Uniform Attention Maps: Boosting Image Fidelity in Reconstruction and Editing

Wenyi Mo[1,2], Tianyu Zhang[3], Yalong Bai[3], Bing Su[1,2†], Ji-Rong Wen[1,2]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods
[3]Du Xiaoman Technology
[1]{mowenyi, jrwen}@ruc.edu.cn, subingats@gmail.com
[2]{zhangtianyu, libai}@duxiaoman.com

Abstract

Text-guided image generation and editing using diffusion models have achieved remarkable advancements. Among these, tuning-free methods have gained attention for their ability to perform edits without extensive model adjustments, offering simplicity and efficiency. However, existing tuning-free approaches often struggle with balancing fidelity and editing precision, particularly due to the influence of cross-attention in DDIM inversion, which introduces reconstruction errors. To address this, we analyze reconstruction from a structural perspective and propose a novel approach that replaces traditional cross-attention with uniform attention maps, significantly enhancing image reconstruction fidelity. Our method effectively minimizes distortions caused by varying text conditions during noise prediction. To complement this improvement, we introduce an adaptive mask-guided editing technique that integrates seamlessly with our reconstruction approach, ensuring consistency and accuracy in editing tasks. Experimental results demonstrate that our approach not only excels in achieving high-fidelity image reconstruction but also performs robustly in real image composition and editing scenarios. This study underscores the potential of uniform attention maps to enhance the fidelity and versatility of diffusion-based image processing methods. Code is available at https://github.com/Mowenyii/Uniform-Attention-Maps.

1. Introduction

In recent years, the field of image processing has seen significant advancements, particularly with the development of Denoising Diffusion Probabilistic Models (DDPMs) [4, 11, 27, 28]. These models have revolutionized image composition and editing by enabling more precise and creative control over images [10, 21]. One of the key innovations has been the introduction of tuning-free methods, which allow for effective editing without the need for extensive model adjustments. These methods offer simplicity and efficiency by manipulating latent vectors during the denoising process, unlocking new possibilities for accurate image editing. However, applying these tuning-free techniques to real-world images presents challenges. In practice, the latent vectors of real images are often unknown, making it difficult to directly apply these methods, which limits their practical use.

To overcome this, researchers have developed inversion

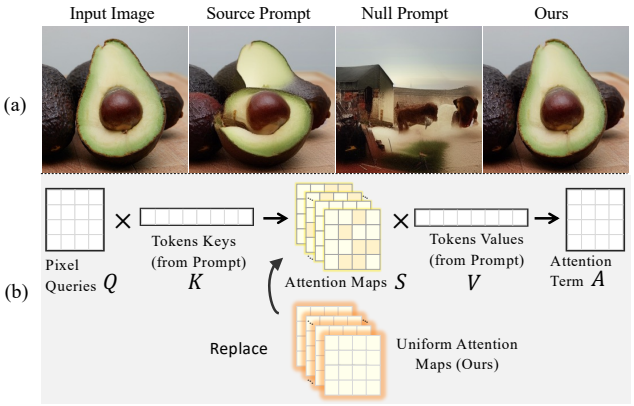

Figure 1. **(a)** Image reconstruction using DDIM with different prompts. The first image shows the input image, followed by the reconstruction using the source prompt "a photo of avocados," the null prompt (an empty string), and the result using Uniform Attention Maps combined with token values from the null prompt. **(b)** Our approach introduces Uniform Attention Maps, where traditional attention maps are replaced with uniform maps that distribute attention weights equally across the token dimension. By combining these uniform maps with the value tokens V, we generate a more balanced attention term A. This method ensures consistent attention, resulting in more accurate image reconstructions, as demonstrated in the final image of part (a).

†Corresponding authors.

Method	Base Model	Structure ↓ Distance$_{\times 10^3}$	PSNR↑	LPIPS$_{\times 10^3}$↓	MSE$_{\times 10^4}$↓	SSIM$_{\times 10^2}$↑
Upper Bound	VQAE [8]	2.39	28.58	34.20	21.57	82.04
Null Prompt	SD 1.4	15.31	22.88	124.35	69.60	72.18
Source Prompt	SD 1.4	11.31	23.89	101.47	55.43	74.45
Zero Cross-Attention Maps	SD 1.4	11.13	24.36	102.83	51.17	74.97
TF-ICON [21]	SD 1.4	5.51	25.57	64.12	37.34	77.70
Uniform Attention Maps (Null)	SD 1.4	4.76	**26.97**	57.29	**28.98**	79.29
Uniform Attention Maps (Src)	SD 1.4	**4.67**	26.96	**54.17**	29.05	**79.33**

Table 1. Reconstruction performance on the PIE benchmark [14] using DDIM Inversion with 20 timesteps under various conditions without CFG. Our method, Uniform Attention Maps, achieves higher fidelity to the original image than others. Additionally, the reconstruction results using token values from source and null prompts are similar, demonstrating the robustness of our approach across different prompts.

methods like Denoising Diffusion Implicit Models (DDIM) Inversion [29], which map images back to their noisy latent vectors using a trained diffusion model. This approach has been particularly effective for unconditional diffusion models. Additionally, recent advances in text-conditioned DDIM inversion [9, 14, 23, 25] have further improved image editing by incorporating classifier-free guidance (CFG) [12] during the generation and editing stages. These enhancements have led to more effective edits, but challenges remain. Current methods still struggle to balance preserving the original image details with making user-defined changes.

Existing methods [3,14,23,25] typically use a dual-branch approach after inverting input images, separating the process into reconstruction (source) and editing (target) branches. While this approach has yielded impressive results, it also introduces challenges, such as discrepancies between noise predictions in the inversion and reconstruction phases in the reconstruction branch, which can lead to the loss of important image details [33]. Various strategies have been proposed to address these issues. Some approaches, like Null-text Inversion [25], use optimization techniques to minimize the distance between the representations between the reconstruction and inversion phases. On the other hand, methods like Proximal Guidance [9] improve reconstruction effectiveness by introducing an extra regularization term, without extensive tuning. Despite these advancements, the reconstruction effectiveness varies significantly with different prompts. As shown in Fig. 1 (a), reconstruction outcomes can differ significantly based on these conditions. This leads us to the core questions of our research: Given the assumption of DDIM inversion with adjacent noise prediction approximation, why do different conditions lead to varied reconstruction outcomes? How can we improve image reconstruction effectiveness in text-conditioned scenarios?

To address these questions, our study focuses on the cross-attention mechanism within the U-Net architecture used in diffusion models. We are the first to analyze DDIM inversion and reconstruction under text-conditioned settings from a structural perspective. Our findings reveal that cross-attention plays a pivotal role in the reconstruction errors observed in current methods. To address this, we propose an improved image reconstruction method that leverages uniform cross-attention to enhance the effectiveness of text-conditioned image reconstruction and composition. Additionally, we introduce an automatic mask generation technique to improve the performance of existing image editing algorithms, making our approach more robust and applicable to a wider range of scenarios.

Our contributions are threefold: (1) We provide a detailed analysis of how cross-attention impacts image reconstruction, (2) We propose an enhanced reconstruction method that shows superior performance in both image composition and editing tasks, and (3) We develop an automatic mask generation technique that significantly improves the accuracy and effectiveness of image editing. Through these innovations, we aim to advance image processing, offering new tools and methods that can be easily adopted in practical applications.

2. Related work

In recent years, significant advancements have been made in the field of text-guided vision tasks, encompassing areas such as vision-language inference [6, 18, 26, 30], text-to-image generation [7, 24, 27, 28], and image editing [3, 10, 14, 23, 25]. While our focus in this paper is on text-conditioned image editing with diffusion-based models, these works highlight the broader importance of effective text guidance in vision-related tasks. The biggest challenge in this task is how to achieve the intention of the guiding texts while ensuring fidelity to the input image. Previous works can be categorized as end-to-end editing models, tuning-based methods, attention-based methods, and sample-based methods. (a) End-to-End Editing Model: Methods like InstructPix2Pix [2] and DiffusionCLIP [17] fine-tune pre-trained text-to-image models to revise images based on simple instructions, allowing for efficient and quick edits without per-example fine-tuning or inversion. (b) Tuning-based methods: Tuning-based methods involve training a set

