Focus on Your Instruction: Fine-grained and Multi-instruction Image Editing by Attention Modulation

Qin Guo[1,2] , Tianwei Lin[2]

[1]Peking University , [2]Horizon Robotics

guoqin@stu.pku.edu.cn, tianwei.lin@horizon.cc

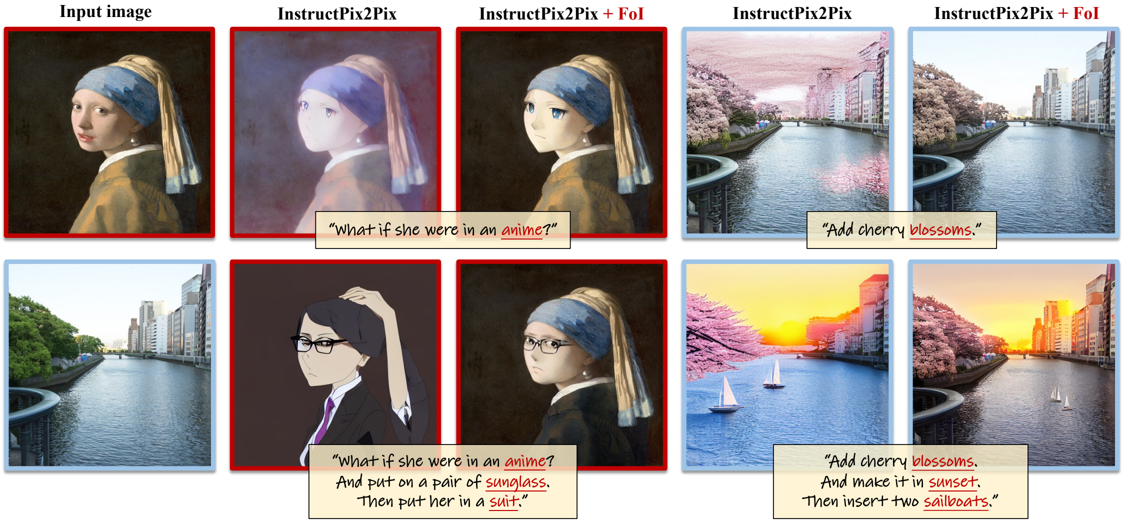


Figure 1. Models like InstructPix2Pix (IP2P) [7] can edit images with given instruction. Yet, they face challenges like over-editing and wrong editing areas, especially with multi-instruction. Our FoI utilizes inherent grounding capability of IP2P to identify precise editing regions, then focuses on them, enabling effective editing. Notably, FoI does not require extra training or test-time optimization.

Abstract

Recently, diffusion-based methods, like InstructPix2Pix (IP2P), have achieved effective instruction-based image editing, requiring only natural language instructions from the user. However, these methods often inadvertently alter unintended areas and struggle with multi-instruction editing, resulting in compromised outcomes. To address these issues, we introduce the ***Focus on Your Instruction (FoI)****, a method designed to ensure precise and harmonious editing across multiple instructions without extra training or test-time optimization. In the FoI, we primarily emphasize two aspects: (1) precisely extracting regions of interest for each instruction and (2) guiding the denoising process to concentrate within these regions of interest. For the first objective, we identify the implicit grounding capability of IP2P from the cross-attention between instruction and image, then develop an effective mask extraction method. For the second objective, we introduce a cross attention modulation module for rough isolation of target editing regions and unrelated regions. Additionally, we introduce a mask-guided disentangle sampling strategy to further ensure clear region isolation. Experimental results demonstrate that FoI surpasses existing methods in both quantitative and qualitative evaluations, especially excelling in multi-instruction editing task. The code is available at* https://github.com/guoqincode/Focus-on-Your-Instruction.

1. Introduction

Large-scale Text-to-Image (T2I) diffusion models [5, 12, 34, 42, 44, 46–48, 64] have achieved remarkable diversity and realism in image generation, garnering widespread attention. Trained on extensive image-text datasets [49], these

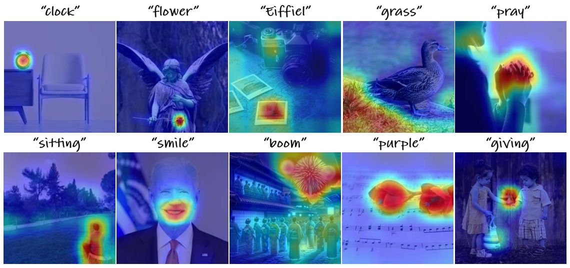


Figure 2. Visualization of cross-attention maps **in the initial denoising step** illustrates the fine-grained implicit grounding capability of IP2P [7] for **nouns**, as well as **verbs** and **adjectives**.

advanced T2I models excel in various generation tasks. However, their direct application to image editing is limited, often lacking the necessary precision for controlling specific objects or attributes within images.

