Towards Efficient Diffusion-Based Image Editing with Instant Attention Masks

Siyu Zou[1*†], Jiji Tang[2†], Yiyi Zhou[1], Jing He[1], Chaoyi Zhao[2], Rongsheng Zhang[2], Zhipeng Hu[2], Xiaoshuai Sun[1‡]

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China
[2] Fuxi AI Lab, NetEase Inc., Hangzhou, China
zousiyu@stu.xmu.edu.cn, tangjiji01@corp.netease.com, zhouyiyi@xmu.edu.cn, blinghe@stu.xmu.edu.cn
{zhaochaoyi, zhangrongsheng, zphu}@corp.netease.com, xssun@xmu.edu.cn

Abstract

Diffusion-based Image Editing (DIE) is an emerging research hot-spot, which often applies a semantic mask to control the target area for diffusion-based editing. However, most existing solutions obtain these masks via manual operations or off-line processing, greatly reducing their efficiency. In this paper, we propose a novel and efficient image editing method for Text-to-Image (T2I) diffusion models, termed *Instant Diffusion Editing* (InstDiffEdit). In particular, InstDiffEdit aims to employ the cross-modal attention ability of existing diffusion models to achieve instant mask guidance during the diffusion steps. To reduce the noise of attention maps and realize the full automatics, we equip InstDiffEdit with a training-free refinement scheme to adaptively aggregate the attention distributions for the automatic yet accurate mask generation. Meanwhile, to supplement the existing evaluations of DIE, we propose a new benchmark called *Editing-Mask* to examine the mask accuracy and local editing ability of existing methods. To validate InstDiffEdit, we also conduct extensive experiments on *ImageNet* and *Imagen*, and compare it with a bunch of the SOTA methods. The experimental results show that InstDiffEdit not only outperforms the SOTA methods in both image quality and editing results, but also has a much faster inference speed, *i.e.*, +5 to +6 times. Our code available at https://github.com/xiaotianqing/InstDiffEdit

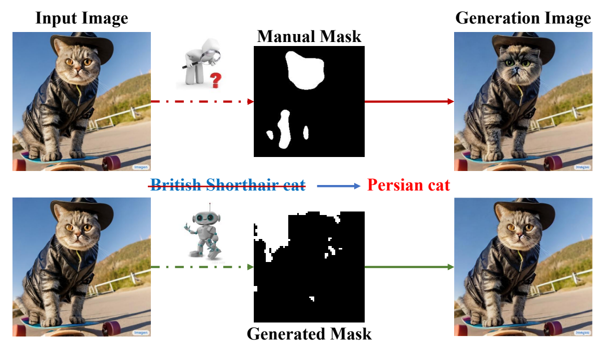


Figure 1: Illustration of existing diffusion-based image editing methods, where a manually or off-line generated mask is often used to control the editing area.

Introduction

For a year or two, diffusion models have gradually become the mainstream paradigm in conditional image generation (Saharia et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; Balaji et al. 2022; Nichol et al. 2021). Compared with *Generative Adversarial Networks* (GAN) (Karras, Laine, and Aila 2019; Karras et al. 2020; Xia et al. 2021), diffusion models yield a completely different generation pipeline, which can obtain more diverse and interpretable generations. The great success of diffusion models also sparks researchers to apply them to the task of semantic image editing (Meng et al. 2021; Kawar et al. 2022).

Semantic image editing (Zhan et al. 2021) aims to modify the target instance of the given image according to the input text description, while the rest image information needs to be preserved as much as possible. Although existing diffusion models (Saharia et al. 2022; Ramesh et al. 2022; Rombach et al. 2022) excel in generation quality and diversity on text-to-image generation, it still lacks precise controls. Therefore, recent diffusion-based editing methods introduce additional information to better control the image manipulation, such as reference image (Meng et al. 2021) or semantic mask (Avrahami, Lischinski, and Fried 2022).

Among these solutions, padding a semantic mask is the most effective way for accurate image editing, which can precisely restrict the target image area and achieve editing via text-to-image diffusions (Avrahami, Lischinski, and Fried 2022), as shown in Fig. 1. However, the mask generation often requires manual intervention (Avrahami, Lischinski, and Fried 2022; Couairon et al. 2022b), greatly limiting the efficiency of these methods for the practical use.