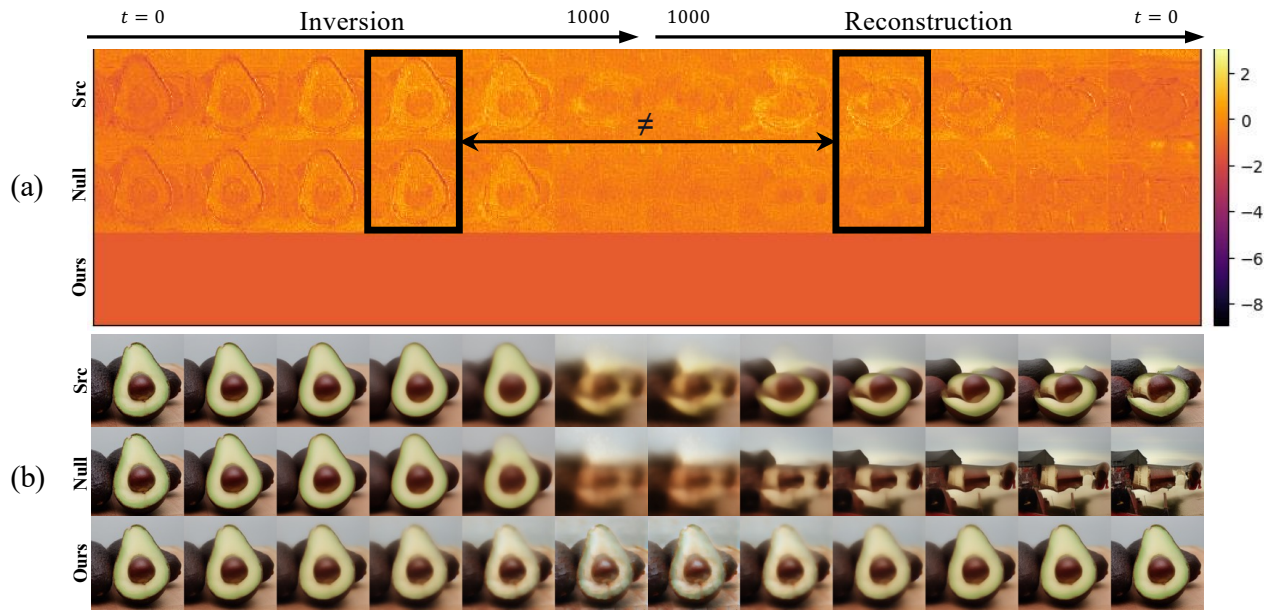

Figure 2. The process of reconstruction using DDIM inversion under various conditions. It visually depicting (a) the cross-attention term $A^{(l)}$ from the U-Net model's layers with output dimensions of 64×64 and (b) the predicted latent representation $\hat{z}_0$ at different stages of both the inversion and reconstruction processes. In (a), discrepancies in the cross-attention maps between the inversion and reconstruction phases are evident, with misalignment causing errors in image fidelity under the source and null prompt conditions. In (b), the reconstructed images show significant distortions under the source and null conditions, whereas our method consistently maintains high image quality throughout the reconstruction process.

of learnable parameters or fine-tuning a model to encapsulate certain concepts. Methods such as Imagic [16] and Unitune [32] specifically fine-tune the model on the input image to achieve high fidelity. These methods are time-consuming and the misalignment of learned variables with the diffusion model's expected input distribution compromises the integrity and quality of edits, limiting their practical use in fast-processing and high-fidelity applications [14]. (c) Attention-based methods: Attention mechanisms allow models to "focus" on specific parts of an image, making it possible to edit certain areas or aspects without affecting the entire image. These methods improve precision, context awareness, and efficiency of image editing, enabling more complex edits. For instance, Prompt-to-Prompt [10] and MasaCtrl [3] focus on integrating attention mechanisms to ensure that edits are contextually aware and maintain the essence of the input image. Our method can be combined with them to help achieve better reconstruction results and enhance editing efficiency. (d) Sample-based methods: Methods like Null-text Inversion [25], Negative-prompt Inversion [23], Proximal Guidance [9], Direct Inversion [14], EDICT [33], and Edit Friendly DDPM [13] focus on refining the reconstruction process to improve the fidelity of the input image during editing. TF-ICON [21] demonstrates that semantically meaningful text within the input prompt introduces deviations in the diffusion process, causing a mismatch between the forward and reverse trajectories in the ODE-based sampling steps. To mitigate this issue, the concept of an "exceptional prompt" is introduced, involving the use of a selected token to stabilize the diffusion process and achieve better image reconstruction. However, this method often struggles to generalize across different generative models due to inherent differences in their architectures, particularly in text encoders. DiffEdit [5] utilize differences in noise predictions to create masks for faithful image editing. We also use masks during editing. The proposed adaptive masks change with each timestep to better coordinate with our proposed reconstruction method and achieve superior editing performance.

3. Method

In this section, we investigate the underlying causes of reconstruction errors associated with different prompts and propose a method to improve reconstruction by reducing the impact of the cross-attention term. We then introduce an automatic mask generation technique that integrates this method into existing image editing algorithms.

3.1. Preliminaries

DDIM Inversion. Denoising Diffusion Implicit Models (DDIMs) [29] are an extension of Denoising Diffusion Probabilistic Models (DDPMs) [4, 11, 27, 28], designed to offer

a deterministic sampling process. The reverse process in DDIM can be described as follows:

$$z_{t-1} = \sqrt{1-\alpha_{t-1}} \cdot \boldsymbol{\epsilon}_\theta(z_t, t) + \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{z_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(z_t, t)}{\sqrt{\alpha_t}} \right)}_{\text{“ predicted } \hat{z}_{0,t} \text{ ”}}, \quad (1)$$

where z_{t-1} represents the latent vector at the previous timestep, derived from z_t at the current timestep. $\hat{z}_{0,t}$ denotes the estimated clean image at timestep t. The parameters α_t are derived from the forward diffusion process, and the function $\boldsymbol{\epsilon}_\theta(z_t, t)$ estimates the noise at each timestep. To make this process more practical for image editing, we can rearrange Eq. (1) as:

$$z_t = \sqrt{1-\alpha_t} \cdot \boldsymbol{\epsilon}_\theta(z_t, t) + \sqrt{\alpha_t} \left(\frac{z_{t-1} - \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_\theta(z_t, t)}{\sqrt{\alpha_{t-1}}} \right). \quad (2)$$

When applying this model to real images, the goal is to obtain the initial noise vector z_T from a given image representation z_0 as the starting point for further editing. However, directly computing z_t requires the noise prediction $\boldsymbol{\epsilon}_\theta(z_t, t)$, which is not always accessible. Therefore, during the inversion process, an approximation is made by using the noise prediction from the previous timestep $\boldsymbol{\epsilon}_\theta(z_{t-1}, t-1)$ [33]. This approach results in a sequence of latent variables, $\{z_t^*\}_{t=1}^T$, that traces back through the diffusion process:

$$z_t^* = \sqrt{1-\alpha_t} \cdot \boldsymbol{\epsilon}_\theta(z_{t-1}^*, t-1) + \sqrt{\alpha_t} \underbrace{\left(\frac{z_{t-1}^* - \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_\theta(z_{t-1}^*, t-1)}{\sqrt{\alpha_{t-1}}} \right)}_{\text{“ predicted } \hat{z}_{0,t} \text{ ”}}. \quad (3)$$

Cross-attention mechanism. In diffusion models implemented using U-Net, the text condition is typically incorporated through a cross-attention mechanism [27, 28]. When predicting $\boldsymbol{\epsilon}_\theta(z_t, t, \mathbf{c})$, where $\mathbf{c} \in \mathbb{R}^{N \times d_c}$ represents the input text and N is the token number of the input text, the flattened intermediate representation of the l^{th} layer of the model $\boldsymbol{\epsilon}_\theta$ at time step t, denoted as $x_t^{(l)} \in \mathbb{R}^{M^{(l)} \times d_x^{(l)}}$, is updated via cross-attention as follows:

$$\tilde{x}_t^{(l)} = x_t^{(l)} + A_t^{(l)}, \quad (4)$$

where $\tilde{x}_t^{(l)}$ is the updated representation, and $A_t^{(l)}$ represents the cross-attention term (or update term), calculated as:

$$A_t^{(l)} = S_t^{(l)} \cdot V^{(l)}, \quad (5)$$

with the score map $S_t^{(l)} \in \mathbb{R}^{M^{(l)} \times N}$ defined by:

$$S_t^{(l)} = \text{softmax}\left(\frac{Q_t^{(l)} (K^{(l)})^T}{\sqrt{d}} \right), \quad (6)$$

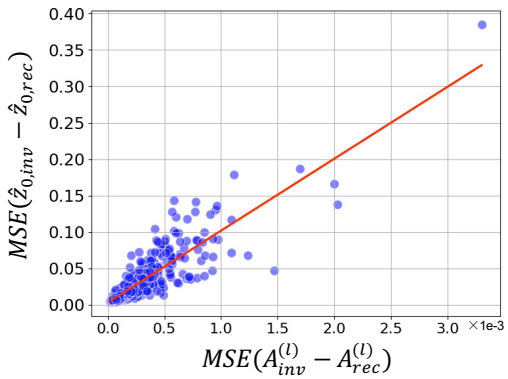

Figure 3. Correlation between MSE of cross-attention term $A^{(l)}$ and clean image prediction $\hat{z}_0$ during inversion and reconstruction. The scatter plot shows that discrepancies in the cross-attention term $A_t^{(l)}$ from all U-Net model's layers with output dimensions of 64×64 during the inversion and reconstruction phases contribute significantly to the Mean Squared Error (MSE) in the predicted clean image $\hat{z}_{0,t}$, as evidenced by the positive correlation across 700 images from the PIE benchmark [14].

where $Q_t^{(l)} \in \mathbb{R}^{M^{(l)} \times d^{(l)}}$ is the linear transformation of $x_t^{(l)}$, and $K^{(l)}, V^{(l)} \in \mathbb{R}^{N \times d^{(l)}}$ are the linear transformations of $\mathbf{c}$. Note that $K^{(l)}$ and $V^{(l)}$ are independent of the time step.