When editing images, a visual creator typically begins by identifying the regions to be edited and then focuses on modifying these regions. For multiple edits, ensuring the collective result is cohesive is crucial. Despite recent remarkable advances in text-based image editing [7, 10, 17, 25, 31, 33, 34, 39, 54, 57, 67], the precisely editing of targeted areas without affecting unrelated regions remains a significant challenge. These methods often struggle to accurately pinpoint the editing areas, leading to unintended modifications in non-targeted areas and resulting in suboptimal outcomes. Furthermore, they typically struggle to simultaneously execute edits in multiple directions, further limiting their utility in complex editing tasks.

IP2P [7] offers an intuitive and fidelity-preserving approach for instruction-based image editing, bypassing the need for extensive descriptions of input and output images. However, as shown in Fig. 1, IP2P has a propensity for over-editing, which is also indicated in recent studies [21, 32]. In our analysis of IP2P, we unveil its powerful implicit grounding ability developed through training on a synthetic pairwise dataset. As shown in Fig. 2, in the cross-attention map of initial denoising step, we can observe precise alignment between keywords and their spatial locations in the image. This effective grounding extends to even adjectives and verbs. This sharply contrasts with the evolving attention maps in models like Stable Diffusion [2, 9, 17, 52]. However, as depicted in Fig. 3, while IP2P effectively locates items like a "hat", other instruction words may inadvertently affect unrelated areas, leading to unintended edits. To our knowledge, no existing methods have harnessed IP2P's potent grounding ability to enhance its editing ability.

To address the limitations of current image editing methods and align with the editing paradigm of visual creators, we introduce **Focus on Your Instruction (FoI)**, a method developed atop the IP2P framework. FoI is specifically designed for precise and harmonious multi-instruction editing, and notably, it achieves this without requiring additional training or test-time optimization. **First**, we utilize IP2P's implicit grounding ability to identify the areas of interest for each instruction. **Then**, we introduce cross-condition attention modulation, leveraging null-instruction cross-attention to modulate the cross-attention calculation with instruction, focusing each instruction on its corresponding area and implicitly reducing interference between different instructions. **Finally**, we propose a mask-guided disentangle sampling strategy, aimed at accurately separating editing and non-editing regions, disentangling the overall editing direction from preserving the original image's direction, and enhancing the model's robustness in hyperparameter selection. Experimental results demonstrate that FoI outperforms existing methods in both quantitative and qualitative evaluations, especially in multi-instruction editing tasks.

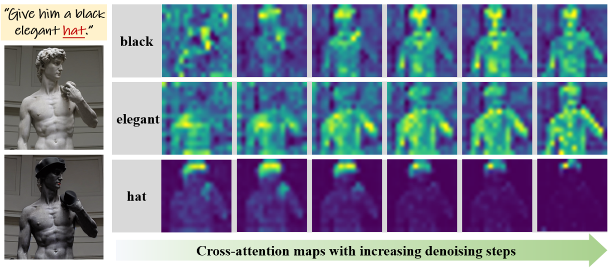

"Give him a black elegant hat."
black
elegant
hat
Cross-attention maps with increasing denoising steps

Figure 3. Visualization of cross attention maps obtained from IP2P [7], associated with different words. Two key observations: (a) the placement of *"hat"* is accurately identified early on, and (b) the attention maps for adjectives *"black"* and *"elegant"* are excessively disperse, leading to over-editing.

Our contributions can be summarized as follows:

- We introduce FoI, a method that leverages the grounding ability of IP2P for precise and harmonious multi-instruction editing, without the need for extra training or test-time optimization.
- We propose cross-condition attention modulation to ensure each instruction is focused on its corresponding area, thereby reducing interference. This method employs cross-attention with null-instruction to modulate the cross-attention calculation with instruction.
- Development of a mask-guided disentangle sampling strategy, isolating editing regions and distinguishing between editing and preserving directions.
- Demonstrated excellence of FoI in experiments, outperforming existing methods quantitatively and qualitatively, particularly in multi-instruction editing tasks.