Recent advance has aspired to automate the editing process via reducing the manual efforts or including the mask generation in diffusion models. For instance, PtP (Hertz et al. 2022) proposes a semi-automated method, which can directly obtain mask by manually setting some parameters. More recently, DiffEdit (Couairon et al. 2022b) proposes a fully automatic method, which can embed the mask generation into the diffusion framework, but its mask generation and image editing are still time consuming. Overall, exist-

*This work was done during her internship at Fuxi AI Lab.
†These authors contributed equally.
‡Corresponding author

ing solutions still exhibit obvious shortcomings in terms of either manual intervention or computation efficiency.

In this paper, we propose a novel yet efficient image editing method for diffusion models, termed *Instant Diffusion Editing* (InstDiffEdit). The feasibility of InstDiffEdit is attributed to the superior cross-modal alignment of existing diffusion models. In the advanced diffusion models like Stable Diffusion (Rombach et al. 2022), an effective multimodal space has been well established by learning numerous image-text pairs, and these models also involve excellent cross-attention mapping. In this case, we can leverage the hidden attention maps in diffusion steps to facilitate instant mask generation. However, these hidden attention maps are intractable to directly use, and they are often full of noise. For instance, the semantic attentions of start token are much more noisier than that of "*cat*" in Fig. 2. Thus, we also equip InstDiffEdit with a learning-free mask refinement scheme, which can adaptively aggregate the attention distributions according to the editing instruction. Notably, the proposed InstDiffEdit is a plug-and-play component for most diffusion models, which is also training-free.

To validate InstDiffEdit, we apply it to Stable Diffusion v1.4 (Rombach et al. 2022), and conduct extensive experiments on two benchmark datasets, namely ImageNet (Deng et al. 2009) and Imagen (Saharia et al. 2022). Meanwhile, to better measure the local editing ability and mask accuracy of existing methods, we also propose a composite benchmark called *Editing-Mask*, as a supplementary evaluation to DIE. The experimental results on ImageNet and Imagen show that compared with existing methods, InstDiffEdit can achieve the best trade-off between computation efficiency and generation quality for semantic image editing. For instance, compared with the recently proposed DiffEdit, our method can obtain competitive editing results while improving the inference speed by 5 to 6 times. The results on Editing-Mask confirm the superiority of our method in background preservation. Furthermore, we also provide sufficient visualizations to examine the ability of InstDiffEdit.

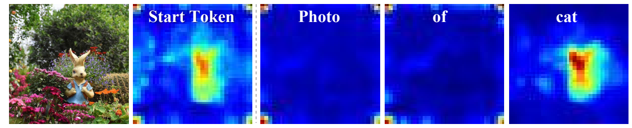
Conclusively, the contribution of this paper is three-fold:

- We propose a novel and efficient image editing method for diffusion-based models, termed *InstDiffEdit*, which obtains instant mask guidance via exploiting the cross-modal attention in diffusion models.
- As a plug-and-play component, InstDiffEdit can be applied to most diffusion models for semantic image editing without further training or human intervention, and its performance is also SOTA.
- We propose a new image editing benchmark, termed Editing-Mask, containing 200 images with human-labeled masks, which can be used for the evaluation of mask accuracy and local editing ability.

Related Work

Text-to-Image Diffusion

In the past few years, a lot of diffusion-based methods (Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022) has been proposed, which also demonstrate superior performance in terms of image quality and diversity compared to GAN. (Karras et al. 2020; Xia et al. 2021). Some recent works (Avrahami, Lischinski, and Fried 2022) also explore the combination of diffusion models with *Contrastive Language-Image Pre-Training* (CLIP) (Radford et al. 2021). For example, Stable Diffusion (Rombach et al. 2022) leverages CLIP's text encoder to guide the image generation process. By incorporating cross-attention between text and noisy images, the model generates images that are semantically aligned with the textual description.



(a) Input Image **(b) The attention maps of LDM with different Tokens**

Figure 2: The visualization of the attention maps in Stable Diffusion. The target word of "*cat*" has the best attention map, but it needs to be manually identified during applications. The start token is relevant but still very noisy.

Semantic Image Editing

A plethora of GAN-based semantic image editing approaches (Goodfellow et al. 2014; Xu et al. 2018; Xia et al. 2021) have been proposed with remarkable outcomes. The emergence of large-scale GAN networks, such as the StyleGAN family (Karras, Laine, and Aila 2019; Karras et al. 2020, 2021), significantly enhances the editing capabilities. Meanwhile, Transformer (Vaswani et al. 2017) has demonstrated remarkable performance in text-driven image editing tasks. ManiTrans (Wang et al. 2022) use Transformers to predict the content of covered regions, which enables semantic editing only performing on a certain image region.