3.2. The Devil in Reconstruction: Non-uniform Cross-attention

DDIM inversion operates under the assumption that the noise predictions at adjacent timesteps, $\boldsymbol{\epsilon}_\theta(z_t, t)$ and $\boldsymbol{\epsilon}_\theta(z_{t-1}, t-1)$, are approximately equal. When conditioned on a prompt $\mathbf{c}$, this assumption breaks down due to variations in the cross-attention mechanism, which introduces discrepancies between $\boldsymbol{\epsilon}_\theta(z_t, t, \mathbf{c})$ and $\boldsymbol{\epsilon}_\theta(z_{t-1}, t-1, \mathbf{c})$. These discrepancies arise because the cross-attention term $A^{(l)}$, which integrates semantic guidance from the prompt into the intermediate latent representation, is misaligned between the inversion and reconstruction processes.

To quantify this phenomenon, we analyze 700 images from the PIE benchmark to explore the relationship between the Mean Squared Error (MSE) of the predicted clean image $\hat{z}_{0,t}$ and the cross-attention term $A_t^{(l)}$ during the inversion and reconstruction phases. Detailed experimental settings can be found in the supplementary materials. As illustrated in Fig. 3, the scatter plot highlights this relationship, with a clear positive correlation shown by the red trend line. This indicates that discrepancies in the cross-attention term $A_t^{(l)}$ contribute to errors in the reconstructed image $\hat{z}_{0,t}$.

This observation is further supported by the visualization experiments presented in Fig. 2, which track the inversion and reconstruction trajectories for an avocado example. At each timestep, we first compute the update term $A^{(l)}$ from

the U-Net model's layers, as illustrated in Fig. 2 (a). Following this, the clean predicted image $\hat{z}_{0,t}$ is generated, as shown in Fig. 2 (b). Fig. 2 (a) highlights a clear mismatch between inversion and reconstruction, particularly under source and null prompt conditions (black-boxed regions), suggesting that misalignment in the cross-attention mechanism contributes to these distortions. The observed misalignment of the update term $A^{(l)}$ across both trajectories at the same timestep in Fig. 2 suggests that cross-attention is responsible for the reconstruction errors. Experiment details can be found in the appendix.

3.3. Our solution

3.3.1 Uniform Cross-attention Maps

Experimentally, the interaction between text prompts and the model's intermediate representation using the attention mechanism introduces inconsistencies that degrade the quality of the final image reconstruction.

Building on our experiments and analyses, we propose Uniform Cross-attention Maps to enhance stability and consistency across various prompts and models. Instead of relying on traditional cross-attention maps, which vary significantly depending on the input prompt, we introduce uniform attention maps where each element is assigned a fixed value of $1/N$:

$$S^{(l)}_{uniform} = \frac{1}{N}\mathbf{1}_{M^{(l)}\times N}, \tag{7}$$

Here, $\mathbf{1}_{M^{(l)}\times N}$ denotes an $M^{(l)} \times N$ matrix with all elements equal to 1, with $M^{(l)}$ being the number of visual tokens and N the number of conditioning tokens. This uniform distribution of attention reduces the variance introduced by different text prompts, ensuring that the model's focus remains balanced across all tokens in $x^{(l)}$. As demonstrated in Fig. 1 (a), our approach effectively mitigates the deviations caused by semantic variations in text prompts, resulting in more reliable and consistent image reconstructions, as evidenced by the improved performance metrics in Tabs. 1 and 2. In contrast, Zero Cross-Attention Maps, which replace the cross-attention term $A^{(l)}$ with zeros, eliminate all semantic guidance from text prompts. While this ensures consistency, it leads to overly simplistic reconstructions and disrupts the pretraining distribution of latent features $x^{(l)}$, which were optimized to interact with cross-attention. This deviation significantly degrades the model's ability to preserve fine-grained details and complex structures. These limitations underscore the importance of Uniform Attention Maps, which not only reduce prompt variance but also maintain compatibility with the pretraining distribution to achieve high-fidelity reconstructions.

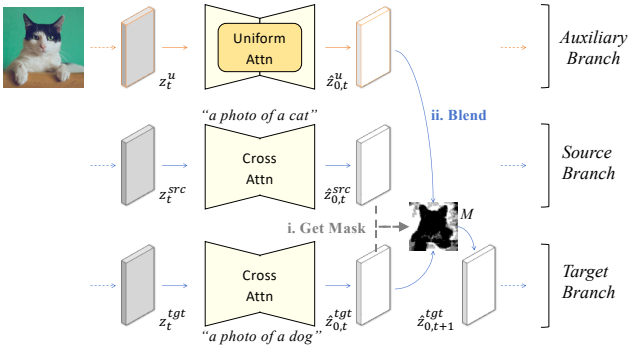

Figure 4. The proposed tuning-free image editing framework. We find that using Uniform Cross-attention Maps yields excellent reconstruction results, as shown in Tab. 1. We introduce an auxiliary branch and generate masks based on the differences between the source branch and the target branch to blend the results of the auxiliary branch. Our method effectively enhances the performance of existing image editing algorithms. The process of using Uniform attention maps is shown in Fig. 1 (b).

3.3.2 Adaptive Mask Guided Editing

The direct use of uniform attention maps in current text-driven editing pipelines presents challenges, as these pipelines typically rely on manipulating cross-attention maps to achieve precise edits. However, the exceptional reconstruction performance of uniform attention maps offers a unique opportunity to improve editing tasks. To harness this reconstructive capability, we propose a novel approach, namely adaptive mask-guided editing, which effectively utilizes the strengths of uniform attention maps in editing scenarios. The overall process is illustrated in Fig. 4.

In this method, the input image is processed through three parallel branches: the auxiliary branch, the source branch, and the target branch. The auxiliary branch, which uses a null prompt combined with uniform cross-attention maps, ensures stable reconstruction. The source branch uses the source prompt $\mathbf{c}_{src}$, while the target branch operates with the target prompt $\mathbf{c}_{tgt}$ to apply the desired edits.

To further refine this process, we introduce an adaptive mask generation technique that compares the noise predictions between the source and target branches. This comparison yields a difference, $diff_t = |\hat{z}^{tgt}_{0,t} - \hat{z}^{src}_{0,t}|$, identifying areas requiring modification. A threshold λ is then applied to this difference to create a mask M, which is subsequently refined using a dilation operation with a square kernel to handle minor inconsistencies:

$$M = dilate(|\hat{z}^{tgt}_{0,t} - \hat{z}^{src}_{0,t}| \leq \lambda).$$

After T_{mask} timesteps, this mask is employed to blend the predicted clean images $\hat{z}^{u}_{0,t}$ and $\hat{z}^{tgt}_{0,t}$ from the auxiliary and target branches, ensuring that the model preserves key details

	Method	MAE ↓	LPIPS ↓	SSIM ↑
Upper Bound	VQAE [8]	0.018	0.043	0.919
Diffusion	SD w/ CFG	0.134	0.340	0.637
	SD w/ Cond.	0.126	0.308	0.654
	SD w/ Uncond.	0.126	0.304	0.655
	TF-ICON [21]	0.019	0.047	0.918
	TF-ICON* [21]	0.021	0.045	0.834
	UAM*	**0.019**	**0.041**	**0.839**

Table 2. Quantitative comparison of image reconstruction on CelebA-HQ [15]. Additional experimental results and setting details are in [21]. Methods marked with '*' indicate results on A800.

from the original image while applying targeted edits:

$$\hat{z}_{0,t}^{tgt} = M \odot \hat{z}_{0,t}^{u} + (1 - M) \odot \hat{z}_{0,t}^{tgt}.$$

By selectively blending the clean images using the mask, the algorithm achieves a balance between maintaining the original image's fidelity and incorporating the desired modifications. This approach ensures that critical details are preserved, while the edits are seamlessly integrated into the final output. For a detailed representation of the algorithm, please refer to the pseudo-code in supplementary materials.