2. Related Work

Text-guided Image Editing. Early methods mainly relied on Generative Adversarial Networks (GANs) [13, 15, 30, 40, 63], excelling in specific domains like faces and flowers, but with limited generality. Recently, methods based on diffusion models [19, 50] have showcased unprecedented prowess in image generation and editing [5, 34, 42, 44, 46, 48, 64]. SDEdit [31] leverages these models in a two-step process of noise addition and denoising to align with prompts. Imagic [25] fine-tunes the diffusion model for each image, focusing on generating variants for objects. Prompt2Prompt(P2P) [17] and PnP [54] explore attention and feature injection for improving image editing performance. Compared to P2P, PnP can directly edit real images. To adapt P2P for real image editing, Null-Text Inversion (NTI) [33] proposes updating the null text embedding for precise reconstruction and editing [18]. Blended Diffusion [3, 4] achieves local editing using user-designed masks and prompts. IP2P [7] streamlines image editing by directly applying instructions, removing the need for detailed descriptions or masks. This approach not only bypasses reconstruction flaws in inversion-based methods [10, 17, 51] but also avoids lengthy test-time optimization [33, 39, 54, 57], enhancing image fidelity.

Locating the Targeted Editing Area. Precise editing areas localization is crucial to prevent unintended image changes. Text2Live [6] utilizes CLIP [43] for optimizing additive image layers. FEAT [20] and CoralStyleCLIP [45] leverage StyleGAN's latent codes for domain-specific local editing. Diffedit [10] and Watch Your Steps [32] generate masks by contrasting different noise predictions. InstructEdit [59] and OIR [65] use text-conditioned segmentation models [26, 29] for identifying ***existing objects*** specified for editing but struggle with fine-grained editing. LPM [41] clusters self-attention maps to pinpoint objects based on cross-attention values, primarily focusing on object-level shape variations. However, most open domain visual editing works face challenges in detailed editing and preserving the original image's fidelity. For instance, with an instruction like *"put a Disney headband on her."*, the aim is to simply add the headband, yet typical methods often alter identity features or other image areas. By leveraging IP2P's implicit grounding ability, our method accurately targets the most relevant areas for each instruction, achieving finer granularity than prior methods, with only a minimal increase in computational overhead.

Multi-instruction Image Editing. A key challenge involves effectively guiding models to target specific editing areas for each instruction, while minimizing interference among instructions to ensure harmonious multi-instruction outcomes. Recent work such as EMILIE [21] focuses on iterative multi-instruction editing but overlooks IP2P's over-editing issues, mainly addressing image quality decline in successive edits. Existing instruction-based methods [7, 11, 14, 66, 68] often struggle with multi-instruction tasks. Contemporary methods [58, 65] in multi-object editing, using inversion-based techniques [51], focus mainly on object-level replacement. These methods, requiring complex optimization, struggle with fine-grained editing and tend to be time-consuming. In contrast, our approach avoids additional training or test-time optimization, and can be easily integrated with existing instruction-based models.

3. Preliminaries

InstructPix2Pix. Given an image I, IP2P edits it following given editing instruction T. IP2P undergoes supervised training on a dataset synthesized using P2P [17] and GPT-3 [8]. Each entry in the dataset includes the original image I, the editing instruction T, and the corresponding edited result I_{e}. IP2P is constructed upon on the Stable Diffusion framework [46], employing a VQ-VAE [55] with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ to enhance efficiency and quality. For training, noise $\epsilon \sim \mathcal{N}(0,1)$ is added to $z = \mathcal{E}(I_{\text{e}})$ to create noisy latent z_t, with the noise level set by a random timestep $t \in T$. The denoiser, ϵ_θ, initially with Stable Diffusion weights [46], is fine-tuned to minimize $\mathbb{E}_{I_{\text{e}},I,\epsilon,t}\big[\|\epsilon - \epsilon_\theta(z_t,t,I,T)\|_2^2\big]$. Conditions are intermittently omitted during training [28] by setting $I = \emptyset_I$ or $T = \emptyset_T$. The vanilla IP2P score estimate is as follows:

$$\begin{aligned}\tilde{\epsilon}_\theta(z_t,t,I,T) &= \epsilon_\theta(z_t,t,\emptyset_I,\emptyset_T)\\ &+ s_I\big(\epsilon_\theta(z_t,t,I,\emptyset_T) - \epsilon_\theta(z_t,t,\emptyset_I,\emptyset_T)\big)\\ &+ s_T\big(\epsilon_\theta(z_t,t,I,T) - \epsilon_\theta(z_t,t,I,\emptyset_T)\big)\end{aligned} \tag{1}$$