Recently, with the developments of diffusion models, practitioners also explore their application in semantic image editing. SDEdit (Meng et al. 2021) accomplishes this by retaining a portion of the reference image information during the diffusion process. CycleDiffusion (Wu and De la Torre 2022) proposes an inversion model to get a better latent from the input image, thus improving the edit quality. PtP (Hertz et al. 2022) and PnP (Tumanyan et al. 2022) operate editing via modifying attention maps in diffusion models. More recently, to prevent unbounded edits from global image editing, some methods resort to local editing techniques. For example, Blended Diffusion (Avrahami, Lischinski, and Fried 2022) and RePaint (Lugmayr et al. 2022) implement local editing on real images with manual mask. However, the acquisition of manual masks is time-consuming and labor-intensive, and hinders the developments of automated semantic editing.

Therefore, some methods have begun to explore automated mask generation. DiffEdit (Couairon et al. 2022b) is better suited to the requirements of automated editing as it obtains the mask by contrasting variations in model predictions with different text prompts. However, because of the stochastic randomness of the diffusion model, DiffEdit requires multiple iterations to stabilize the ultimate output, which leads to inefficiencies in terms of time.

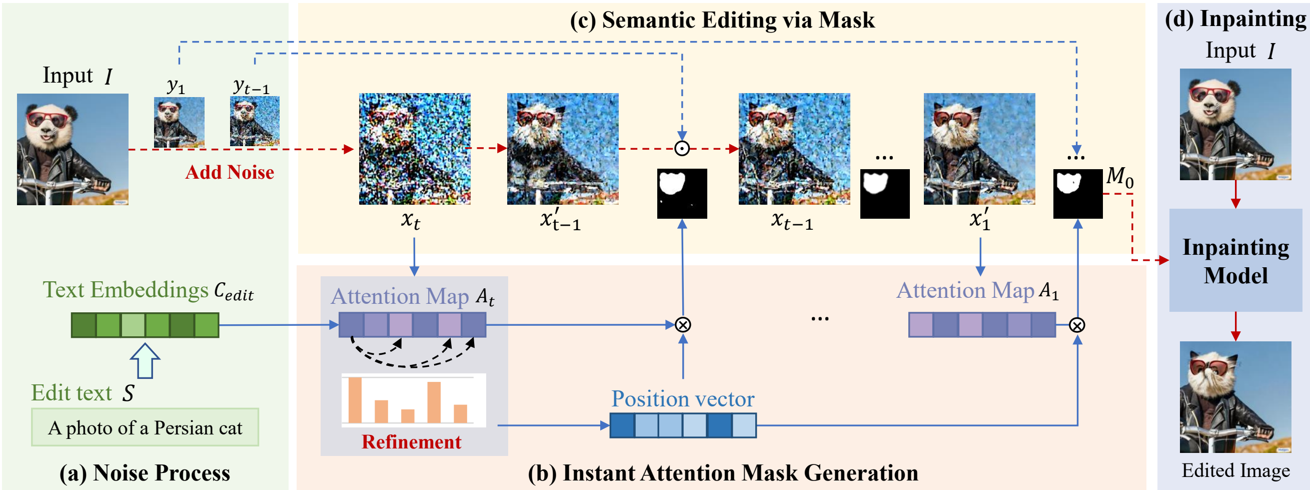


Figure 3: The framework of the *Instant Diffusion Editing* (InstDiffEdit). InstDiffEdit involves instant mask generation at each denoising step based on the attention maps. This mask can provide instant guidance for the image denoising. The left part (a) illustrates the noise process, and (b) depicts the generation of semantic mask at each step, based on which the diffusion-based image editing is performed (c). Lastly, the inpainting model is further applied to accomplish the generation (d).

Preliminary

Latent Diffusion Models

Traditional diffusion models (Ho, Jain, and Abbeel 2020) typically operate the diffusion process on high-resolution image space, which significantly limits training and generation speed. In order to achieve more efficient training and generation, Latent Diffusion Models (LDMs) (Rombach et al. 2022) perform the diffusion process on the latent space rather than the resolution space, thereby improving the efficiency of training and inference.