4. Experiments

In our experiment, for the image composition task, we follow the experimental setting and composition process of [21], using Stable Diffusion v2.1 [27] and the 20-step DPM solver sampling method [20]. We use Uniform Attention Maps (UAM) combined with token values from the target prompts in both the inversion and composition processes. For the image editing task, we follow the setup of [14], using the DDIM solver sampling method [29] with 50 steps. The experiments are conducted on a single setup with an A800 GPU, where our method efficiently uses up to 13.7 GB of GPU memory. Additionally, we set the threshold λ at the 50% quantile of the $diff_t$ and T_{mask} to 200, using UAM combined with token values from the null prompts.

4.1. Experimental Setup

Data Set. To conduct an objective evaluation of the effectiveness of our method for image editing, we conduct experiments using PIE benchmark [14], which has 700 images and a diverse set of complex image editing tasks, including object addition or removal, color changes, and so on. For image composition task, we use the TF-ICON bench mark [21]. In addition, CelebA-HQ dataset [15] and PIE benchmark [14] are used to verify the reconstruction effect of our UAM.

Comparison to other methods. For the image editing task, we consider several baselines, including DDIM [29], Null-Text (NT) [25], Negative Prompt (NP) [23], StyleDiffusion (StyleD) [19] and Direct Inversion (DI) [14]. Additionally, we consider two editing methods: (1) Prompt-to-Prompt (P2P) [10] and (2) MasaCtrl (Masa) [3]. For the image composition task, we compared our approach with the current state-of-the-art, TF-ICON [21].

Metrics. The primary goal of semantic image editing is to accurately modify specific objects or scenes in an image as described in the target text. This process ensures that only the intended part of the image is altered while retaining unmodified parts as much as possible. To assess the effectiveness of our methods, we utilize metrics from prior work [14]. We report the following metrics: (1) Peak Signal-to-Noise Ratio (PSNR) and Mean Squared Error (MSE): These metrics evaluate the faithfulness of the generated images by comparing them to the input images. (2) LPIPS [37]: LPIPS is a deep learning-based metric that assesses perceptual similarity between images, aligning more closely with human perception than traditional metrics. (3) SSIM [34]: SSIM measures the similarity between the two images, focusing on changes in structural information, luminance, and contrast. (4) CLIP Score [26]: We employ a combination of CLIP image and text models to calculate the similarity between generated images and corresponding texts, measuring the alignment between the generated image and the target text. We report CLIP Score for both the entire image (Whole) and within the editing mask (Edited), where regions outside the mask are blacked out. (5) Structural Distance [31]: This metric assesses structural changes in images.

4.2. Image Reconstruction

In Tab. 1 and Tab. 2, our method demonstrates superior reconstruction capabilities, achieving the best results in comparison to the baselines. This further supports the robustness of our approach in generating high-quality images that faithfully adhere to the input specifications.

4.3. Image Composition

Qualitative Evaluation. As shown in Fig. 5, our method achieves a superior balance between semantic expression and fidelity when compared to TF-ICON [21]. The visual comparison highlights that our approach not only maintains higher fidelity to the reference images but also produces more coherent and realistic results across diverse contexts, including natural photographs and artistic styles. For instance, in scenarios requiring complex interactions between foreground and background elements, our method successfully preserves the contextual integrity and stylistic consistency, leading to a more harmonious and visually appealing composition. This indicates that our method is particularly effective in handling the subtleties of image composition, where both the content and style need to be accurately represented.

Quantitative Analysis. In Tab. 5, our method consistently outperforms existing approaches across multiple metrics, confirming its effectiveness in image composition

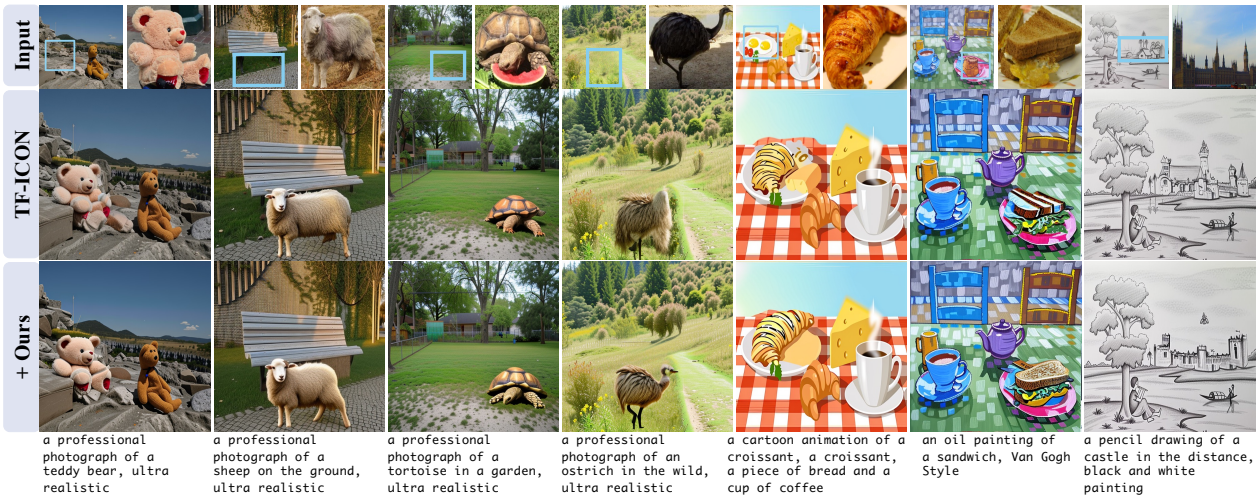

Figure 5. Qualitative comparison with SOTA and baselines in image composition task on TF-ICON bench mark. Our method generates images with higher fidelity to the reference images and produces more realistic results.

Method	Structure ↓ Distance $\times 10^3$	Background Preservation PSNR↑	LPIPS $\times 10^3$↓	MSE $\times 10^4$↓	SSIM $\times 10^2$↑	CLIP Score Whole↑	Edited↑
DDIM	28.38	22.17	106.62	86.97	79.67	23.96	21.16
+ Ours	24.80$_{13\%\downarrow}$	**22.96**$_{3.6\%\uparrow}$	91.56$_{14.1\%\downarrow}$	**76.17**$_{12.4\%\downarrow}$	81.19$_{1.9\%\uparrow}$	24.29$_{1.4\%\uparrow}$	21.21$_{0.2\%\uparrow}$
DI	24.70	22.64	87.94	81.09	81.33	24.38	21.35
+ Ours	**24.60**$_{0.4\%\downarrow}$	22.68$_{0.2\%\uparrow}$	**87.39**$_{0.6\%\downarrow}$	80.63$_{0.6\%\downarrow}$	**81.52**$_{0.2\%\uparrow}$	**24.59**$_{0.9\%\uparrow}$	**21.46**$_{0.5\%\uparrow}$

Table 3. Quantitative comparison of image editing on the PIE benchmark. The methods are compared using the Masactrl attention control [3].

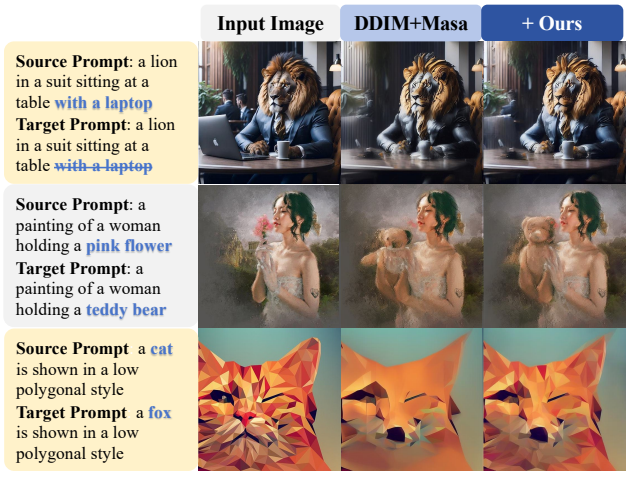

Figure 6. Examples of editing some images using DDIM+Masa on the PIE benchmark.

tasks. Specifically, our approach achieves the lowest LPIPS scores [37] for both background (LPIPS$_{BG}$) and foreground (LPIPS$_{FG}$), which indicates a closer perceptual match to the reference images and, therefore, superior visual quality. Additionally, our method exhibits significant improvements in CLIP scores [26], with higher CLIP$_{Image}$ and CLIP$_{Text}$ values reflecting better alignment between the generated images and the input descriptions. These enhancements suggest that our approach not only excels in producing visually appealing images but also in ensuring that the generated content is semantically coherent and contextually relevant.