Studies [7, 16, 32] highlight the importance of balancing image guidance s_I and text guidance s_T. An increase in s_I preserves image details but reduces the impact of instructions, while an increase in s_T poses risks of over-editing. Consequently, $\epsilon_\theta(z_t,t,I,\emptyset_T)$ estimates scores for image preservation, and $\epsilon_\theta(z_t,t,I,T)$ for applying edits.

Cross Attention in InstructPix2Pix. IP2P incorporates textual features in image editing through a cross-attention mechanism [56]. This process generates cross-attention maps $\mathcal{A}_t \in \mathbb{R}^{r\times r\times N}$ at each denoising step t for every token (N tokens tokenized using CLIP [43]'s tokenizer) in the input instruction, where $r \in \{64, 32, 16, 8\}$ [2, 9, 17]. Because IP2P integrates the original image into the input channels of its U-Net, the behavior of its attention mechanism exhibits distinctions compared to Stable Diffusion [46]. We denote the cross attention map in $\epsilon_\theta(z_t,t,I,T)$ as $\mathcal{A}_{t,\text{ins}}$.

4. Method

Given the input image I and the composite instruction T, composed of k sub-instructions $\{T_1, T_2, \ldots, T_k\}$, our goal is to edit I with **(1)** precise execution of each sub-instruction in T, and **(2)** harmonious execution of T as a whole. We believe the core challenge here is ***how to precisely directing instructions towards their corresponding areas of interest.***