First of all, LDMs leverages an automatic encoder framework E_I, such as VAE (Kingma and Welling 2013), to map the image features I to low-dimensional latent spaces x_0 and generate noisy image features x_t through the diffusion forward process:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_t, x_0 = E_I(I), \tag{1}$$

where t denotes the time-step, which is determined by noise strength r. The noise term ϵ_t is sampled from a *standard normal distribution*. α_t is a decreasing schedule of diffusion coefficients that controls the strength of noise ateach step.

Subsequently, the text sequence S is mapped to a feature space using a text encoder E_T such as CLIP (Radford et al. 2021), recorded as $C_{edit} = E_T(S)$. The diffusion process period is operated on latent space, denoted as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, c, t)) + \sigma_t z. \tag{2}$$

Finally, a decoder D_I, which corresponds to the encoder E_I, is employed to reconstruct the image from the latent dimension with $I_{rec} = D_I(x_0)$.

Cross-Attention in LDMs

In LDMs, text-to-image generation is accomplished by modifying the latent representations using cross-attention alignments. Specifically, for each text S which consists of N tokens, the pre-trained text encoder $CLIP_T$ is utilized to transform it into the text feature $c = \{c_1, c_2, \ldots, c_N\}$. Similarly, input image is transformed into image latent x_0 and the noisy image latent x_t is obtained according to Eq. 1.

Subsequently, the text features and image latent are projected by three trainable linear layers, denoted as f_Q, f_V, and f_K. Next, the spatial attention maps A is generated for each text token by:

$$A = Softmax(\frac{QK^T}{\sqrt{d_k}}), Q = f_Q(z_t), K = f_K(t), V = f_V(t) \tag{3}$$

where d_k denotes the feature dimension of K. And the attention maps A is then combined with the value matrix V to obtain the final output of the cross-attention layer with $V \cdot A$.

Generally, the attention maps in Stable Diffusion can indicate the correspondence between text words and image regions. However, due to the noise contained in image latent , it is challenging to directly obtain the desired target instance from the attention maps, and these hidden attention maps are still of noisy, as shown in Fig. 2.

Methodology

Overview

In this paper, we propose a novel and efficient image editing method based on text-to-image diffusion models, termed *Instant Diffusion Editing* (InstDiffEdit), of which structure is illustrated in Fig. 3.

Concretely, similar to exisitng methods (Avrahami, Lischinski, and Fried 2022), we aim to achieve the target image editing by padding a semantic mask to input image, based on which the diffusion steps are conducted to achieve target edition. This process can be defined by:

$$x_t = M \cdot x'_t + (1-M) \cdot y_t, \tag{4}$$

where, x'_t and y_t denote the predicted noisy latent and the latent representation of the noisy image at step t, and M is the mask. Then, we can get the noisy latent x_t for editing.

This mask-based editing is supported by recent advances in diffusion models (Avrahami, Lischinski, and Fried 2022), which can restrict editing areas using mask and replace the non-masked area of the predicted image with noise image at the current timestep. This allows mask-based methods to preserve the background in the non-masked area while editing. However, the generation of this semantic mask often requires manual efforts (Hertz et al. 2022; Patashnik et al. 2023) or off-line processing (Avrahami, Lischinski, and Fried 2022; Lugmayr et al. 2022). In this case, InstDiffEdit resorts to the attention maps in LDMs for instant mask generation during diffusions. As shown in Fig. 2, the attention maps in LDMs capture the semantic correspondence between the image and text well.

However, it also encounters some problems. To specify the attention map of the editing target, *e.g., "cat"* in Fig. 2, the method still requires manual efforts, since we do not know the length and content of user's instruction during application. And directly using the map of *"start token"* as a trade-off is still too noisy for efficetive edition.

In this case, we equip InstDiffEdit with an automatic refinement scheme for mask generation. As shown in Fig. 3, given an input image latent feature x_t, and a text feature C_{edit}, we can get the hidden attention maps A in denoise process from Eq. 3. Then, we propose a parameter-free attention mask generation module $G(\cdot)$ to obtain the semantic mask $M_t = G(x_t, C_{edit})$. Later, with this instant mask, we can directly perform target image editing during the diffusion steps, which can be re-written by:

$$x_{t-1} = M_t \cdot \epsilon_\theta(x_t, t, C_{edit}) + (1 - M_t) \cdot y_t. \tag{5}$$

where, M_t is the mask computed by the attention mask module in timestep t and ϵ_θ denotes the diffusion model.

Lastly, in order to achieve better generation results, we adopt a strategy of using the mask generated in the last denoising step as the final mask, and generating the final editing results through the inpainting way in LDMs.

In the next subsection, we will give the detail definition of the proposed attention mask generation module.

