T., Toisoul, A.: Move anything with layered scene diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6380–6389 (2024) 4

32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1, 6, 12
33. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023) 4
34. Sajnani, R., Vanbaar, J., Min, J., Katyal, K., Sridhar, S.: Geodiffuser: Geometry-based image editing with diffusion models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 472–482. IEEE (2025) 3, 4, 11
35. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022) 2
36. Shi, Y., Xue, C., Liew, J.H., Pan, J., Yan, H., Zhang, W., Tan, V.Y., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8839–8849 (2024) 5, 6
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 5, 6
38. Song, W., Jiang, H., Yang, Z., Quan, R., Yang, Y.: Insert anything: Image insertion via in-context editing in dit. arXiv preprint arXiv:2504.15009 (2025) 3, 11
39. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2149–2159 (2022) 11
40. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6232–6242 (2024) 1
41. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023) 4
42. Yu, X., Wang, T., Kim, S.Y., Guerrero, P., Chen, X., Liu, Q., Lin, Z., Qi, X.: Objectmover: Generative object movement with video prior. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17682–17691 (2025) 4
43. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023) 4
44. Zhang, L., Rao, A., Agrawala, M.: Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In: The Thirteenth International Conference on Learning Representations (2025) 15
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 12
46. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22490–22499 (2023) 4
47. Zhou, S., Xiao, T., Yang, Y., Feng, D., He, Q., He, W.: Genegan: Learning object transfiguration and attribute subspace from unpaired data. In: Proceedings of the British Machine Vision Conference (BMVC) (2017) 3