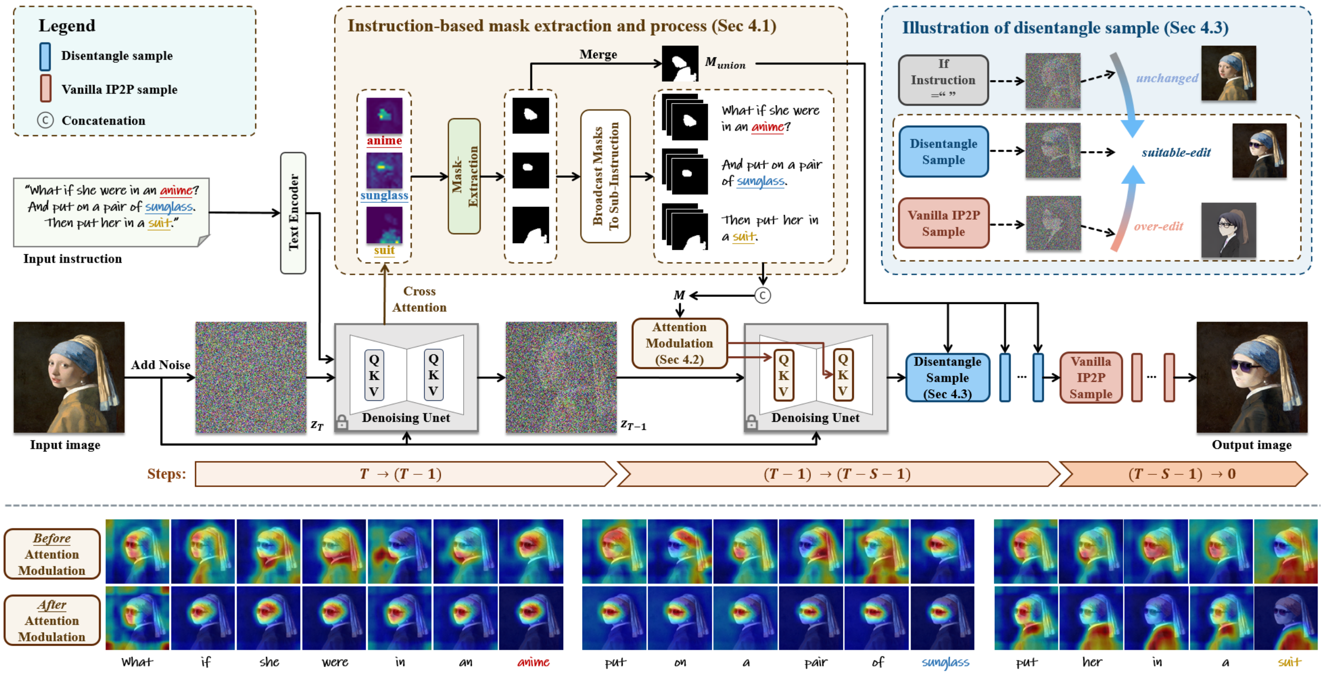


Figure 4. **Framework of FoI.** FoI is designed to perform precise single-instruction edits and coordinated multi-instruction edits, all within a single forward pass. Firstly, a unique mask for each sub-instruction is extracted at the start of the denoising step, as described in Sec. 4.1. Next, we use cross-condition attention modulation to focus each instruction on its interest area and reduce interference, elaborated in Sec. 4.2(the bottom figure illustrates the cross-attention map before and after attention modulation). Finally, a disentangle sample method isolates editing areas, detailed in Sec. 4.3.

To solve this challenge, we propose FoI, with overall framework illustrated in Fig. 4. In this section, we first discuss how to find precise area of interest for each sub-instruction (Sec. 4.1). Then, we introduce how to guide the denoising process to proper direction where each instruction focus on its own interest area, with cross-attention modulation (Sec. 4.2) and disentangle sampling strategy (Sec. 4.3).

4.1. Extracting Instruction-Based Masks

Inspired by the segmentation capabilities of large-scale diffusion models [23, 53, 60, 61], our analysis of IP2P uncovers its precise location-finding ability in early denoising steps, evident from cross-attention maps in Fig. 2. Demonstrated in Fig. 3, IP2P quickly identifies where objects, like “hat” should be placed. We harness this robust grounding capability to extract areas of interest for each instruction from IP2P’s cross-attention maps.

Previous studies [2, 9, 17, 62] have shown that attention maps with a resolution of 16×16 capture the most detailed semantic information. Accordingly, we use attention maps with a resolution of $r = 16$ for extracting masks.

In each sub-instruction T_i, we identify a keyword e_i, which represents either the target object for editing, an object to be added, or an object inferred from the context, as specified in the sub-instruction. We begin by applying a Gaussian filter [9] to the corresponding cross-attention map $\mathcal{A}_t[\boldsymbol{e_i}] \in \mathbb{R}^{r \times r}$. This step ensures that each patch in the map is a linear combination of its neighboring patches in the original map. We then use a direct and effective algorithm for mask extraction, enhancing the cross-attention map $\mathcal{A}_t[\boldsymbol{e_i}]$ iteratively. The algorithm operates through a sequence of operations, repeated γ times. In each cycle, the map is squared and then normalized via min-max scaling to the $[0, 1]$ range. This iterative approach is designed to incrementally heighten the contrast between target regions and surrounding areas, as detailed in the following equation:

$$\mathcal{A}_t[\boldsymbol{e_i}] = \underbrace{\text{norm}\left(\text{norm}\left(\cdots\text{norm}\left(\mathcal{A}_t[\boldsymbol{e_i}]^2\right)\cdots\right)^2\right)^2}_{\gamma\text{ times}} \quad (2)$$

Here, norm denotes the min-max normalization process, scaling the values within the map to a [0,1] range. Upon completing the γ iterations, we apply a threshold τ to compute the mask $\mathcal{M}_{e_i} = \mathbb{1}(\mathcal{A}_t[\boldsymbol{e_i}] \geq \tau)$. This mask, denoting the area of interest for the i-th sub-instruction, has dimensions $\in \mathbb{R}^{r \times r}$. Fig. 4 displays the effective results of our method in extracting masks for each sub-instruction.

4.2. Cross Condition Attention Modulation

For fine-grained editing, confining each instruction within its mask is essential. We introduce the cross-condition attention modulation. This method utilizes the cross-attention map with null-instruction to modulate the cross-attention calculation with instruction, thereby reducing the impact of instruction on irrelevant areas and decreasing interference between different instructions when multiple instructions are present. To be specific, we preserve the masked region's attention in the computation of $\epsilon_\theta(z_t, t, I, T)$, while substituting attention in the external regions with that from $\epsilon_\theta(z_t, t, I, \emptyset_T)$. The modified cross attention function is defined as:

$$\mathcal{A}'_{t,ins} = \text{softmax}\left(\frac{(\mathcal{X} + \Delta\mathcal{X}) \odot \mathcal{M} + \mathcal{Y} \odot (1 - \mathcal{M})}{\sqrt{d}}\right) \tag{3}$$