Instant Attention Mask Generation

In InstDiffEdit, we use the attention maps generated in the denoising process as the information source for mask generation. However, the input text often consists of multiple tokens, and the attention information of each token has its own focus and varies vastly with the change of sentence length and word composition. Therefore, it is difficult for the model to automatically locate attention results of the target words.

In practice, we use the attention maps of the start token as the base information for further attention mask refinements. To explain, in a well pre-trained T2I diffusion model, the start token often expresses the semantics of the whole sentence. As shown in Fig. 2, the focus region of attention corresponding to the start token overlaps highly with the edit region of the semantic description. However, the start token

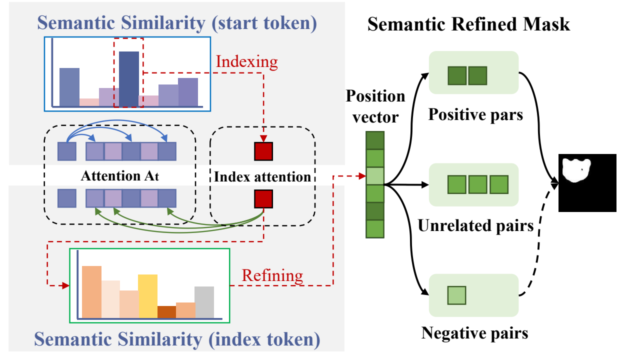


Figure 4: The proposed instant mask generation. An indexing process is first performed based on the semantic similarities between the start token and the other ones (upper left). Refinement is then operated between the index and the remaining ones (lower left). Finally, the mask is obtained via the adaptive aggregation of all attention maps.

contains the whole sentence as well as part of the original image information, so its attention distribution is still messy.

In this case, we adopt the idea of key information extraction to eliminate the noisy information and obtain the most relevant content with semantic information. Assuming a noise strength of r, the denoise process starts at time-step τ ($\tau = r * T, T = 1000$), and the corresponding attention maps A_τ can be obtained using Eq. 3. Specifically, we leverage the attention map of the start token $A^\tau_{start} \in R^{16\times16}$ as the reference information, and subsequently retrieve the attention $A^\tau_{index} \in R^{16\times16}$ by computing all similarities with the reference map. This enables us to identify the location of the object that requires modification:

$$A^\tau_{index} = argmax \sum_{i\in[1,N]} cosine(A^\tau_i, A^\tau_{start}), \tag{6}$$

where $cosine(\cdot)$ denotes semantic similarity and N is the length of all tokens in sentence.

To obtain more accurate mask information, we further aggregate the concept-related information and eliminate irrelevant information. Specifically, we compute the similarities between the obtained A^τ_{index} and the attention maps of the text tokens to obtain a similarity vector $S \in R^{1\times N}$:

$$S_i = cosine_{i\in[1,N]}(A^\tau_i, A^\tau_{index}). \tag{7}$$

In principle, the similarity of the attention maps at each token is closely related to the semantic similarity of the sentence. As the attention maps are associated with the core semantic, the similarities will be larger, and *vice versa*.

Afterwards, we can get a position vector to weight the attention information via filtering the similarity vector with two thresholds:

$$P_{i\in[1,N]} = \begin{cases} 1 & S_i > \gamma_1, \\ -1 & S_i < \gamma_2, \\ 0 & others. \end{cases} \tag{8}$$

Computing semantic similarities at each step of the denoising process can be time-consuming due to the large dimensionality of the attention maps. To mitigate this issue, we propose to compute the position vector P only in the first step τ of the denoising process.

Finally, we obtain the refined attention map A_t^{ref} with attention maps A_t and P at timestep $t \in \{\tau, \dots, 0\}$ ($A_t^{ref} = P \cdot A_t$), which is then processed using Gaussian filtering and *binarized* with a threshold φ to obtain the final mask M_t:

$$M_t(x, y) = \begin{cases} 1 & A_t^{ref}(x, y) > \varphi, \\ 0 & others. \end{cases} \tag{9}$$

here, (x, y) refers to a point in the latent space of the image. Notably, the above instant attention mask generation module is training free, and thus it can be directly plugged into most existing T2I diffusion models. Meanwhile, through the refine processing, the obtained mask is much superior than the ones before refining.

Semantic Editing via Mask

Through the mask generation module, we obtain a mask at each step of the image denoising process. Thus, by blending the mask, guidance can be provided to denoising by Eq. 4.