4.4. Image Editing

Qualitative Evaluation. As shown in Fig. 6, our method demonstrates a superior balance between semantic expression and image fidelity when applied to both real and generated images, outperforming the DDIM+Masa approach. For instance, in the first row, where a lion in a suit is depicted, DDIM+Masa fails to accurately remove the laptop, leaving artifacts that detract from the overall image quality. In contrast, our method successfully preserves the integrity of the original image while effectively applying the desired edits. Similarly, in the second and third rows, our approach maintains the delicate balance between the new and original elements, ensuring that the edits are both contextually appropriate and visually coherent. These examples illustrate that our method better preserves critical image information and mitigates common mismatches or artifacts seen with DDIM+Masa, leading to more realistic and visually appeal-

Method	Structure ↓ Distance $\times 10^3$	Background Preservation				CLIP Score	
		PSNR↑	LPIPS $\times 10^3$↓	MSE $\times 10^4$↓	SSIM $\times 10^2$↑	Whole↑	Edited↑
NT	13.44	27.03	60.67	35.86	84.11	24.75	21.86
NP	16.17	26.21	69.01	39.73	83.40	24.61	21.87
StyleD	11.65	26.05	66.10	38.63	83.42	24.78	21.72
DDIM	69.43	17.87	208.80	219.88	71.14	25.01	**22.44**
+ Ours	$49.78_{28.3\%\downarrow}$	$18.97_{6.2\%\uparrow}$	$180.85_{13.4\%\downarrow}$	$181.95_{17.2\%\downarrow}$	$73.33_{3.1\%\uparrow}$	$25.09_{0.3\%\uparrow}$	22.23
DI	11.65	27.22	54.55	32.86	84.76	25.02	22.10
+ Ours	$\mathbf{11.05}_{5.2\%\downarrow}$	$\mathbf{27.44}_{0.8\%\uparrow}$	$\mathbf{52.17}_{4.4\%\downarrow}$	$\mathbf{31.46}_{4.3\%\downarrow}$	$\mathbf{85.15}_{0.5\%\uparrow}$	$\mathbf{25.17}_{0.6\%\uparrow}$	$22.14_{0.2\%\uparrow}$

Table 4. Quantitative comparison for image editing on the PIE benchmark. The methods are compared using P2P attention control [10].

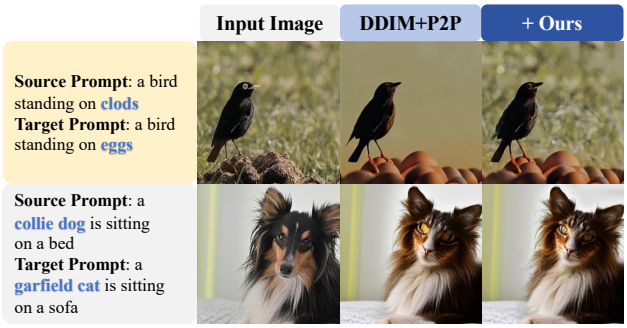

Figure 7. Examples of editing some images using DDIM+P2P on the PIE benchmark.

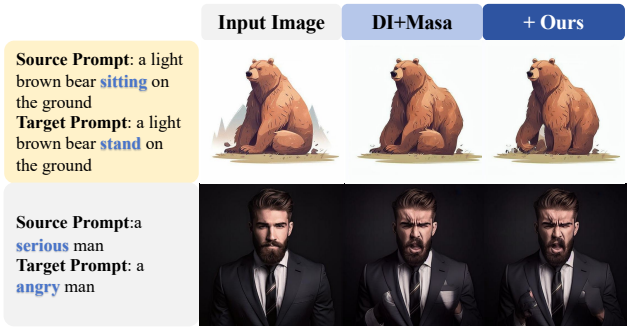

Figure 8. Examples of editing some images using DI+Masa on the PIE benchmark.

Method	LPIPS$_{(BG)}$ ↓	LPIPS$_{(FG)}$ ↓	CLIP$_{(Image)}$ ↑	CLIP$_{(Text)}$ ↑
SDEdit (0.4) [22]	0.35	0.62	80.56	27.73
SDEdit (0.6) [22]	0.42	0.66	77.68	27.98
Blended [1]	0.11	0.77	73.25	25.19
Paint [35]	0.13	0.73	80.26	25.92
DIB [36]	0.11	0.63	77.57	26.84
TF-ICON [21]	0.10	0.60	82.86	28.11
TF-ICON* [21]	0.09	0.51	80.78	31.33
+ UAM*	**0.07**	**0.50**	**81.10**	**31.70**

Table 5. Quantitative comparison of image composition on TF-ICON benchmark [21]. Additional experimental results and details are in [21]. Methods marked with '*' indicate results on A800.

ing results. More results are shown in Figs. 7 and 8.

Quantitative Analysis. In Tab. 4, methods enhanced with our approach exhibit superior performance across a range of metrics compared to their baseline counterparts. Specifically, our methods significantly reduce the Structural Distance [31], indicating a closer visual resemblance to the original images and thereby enhancing fidelity. Moreover, our approach yields improvements in Background Preservation metrics, as evidenced by increased PSNR and SSIM [34] values and decreased LPIPS and MSE scores. These improvements suggest that our method better maintains the original background's integrity while applying the desired edits. Additionally, the CLIP Score for both the whole image and the edited regions shows notable gains, reflecting a more accurate alignment between the generated content and the text prompts. These enhancements collectively underscore the effectiveness of our method in preserving essential image characteristics while performing precise and contextually appropriate edits, thereby achieving a higher quality of image editing compared to existing methods.

4.5. Visualization of Generated Mask

In Fig. 9, we illustrate the masks for the cat as shown in Fig. 4. The masks highlight the areas that need modification, and adaptive selection at different time steps ensures that the modifications are not limited to a specific range, resulting in more realistic images. The masks change with each time step, indicating the areas requiring modifications.

4.6. Ablation Study

Threshold λ. As shown in Tab. 6 (a), the edited images result from setting the threshold λ to different quantiles of the $diff_t$. With an increase in the quantile, the edited image becomes more similar to the original, potentially compromising the desired semantic change. Consequently, a quantile of 0.5 is the chosen setting for subsequent experiments because it offers a balance by sufficiently reflecting the target text while preserving a close resemblance to the original image.
Mask Steps T_{mask}. As shown in Tab. 6 (b), we experiment with T_{mask} values of 0, 200, and 400 for image editing. Notably, $T_{mask} = 200$ emerges as the optimal setting, preserving the original image's details while effectively intro-

Figure 9. The adaptive masks generated by our methods.

	Settings	Structure ↓ Distance×10^3	Background Preservation PSNR↑	LPIPS×10^3↓	MSE×10^4↓	SSIM×10^2↑	CLIP Score Whole↑	Edited↑
(a)	$quantile = 0.7$	21.92	23.72	80.51	66.52	82.47	24.15	20.77
	$quantile = 0.6$	23.35	23.30	86.37	71.69	81.78	24.27	21.15
	$quantile = 0.5$	24.80	22.96	91.56	76.17	81.19	24.29	21.21
	$quantile = 0.4$	26.00	22.66	96.44	80.19	80.69	24.33	21.24
	$quantile = 0.3$	26.92	22.43	100.62	83.29	80.30	24.31	21.24
(b)	$T_{mask} = 0$	28.38	22.17	106.62	86.97	79.67	23.96	21.16
	$T_{mask} = 200$	24.80	22.96	91.56	76.17	81.19	24.29	21.21
	$T_{mask} = 400$	24.70	22.96	91.85	76.03	81.11	24.28	21.18

Table 6. (a) Ablation study on the influence of λ in the editing process using DDIM + Masa with our method when $T_{mask} = 200$. (b) Ablation study on the influence of T_{mask} when $quantile = 0.5$.

ducing the intended semantic changes. This balance ensures that key features, such as the bear's texture, remain intact while still reflecting the desired alterations. In contrast, when $T_{mask} = 0$, the edited image deviates significantly from the original, underscoring the mask's importance. Therefore, we adopt $T_{mask} = 200$ for subsequent experiments.

5. Conclusion

In this work, we introduce Uniform Attention Maps to replace traditional cross-attention in DDIM-based image reconstruction and editing. Our approach significantly improves the fidelity of image reconstructions while maintaining robustness across different text prompts. We also develop an adaptive mask-guided editing technique that seamlessly integrates with our reconstruction method, enhancing the consistency and accuracy of edits. Experimental results demonstrate that our method outperforms existing approaches in image composition and editing tasks. These findings suggest that Uniform Attention Maps hold strong potential for broader applications in image processing.