Here, d represents the latent projection dimension. The terms are defined as follows:

$$\mathcal{X} = Q_{I,T} K_{I,T}^T, \tag{4}$$

$$\mathcal{Y} = Q_{I,\emptyset_T} K_{I,\emptyset_T}^T, \tag{5}$$

$$\Delta\mathcal{X} = \boldsymbol{\alpha} \odot \xi(t) \odot (\max(Q_{I,T} K_{I,T}^T) - Q_{I,T} K_{I,T}^T) \tag{6}$$

where $Q_{I,T}$ and $K_{I,T}$ are the query and Key in $\epsilon_\theta(z_t, t, I, T)$ respectively, and $Q_{I,\emptyset_T}$ and $K_{I,\emptyset_T}$ are the query and key in $\epsilon_\theta(z_t, t, I, \emptyset_T)$ respectively. The attention mask $\mathcal{M}$ is constructed by initially ***broadcasting*** the mask $\mathcal{M}_{e_i}$ of each keyword across its corresponding sub-instruction T_i. Subsequently, these broadcasted masks are concatenated for all sub-instructions, resulting in an initial dimension of $\mathcal{M}$ being $\mathbb{R}^{(r \times r) \times N}$. This mask is then adaptively interpolated across each cross-attention layer.

We employ $\Delta\mathcal{X}$ to subtly enhance attention values within the mask. This allows for precise control over the relative strengths of different sub-instructions, enabling fine-grained control over the intensity of each sub-instruction. This is achieved by selectively adjusting the values in the coefficient vector $\boldsymbol{\alpha}$, which is initially set to all ones. Through strategic modifications to specific values within $\boldsymbol{\alpha}$, we can finely tune the intensity of each sub-instruction. This method directs attention within the mask and enables flexible adjustment of each sub-instruction's relative intensity, ensuring focused and controlled effects during the editing process.

In Eq. (6), the timestep-related weight term is:

$$\xi(t) = 0.05 * t^4 \tag{7}$$

and the timestep $t \in [0, 1]$ has been normalized.

After mask extraction, we apply cross-condition attention modulation across all remaining denoising steps. The bottom part of Fig. 4 illustrates the cross-attention maps before and after the application of cross-condition attention modulation. It's observable that compared to *before modulation*, each sub-instruction becomes more concentrated within its respective area of interest.

4.3. Mask Guided Disentangle Sample

While restricting the area of interest at the cross-attention level is useful, it is insufficient for fine-grained editing due to the low resolution of semantically rich layers in cross attention [9, 56, 62]. Additionally, disentangling the different noise estimates in Eq. (1) is challenging, leading to a lack of robustness in the arbitrary selection of s_I and s_T. Therefore, we suggest modifying the noise estimation to isolate the editing area from irrelevant regions during the sampling process and disentangle the directions of editing and preserving the original image. To achieve this, we first combine the masks corresponding to all sub-instructions to obtain $\mathcal{M}_{union}$:

$$\mathcal{M}_{union} = Upsample(\bigvee_i \mathcal{M}_{e_i}) \tag{8}$$

where $Upsample$ denotes the operation of upsampling the mask to match the resolution of the latent space. Following this, new score estimates are formulated:

$$\begin{aligned}\tilde{\epsilon}_\theta(z_t, t, I, T) &= \epsilon_\theta(z_t, t, \emptyset_I, \emptyset_T) \\ &+ s_I\big(\epsilon_\theta(z_t, t, I, \emptyset_T) - \epsilon_\theta(z_t, t, \emptyset_I, \emptyset_T)\big) \\ &+ s_T\big(\epsilon_\theta(z_t, t, I, T) - \epsilon_\theta(z_t, t, I, \emptyset_T)\big) \odot \mathcal{M}_{union}\end{aligned} \tag{9}$$

We refer to the sampling using the above score estimates as disentangle sampling. In practice, we employ the disentangle sampling for the initial 75% steps, and switch to the standard IP2P sampling for the remaining 25%,.

5. Experiments

5.1. Experimental Settings

Dataset. For single-instruction editing, we filter 5,000 localized edit-type images from the IP2P dataset [7] using GPT4 [35], each tagged with specific object edits. For multi-instruction editing, we gather 100 real images. For each image, GPT-4V(ision) [1, 35, 36] is used to create 2-4 instructions, along with original and target descriptions, and marking the objects to edit.