However, since all the information in the masked area is essentially discarded, the resulting image often has local semantic consistency but does not consider global semantics, leading to artifacts. Additionally, when the noise level is low, some editing operations cannot be achieved, such as color modification. Thus, we also equip InstDiffEdit with an inpainting based method for semantic image editting.

The inpainting method (Rombach et al. 2022) initializes the information in the masked area with completely random noise and considers global information during generation, thus eliminating artifacts and editing failures caused by the original image information. Nevertheless, the performance of inpainting is highly dependent on the accuracy of mask.

Therefore, we combine the advantages of the two methods by using attention maps to generate mask in the denoising process, thereby guiding image generation and obtaining more accurate mask during denoising.

Finally, we use the inpainting method on the mask generated in the last step of denoising to generate an image that is artifact-free and more consistent with the remaining information in the original image. Notably, the combination of two mask editing methods only slightly increases the computation cost of semantic image editing.

Experiments

Experiment Setting

Datasets We use ImageNet, Imagen and Editing-Mask to evaluate the performance of semantic editing task.

- **ImageNet** Followed the evaluation of Flexit (Couairon et al. 2022a). A total of 1092 images in ImageNet (Deng et al. 2009) are included, covering 273 categories. For each image, the edit text is another similar category .
- **Imagen** We construct an evaluation dataset for semantic editing by utilizing the generations from the Imagen (Saharia et al. 2022) model. Specifically, we randomly selected a short text which not in the input text as the edit text, such as replacing "British shorthair cat" with "Shiba Inu dog", resulting in a dataset of 360 paired samples.
- **Editing-Mask** A new dataset, which comprises 200 images randomly selected from Imagen and ImageNet. Each sample includes an image, input text, edit text, and a human-labelled mask that corresponds to the semantics of the edit text. Our proposed dataset enables direct evaluation of the performance of editing tasks, particularly in regions where editing is necessary.

Metrics We evaluate the performance of editing methods in terms of time efficiency and generation quality. Specifically, we measure the average editing time of an image at a resolution of 512 to assess the time consumption of each method. Additionally, we used the *Learned Perceptual Image Patch Similarity* (**LPIPS**) (Zhang et al. 2018) metric to quantify the difference between the generated image and the original image, which reflects the degree of modification made by the editing method. Furthermore, we employed the *Classwise Simplified Fréchet Inception Distance* (**CSFID**) (Couairon et al. 2022a) metric, which is a category FID metric that measures the distance between generated and original images. We also use **CLIPScore** (Hessel et al. 2021) to measure the semantic similarities between the edit texts and generated images. It is noted that all of these metrics evaluate the generated image quality rather than the editing performance. Therefore, in our proposed human-labeled mask daraset, we use *Intersection over Union* (**IOU**) to assess the quality of the generated masks, $\boldsymbol{C_m}$ and $\boldsymbol{C_{non}}$ to represented the modifications of the image in the mask and non-mask areas. The metrics on Editing-Mask provides a more direct evaluation of editing performance.

Implementation The framework of InstDiffEdit is based on Stable Diffusion v1.4. We use 50 steps of LDMScheduler sampler with a scale 7.5, and set noise strength to $r = 0.5$, threshold of binarization to $\varphi = 0.2$, and the thresholds for attention refinement defined in Eq. 8 are 0.9 and 0.6 by default, respectively. We maintain $n = 3$ rounds of denoising on the input image in parallel throughout the entire denoising process. Finally, we use the inpainting mode in Stable Diffusion to get the target image.

Experimental Results

Quantitative Analysis In this section, we present quantitative results on three datasets.

Comparison With Existing Methods. To validate the effectiveness of the proposed InstDiffEdit, we compare it with five diffusion-based methods, of which results are given in Tab. 1 and Fig. 5. The latent-based methods, *i.e.*, SDEdit (Meng et al. 2021) and CycleDiffusion (Wu and De la Torre 2022), which rely on the association between the generated image's latent and the original image's latent. These methods offer the advantage of low time cost for editing. However, their performance is much worse than the other methods. Meanwhile, attention-based methods, *i.e.*, PtP (Hertz et al. 2022) and PnP (Tumanyan et al. 2022), infer on the latent representation of real images, resulting in lower time efficiency and heavy reliance on the performance of inversion. As a mask-based model, DiffEdit (Couairon et al. 2022b) achieves significant improvements over all datasets, indicating the effectiveness of generated masks in diffusion-based