Acknowledgment This work was supported in part by the National Natural Science Foundation of China No. 62376277, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Public Computing Cloud, Renmin University of China.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4):149:1–149:11, 2023. 8

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18392–18402. IEEE, 2023. 2

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 22503–22513. IEEE, 2023. 2, 3, 6, 7

[4] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 1, 3

[5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[6] Zhenyu Cui, Yuxin Peng, Xun Wang, Manyu Zhu, and Jiahuan Zhou. Continual vision-language retrieval via dynamic knowledge rectification. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 11704–11712. AAAI Press, 2024. 2

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow trans-

formers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 2

[8] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12873–12883. Computer Vision Foundation / IEEE, 2021. 2, 6

[9] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, Di Liu, Qilong Zhangli, Jindong Jiang, Zhaoyang Xia, Akash Srivastava, and Dimitris N. Metaxas. Proxedit: Improving tuning-free real image editing with proximal guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 4279–4289. IEEE, 2024. 2, 3

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 1, 2, 3, 6, 8

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 3

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[13] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly DDPM noise space: Inversion and manipulations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 12469–12478. IEEE, 2024. 3

[14] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 2, 3, 4, 6

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 6

[16] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6007–6017. IEEE, 2023. 3

[17] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 2416–2425. IEEE, 2022. 2

[18] Jiangmeng Li, Wenyi Mo, Wenwen Qiang, Bing Su, and Changwen Zheng. Supporting vision-language model inference with causality-pruning knowledge prompt. *CoRR*, abs/2205.11100, 2022. 2

[19] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. volume abs/2303.15649, 2023. 6

[20] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 6

[21] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. TF-ICON: diffusion-based training-free cross-domain image composition. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2294–2305. IEEE, 2023. 1, 2, 3, 6, 8

[22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 8

[23] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. volume abs/2305.16807, 2023. 2, 3, 6

[24] Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. Dynamic prompt optimizing for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26617–26626. IEEE, 2024. 2

[25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6038–6047. IEEE, 2023. 2, 3, 6

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 6, 7

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 1, 2, 3, 4, 6

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans,

Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 1, 2, 3, 4

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 3, 6

[30] Hongbo Sun, Xiangteng He, Jiahuan Zhou, and Yuxin Peng. Fine-grained visual prompt learning of vision-language models for image recognition. In Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain, editors, *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 5828–5836. ACM, 2023. 2

[31] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10738–10747. IEEE, 2022. 6, 8

[32] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. volume abs/2210.09477, 2022. 3

[33] Bram Wallace, Akash Gokul, and Nikhil Naik. EDICT: exact diffusion inversion via coupled transformations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22532–22541. IEEE, 2023. 2, 3, 4

[34] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 6, 8

[35] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18381–18391. IEEE, 2023. 8

[36] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 231–240. IEEE, 2020. 8

[37] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 6, 7

Algorithm 1 Edit images with adaptive mask

Input: Given original image z_0, target prompt $\mathbf{c}_{tgt}$, source prompt $\mathbf{c}_{src}$, denoising model ϵ_θ, uniform cross-attention maps $\mathcal{C}$, null prompt $\mathbf{c}_\varnothing$, a dilation operation $dilate(\cdot)$.
$z_T^u \leftarrow \text{Invert}(z_0, \mathcal{C}, \mathbf{c}_\varnothing)$
$z_T^{src} \leftarrow \text{Invert}(z_0, \mathbf{c}_{src})$
$z_T^{tgt} \leftarrow z_T^{src}$
for $t = T$ to 1 **do**
  $\#\ Auxiliary\ Branch$
  $\epsilon_u \leftarrow \epsilon_\theta(z_t^u, \mathcal{C}, \mathbf{c}_\varnothing)$
  $\hat{z}_{0,t}^u \leftarrow \frac{1}{\sqrt{\alpha_t}} z_t^u - \frac{1-\alpha_t}{\sqrt{\alpha_t}} \epsilon_u$
  $\#\ Source\ Branch$
  $\epsilon_{src} \leftarrow \epsilon_\theta(z_t^{src}, \mathbf{c}_{src})$
  $\hat{z}_{0,t}^{src} \leftarrow \frac{1}{\sqrt{\alpha_t}} z_t^{src} - \frac{1-\alpha_t}{\sqrt{\alpha_t}} \epsilon_{src}$
  $\#\ Target\ Branch$
  $\epsilon_{tgt} \leftarrow \epsilon_\theta(z_t^{tgt}, \mathbf{c}_{tgt})$
  $\hat{z}_{0,t}^{tgt} \leftarrow \frac{1}{\sqrt{\alpha_t}} z_t^{tgt} - \frac{1-\alpha_t}{\sqrt{\alpha_t}} \epsilon_{tgt}$
  $M \leftarrow dilate(|\hat{z}_{0,t}^{tgt} - \hat{z}_{0,t}^{src}| \leq \lambda)$
  if $t < T_{mask}$ **then**
    $\hat{z}_{0,t}^{tgt} \leftarrow M \odot \hat{z}_{0,t}^u + (1 - M) \odot \hat{z}_{0,t}^{tgt}$
  end if
  $z_{t-1}^{tgt} \leftarrow \sqrt{\alpha_{t-1}} \hat{z}_{0,t}^{tgt} + \sqrt{1 - \alpha_{t-1}} \epsilon_{tgt}$
  $z_{t-1}^{src} \leftarrow \sqrt{\alpha_{t-1}} \hat{z}_{0,t}^{src} + \sqrt{1 - \alpha_{t-1}} \epsilon_{src}$
  $z_{t-1}^u \leftarrow \sqrt{\alpha_{t-1}} \hat{z}_{0,t}^u + \sqrt{1 - \alpha_{t-1}} \epsilon_u$
end for
return z_0^{tgt}

A. Adaptive Mask-Guided Image Editing: Algorithm Overview

The pseudocode for our adaptive mask method is shown in Algorithm 1. The algorithm takes an input image z_0, a target prompt $\mathbf{c}_{tgt}$, and a source prompt $\mathbf{c}_{src}$. The method starts by inverting the image through auxiliary and source branches and then initializes the target branch from the source branch.

At each timestep t, we compute noise predictions and update the latent variables in the auxiliary, source, and target branches. It generates an adaptive mask M by comparing the clean images $\hat{z}_0$ from the target and source branches and applies a dilation operation to ensure robustness. The mask M is then used to blend the predictions from the auxiliary and target branches, preserving key details of the original image while applying the edits.

The process repeats until the final image z_0^{tgt} is returned, incorporating the original information and the desired modifications.

B. More Examples of Image Reconstruction

Figs. 11 to 14, provide additional examples of image reconstruction using DDIM inversion with 20 timesteps on the PIE benchmark, showcasing the performance of our method in comparison to null prompts and source prompts. In Figs. 11 to 14, we observe the reconstruction of various images. The results using the null prompt often produce blurred or incorrect outputs, while the source prompt reconstructions are better but still show visible artifacts. By leveraging uniform attention maps, our method demonstrates significant improvements, yielding clearer and more accurate reconstructions that align closely with the original input images, preserving important details such as texture and shape. These examples confirm the robustness of our approach across different image types, showing that our method consistently outperforms the baseline approaches in generating high-quality reconstructions that faithfully resemble the input images.

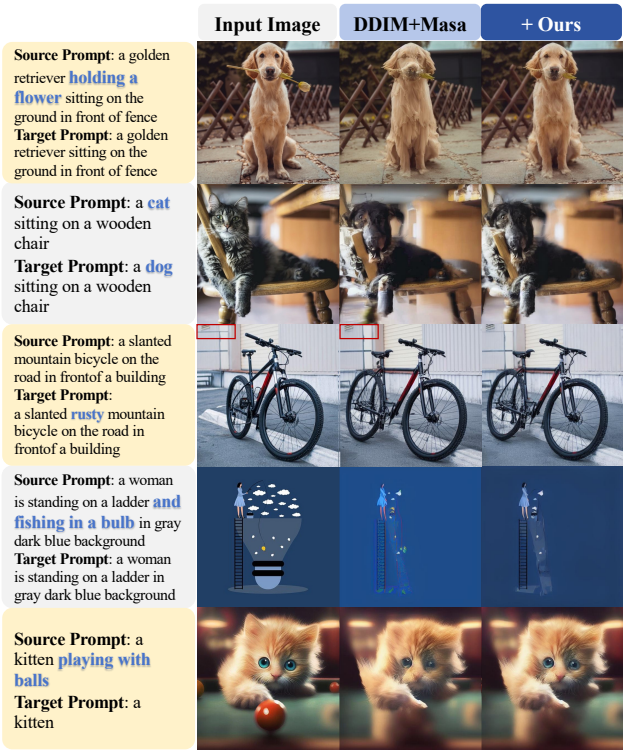

Figure 10. More examples of image editing on the PIE benchmark. Examples of image editing on the PIE benchmark, comparing the DDIM+Masa method with our image editing method.