Metrics. For evaluation, we use *CLIP image similarity* [43] and *Dinov2 image similarity* [37] to measure the cosine similarity between edited and original images. *CLIP text-image direction similarity* [13] evaluates how image changes correspond with changes in their captions. Additionally, *PickScore* [27] evaluates image fidelity based on learned human preferences.

Baseline models. We make comparisons with the state-of-the-art (SOTA) image editing methods, including Diffedit [10], NTI+P2P [33], IP2P [7], MagicBrush [66], and InstructDiffusion (InsDiff) [14]. Diffedit identifies regions for editing based on the differences between the noise

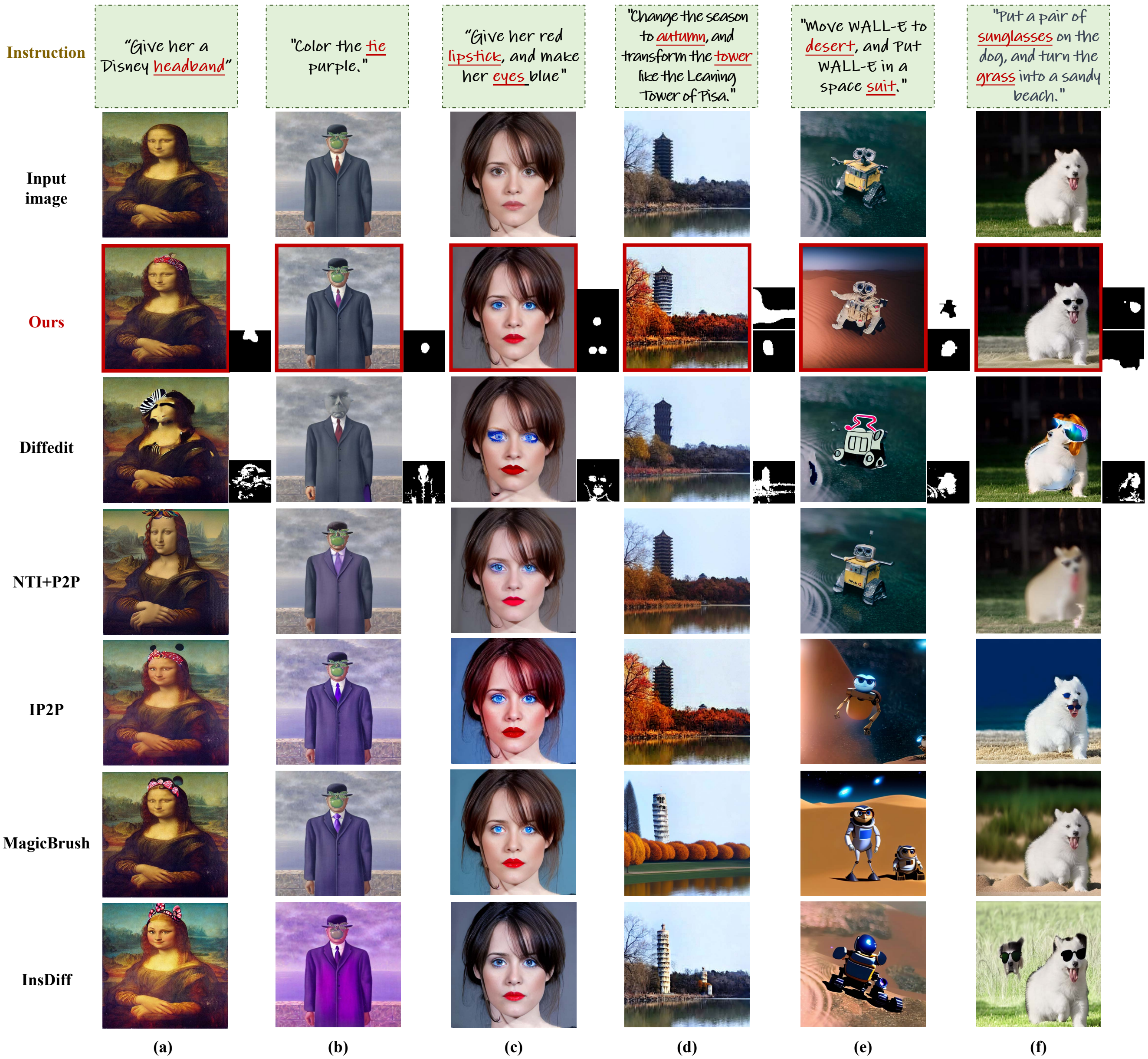

Instruction
"Give her a Disney headband"
"Color the tie purple."
"Give her red lipstick, and make her eyes blue"
"Change the season to autumn, and transform the tower like the Leaning Tower of Pisa."
"Move WALL-E to desert, and Put WALL-E in a space suit."
"Put a pair of sunglasses on the dog, and turn the grass into a sandy beach."
Input image
Ours
Diffedit
NTI+P2P
IP2P
MagicBrush
InsDiff
(a)
(b)
(c)
(d)
(e)
(f)

Figure 5. **Qualitative comparisons.** We provide all baselines with their desired input formats. From top to bottom: input image, our method, Diffedit [10], NTI+P2P [33], IP2P [7], MagicBrush [66], InsDiff [14]. The texts at the top of the images represent edit instructions. Inputs for Diffedit and NTI+P2P include the original and target captions. Additionally, we present masks that highlight the regions of interest identified by FoI and Diffedit, located on the lower right side of the results, organized in accordance with the sequence of sub-instructions. Compared with baseline models, FoI can accurately edit regions of interest.

predictions of original and target prompts. NTI+P2P extends P2P [17] to realize the editing of real images, representing inversion-based image editing methods. IP2P is the basic model of instruction-based editing methods. MagicBrush fine-tunes IP2P on its own high-quality constructed dataset. InsDiff uses the same model structure as IP2P but trains a generalist model on multiple datasets. The closely related Watch Your Steps [32] lacks an available implementation, precluding direct comparison.

Implementation details. In all our experiments, we utilize the pretrained IP2P model [7] with freeze weight. We use the Euler ancestral sampler [24] with a total of 100 denoising steps. The default settings of $s_I = 1.5$ and $s_T = 7.5$ are used unless specified otherwise. For mask extraction described in Sec. 4.1, it is only performed during the first denoising step. The threshold τ is randomly sampled from the range $[0.4, 0.7]$, and the hyperparameter γ is set to 3. Following previous work [32], the initial noise is generated by adding 80% noise to the original image, thus the actual number of denoising steps is 80 in FoI.