Category	Models	Time↓	Editing-Mask				ImagetNet		Imagen		
			IOU↑	$C_m(\%)$↑	$C_{non}(\%)$↓	rate↑	LPIPS↓	CSFID↓	LPIPS↓	FID↓	CLIPScore↑
Latent	SDEdit	3.0	-	11.6	8.4	1.38	31.1	76.5	32.1	75.2	0.238
	CycleDiffusion	5.2	-	12.1	7.3	1.66	31.1	87.5	25.8	63.0	0.246
Attention	PtP	18.2	-	16.8	12.9	1.30	-	-	42.8	85.67	0.240
	PnP	80.0	-	12.1	7.8	1.56	**27.3**	76.8	22.2	61.6	0.240
Mask	DiffEdit	64.0	33.0	19.5	8.0	2.45	27.9	70.9	29.7	58.8	0.247
	InstDiffEdit	10.8	**56.2**	**22.7**	**6.1**	**3.71**	28.6	**65.1**	**17.0**	**55.3**	**0.249**

Table 1: Comparison with existing methods on three datasets. The performance of Mask-based methods are much ahead of other methods. Moreover, InstDiffEdit leads to 70.3% on IOU and 51.4% on changing rate C_m/C_{non} compared with SOTA method DiffEdit. All experiment are conducted on a NVIDIA A100.

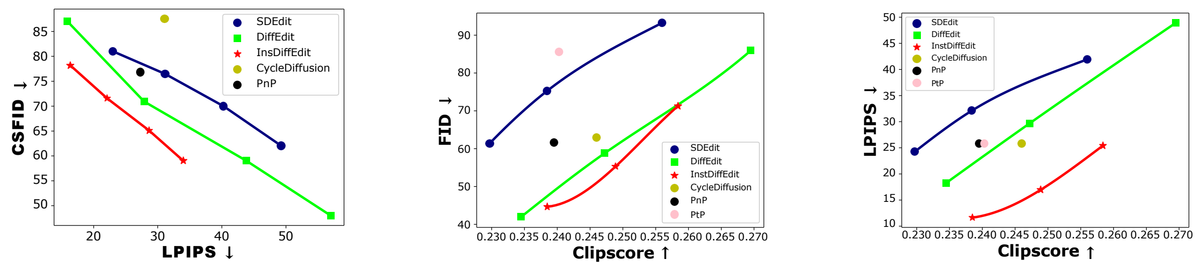


Figure 5: The trade-offs of existing methods between different metrics. We conduct experiments by using two different metrics as the independent and dependent variables respectively. The proposed InstDiffEdit has the best trade-offs.

image editing. Specifically, on our proposed Editing-Mask, DiffEdit's changing rate C_m/C_{non} far exceeds that of latent-based and attention-based methods. However, DiffEdit still requires much longer inference time. In stark contrast, our InstDiffEdit achieves up to 5 to 6 times faster inference speeds than DiffEdit, while obtaining more accurate masks. InstDiffEdit also demonstrates improvements of IOU with ground truth masks, changing rates with 70.3% and 51.4%, respectively. This strongly confirms that the proposed mask generation scheme can generate more accurate masks. Results on ImageNet show that InstDiffEdit generally outperforms DiffEdit in terms of image quality, although its LPIPS score is slightly worse . Additionally, InstDiffEdit's performance on the CSFID benchmark significantly outperforms DiffEdit by +21.1%. Similar results are also observed on the Imagen benchmark, where InstDiffEdit excels in both image quality and image-text matching, achieving a performance increase of +44.8% compared to DiffEdit on LPIPS.

We also depict the performance trade-offs between different metrics in Fig. 5. These results are achieved by tuning the hyper-parameters of each method based on the target metric. From these figures, we can first conclude that the proposed InstDiffEdit can consistently achieve the best trade-offs on all metric pairs. We observe that InstDiffEdit significantly outperforms the other methods under all conditions. These results further confirm the advantages of InstDiffEdit in terms of diffusion-based image editing.