C. More Examples of Image Editing

Fig. 10 showcases the effectiveness of our image editing method compared to the DDIM+Masa baseline. Our method consistently produces more accurate, detailed, and visually coherent edits across various scenarios, such as transforming animals, modifying complex objects, and retaining structural fidelity in abstract compositions, outperforming the baseline in terms of both precision and consistency.

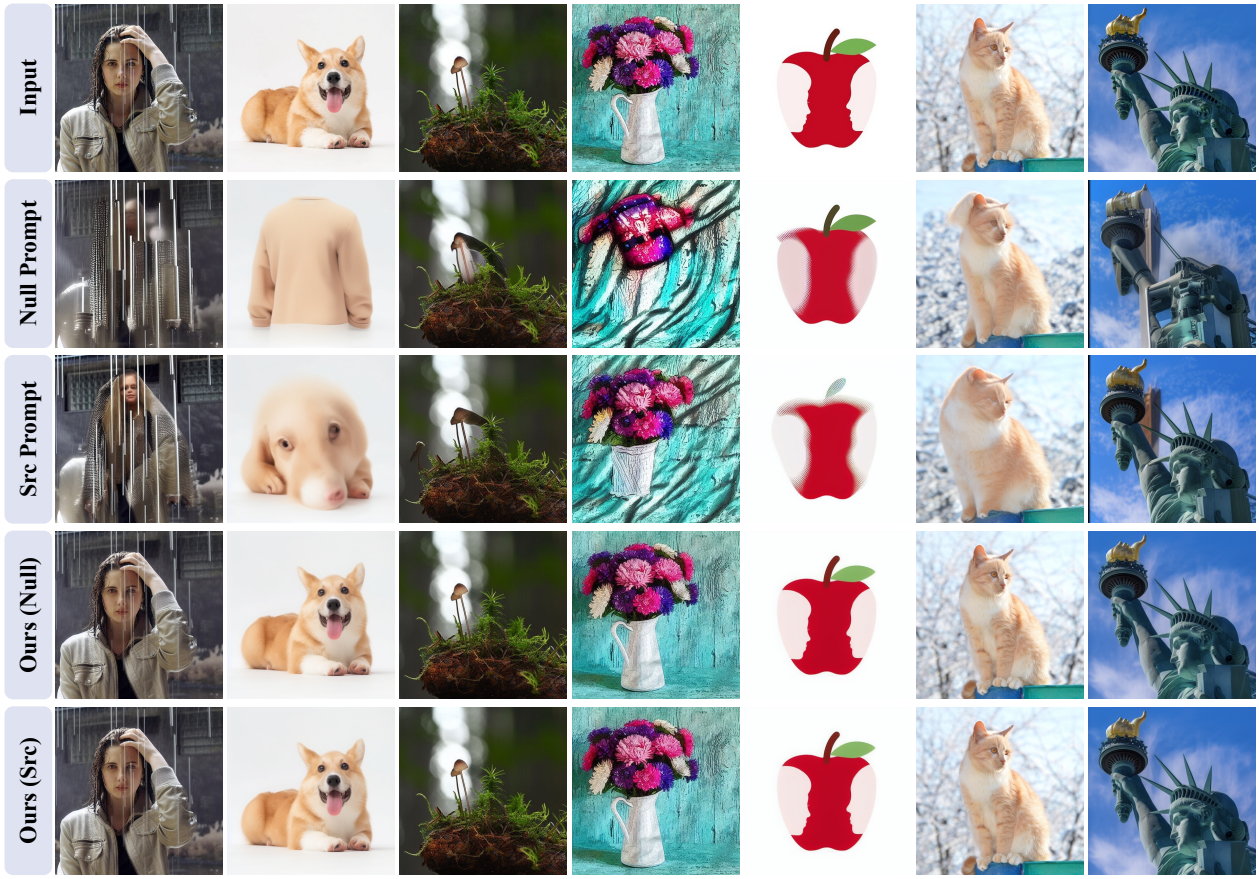

Figure 11. Examples of image reconstruction on the PIE benchmark. The first row shows the input images. The second and third rows display the results using a null prompt (an empty string) and a source prompt from the benchmark, respectively. The fourth and fifth rows show the results from our method with different value tokens, demonstrating superior reconstruction quality and better alignment with the original input images.

D. More Experimental Details

We conduct experiments in Fig. 3 and Fig. 2 using Stable Diffusion v1.4 with DDIM inversion and reconstruction under 20 inference steps. At each timestep, the cross-attention term $A^{(l)}$ is extracted from U-Net layers with an output dimension of 64×64. These terms are analyzed or visualized to examine the impact of cross-attention misalignment during the inversion and reconstruction processes. The clean predicted image $\hat{z}_{0,t}$ is also generated at each timestep to evaluate the reconstruction fidelity. All experiments are conducted using 700 images from the PIE benchmark dataset.

In Fig. 3, the Mean Squared Error of the cross-attention term is computed at the pixel level as the discrepancy between $A^{(l)}_{\text{inv}}$ and $A^{(l)}_{\text{rec}}$, with the results averaged across all pixels. Similarly, the reconstruction error is calculated as the pixel-level MSE between the predicted clean images $\hat{z}_{0,\text{inv}}$ and $\hat{z}_{0,\text{rec}}$. These two MSE metrics are aggregated across all timesteps for each image. The scatter plot in Fig. 3 illustrates a strong positive correlation between the cross-attention discrepancies and the reconstruction errors, demonstrating that misalignment in the cross-attention mechanism is a significant contributor to the errors in the final reconstructed images.

In Fig. 2, the extracted cross-attention terms $A^{(l)}$ are visualized as heatmaps to show their temporal evolution across the inversion and reconstruction processes. Fig. 2 (a) highlights the discrepancies in the cross-attention maps under source prompts, null prompts, and our proposed method. The heatmaps for the source and null conditions reveal significant misalignments between the inversion and reconstruction phases, emphasized by the black-boxed regions. In contrast, our method ensures consistent cross-attention alignment throughout the process. Furthermore, Fig. 2 (b) presents the corresponding clean predicted images $\hat{z}_{0,t}$ at various timesteps, showing that the proposed method maintains high-quality reconstructions, while the source and null prompts result in noticeable distortions.