Ablation Study. Tab. 2 presents ablation results for different settings of the noise strength r in Eq. 1 and the binarization threshold φ. In the firs row, we assess the method's performance without a mask, and the insufficient performance indicates that mask-free methods are inferior in for image editing. Secondly, as the noise strength r increases, the model obtains less information from the original image and tends to generate masks with larger areas, which results in an upward trend of C_m and C_{non} (*Line 2* vs *Line 5* vs *Line 8*). However, the IOU with ground truth mask and change rate exhibits a trend of initially increasing and then decreasing. Additionally, as the binarization threshold φ decreases, there is a tendency for the mask to cover a larger region, resulting in a similar phenomenon as discussed previously. Therefore, we select $r = 0.5$ and $\varphi = 0.2$, which yields the highest IOU and superior performance on the change rate.

r	φ	IOU↑	$C_m(\%)$↑	$C_{non}(\%)$↓	C_m/C_{non}↑
0.5	None	-	11.6	8.4	1.38
0.4	0.1	52.9	26.0	8.2	3.16
	0.2	55.7	21.8	6.0	3.63
	0.3	52.0	17.3	**4.7**	3.68
0.5	0.1	51.9	27.4	8.7	3.16
	0.2	**56.2**	22.7	6.1	3.71
	0.3	54.3	18.2	4.8	**3.81**
0.6	0.1	49.6	**28.1**	9.4	2.98
	0.2	54.6	24.3	6.7	3.60
	0.3	54.2	19.3	5.1	3.76

Table 2: Ablation study of noise strength r and binarization threshold φ on Editing-Mask.

Qualitative Analysis To obtain deep insight into InstDiffEdit, we visualize the editing results of our InstDiffEdit

Figure 6: Visualizations of the generated masks and edited images of InstDiffEdit and the compared methods. Compared with DiffEdit, the masks of InstDiffEdit are closer to the human-labeled ones. Moreover, the comparisons with the latent-based and attention-based approaches also show the merit of the instant mask in our InstDiffEdit. The red boxes refers to failed editions.

and other compared methods on Editing-Mask, as shown in Fig. 6. It can be first seen that both latent-based and attention-based approaches lack explicit constraints on the area to edit, which may result in unexpected generations. For instance, in the case of the *"German Shepherd"* image in the 4th column, DiffEdit and InstDiffEdit successfully modify the object while preserving the background, while other mask-free methods obviously change the background. However, a noteworthy disparity exists between the generated masks of DiffEdit and the human-labeled masks. Specifically, the masks produced by DiffEdit are somewhat inaccurate, and exhibits peculiar shape outlines. In contrast, our generated masks are significantly superior to those generated by DiffEdit, leading better editing results. For instance, in the case of *"speedboat"* image in the 3rd column, our mask accurately encompasses the primary object *"boat"*, whereas the mask generated by DiffEdit is non-representative. Consequently, our approach achieves successful editing, whereas DiffEdit fails to do so. These results are consistent with IOU performance presented in Tab. 1.

Conclusion

In this paper, we propose a novel and efficient method, called InstDiffEdit for diffusion-based semantic image editing. As an plug-and-play component, InstDiffEdit can be directly applied to most diffusion models without any additional training or human intervention. Experimental results not only demonstrate the superior performance of InstDiffEdit in semantic image editing tasks, but also confirm its superiority in computation efficiency, *e.g.*, up to 5 to 6 times faster than DiffEdit.

Acknowledgments

This work was supported by National Key R&D Program of China (No.2023YFB4502804) , the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U22B2051, No. U21B2037, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), the Key Research and Development Program of Zhejiang Province (No. 2022C01011), the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001), and partially sponsored by CCF-NetEase ThunderFire Innovation Research Funding (NO. CCF-Netease 202301).

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18208–18218.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Couairon, G.; Grechka, A.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022a. Flexit: Towards flexible semantic image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18270–18279.

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

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. 2672–2680.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34: 852–863.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2022. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.

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

Patashnik, O.; Garibi, D.; Azuri, I.; Averbuch-Elor, H.; and Cohen-Or, D. 2023. Localizing Object-level Shape Variations with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2303.11306*.

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

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2022. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. *arXiv preprint arXiv:2211.12572*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Lu, G.; Xu, H.; Li, Z.; Xu, C.; and Fu, Y. 2022. ManiTrans: Entity-Level Text-Guided Image Manipulation via Token-wise Semantic Alignment and Generation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10707–10717.

Wu, C. H.; and De la Torre, F. 2022. Unifying Diffusion Models' Latent Space, with Applications to CycleDiffusion and Guidance. *arXiv preprint arXiv:2210.05559*.

Xia, W.; Yang, Y.; Xue, J.-H.; and Wu, B. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2256–2265.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Zhan, F.; Yu, Y.; Wu, R.; Zhang, J.; Lu, S.; Liu, L.; Kortylewski, A.; Theobalt, C.; and Xing, E. 2021. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*.

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