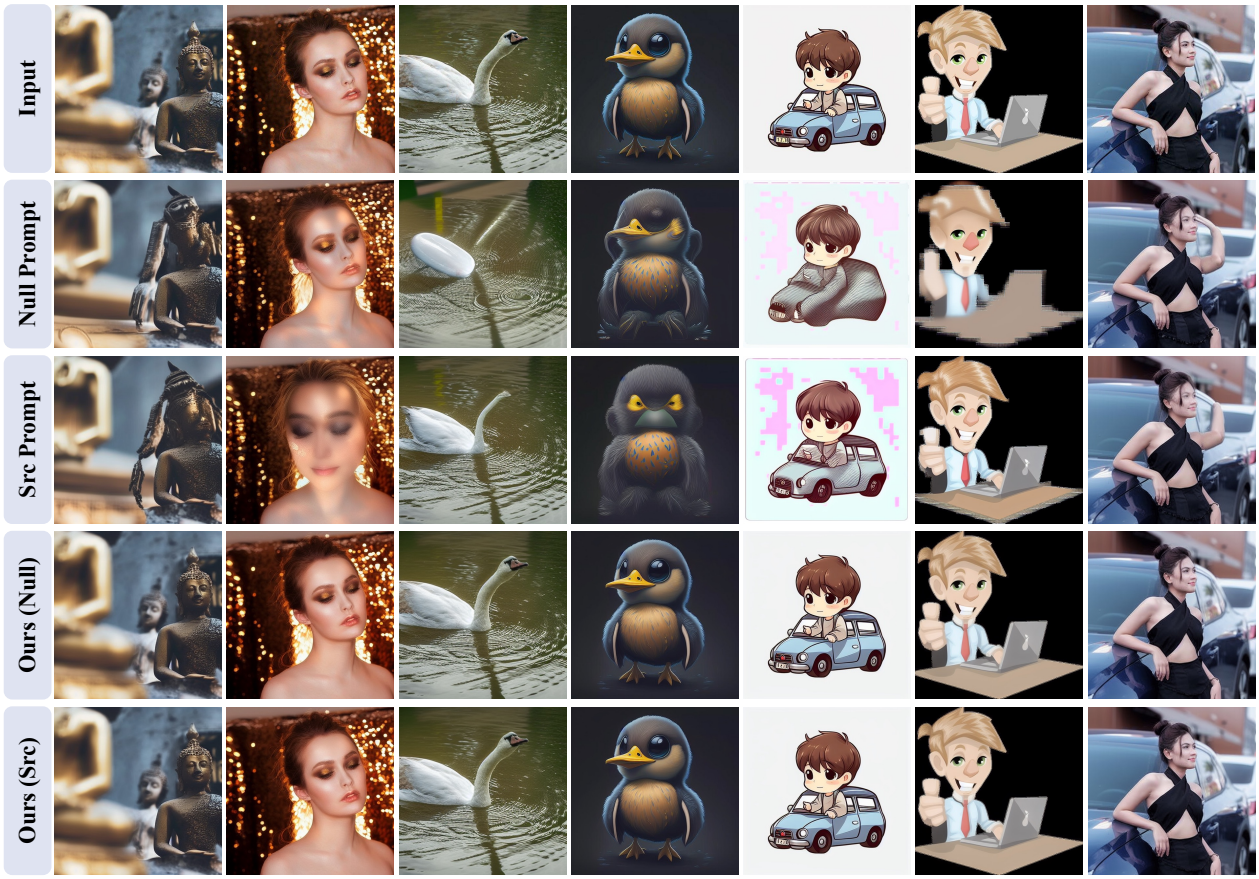

Figure 12. More examples of image reconstruction on the PIE benchmark.

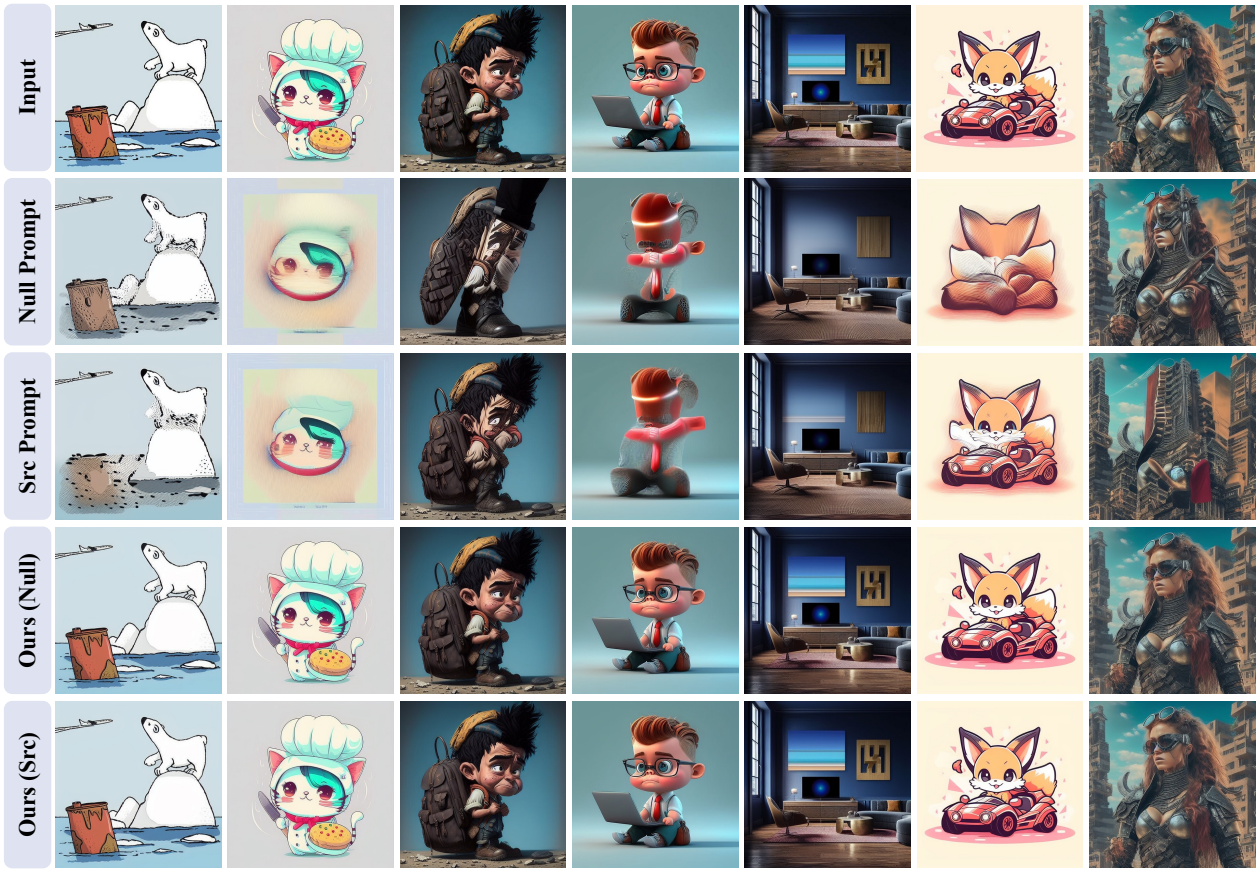

Figure 13. More examples of image reconstruction on the PIE benchmark.

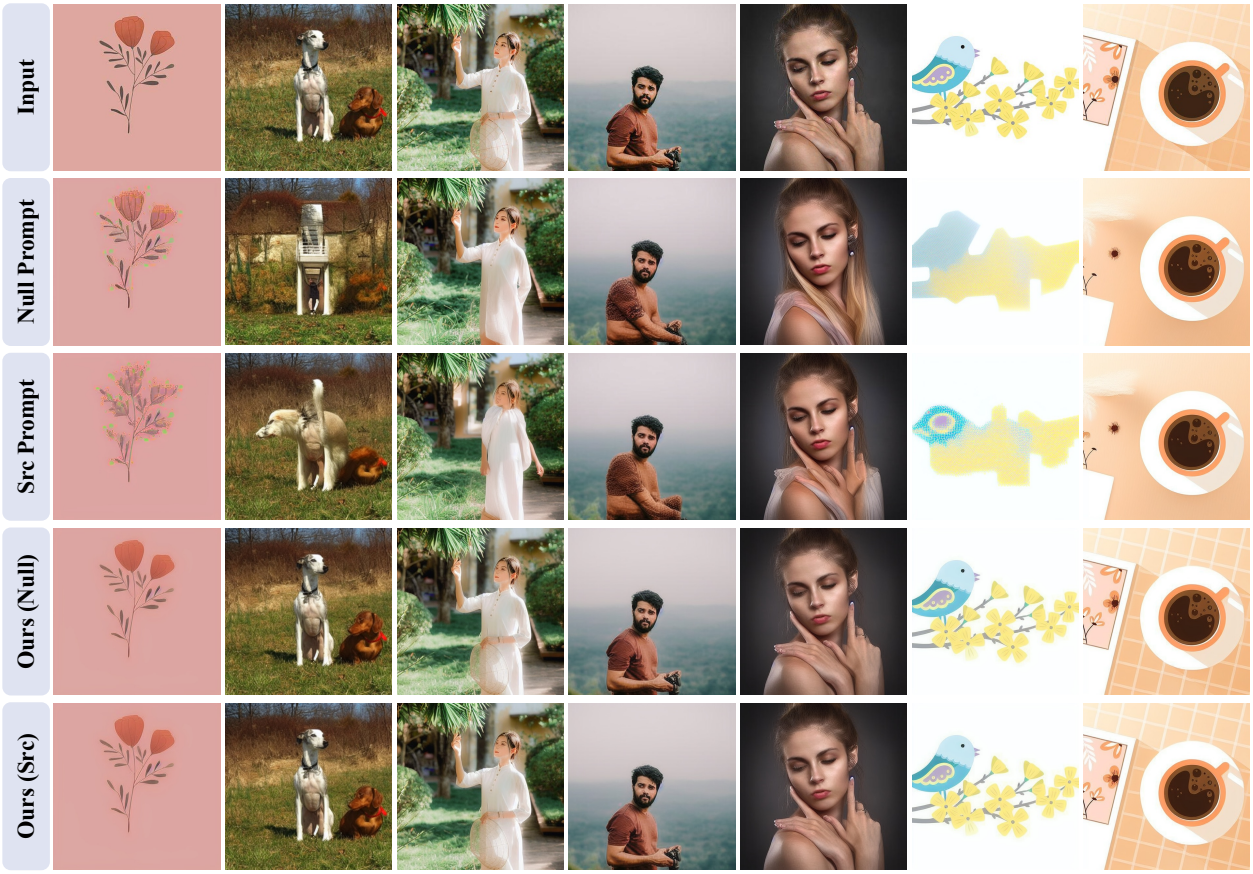

Figure 14. More examples of image reconstruction on the PIE benchmark.