5.2. Main Results

Qualitative evaluation. We show some qualitative experimental results in Fig. 5, From our experiments, we observe the following: Firstly, Diffedit and NTI+P2P often overlook sub-tasks in complex editing scenarios. For example, in Fig. 5 (d), the tower edit is ignored; in Fig. 5 (e), WALL-E remains in water instead of desert; and in Fig. 5 (f), the grass does not change to a sandy bench. These models also lead to unwanted modifications within the editing areas, such as over-modification of Mona Lisa's facial features (Fig. 5 (a)), incorrect facial edits (Fig. 5 (b)), and excessive changes to the dog in Fig. 5 (f). Secondly, instruction-based methods like IP2P, MagicBrush, and InsDiff tend to over-edit. This is observable as IP2P and MagicBrush modify beyond the intended headband area in Fig. 5 (a), affecting Mona Lisa's identity. In Fig. 5 (b), they alter the entire image to purple and in Fig. 5 (c), they change the background to blue. Over-editing is also evident in Fig. 5 (d) and (e) by MagicBrush and InsDiff, while IP2P misses the tower edit in Fig. 5 (d). All three methods cause excessive edits to WALL-E in Fig. 5 (e), leading to a significant departure from the original image. IP2P and MagicBrush fail to appropriately add sunglasses to the dog in Fig. 5 (f), and InsDiff creates an incomplete additional dog head wearing sunglasses. Despite the instruction only requesting the glass to change to a sandy bench, IP2P, MagicBrush, and InsDiff also modify the background.

Compared to Diffedit, our method extracts masks for the area of interest with greater precision and detail, ensuring higher quality and more fine-grained editing. This further evidences the implicit grounding capability of IP2P in enhancing fine-grained editing.

Our method provides more nuanced editing capabilities and does not affect unrelated areas, outperforming baseline models, especially in scenarios with multi-instruction.

Quantitative evaluation. As illustrated in Tab. 1, in single-instruction and multi-instruction evaluations, we achieve state-of-the-art results in CLIP image similarity, Dinov2 image similarity and PickScore, demonstrating that our method best aligns with human perception in terms of fidelity to the original and edited images. Notably, our method shows vastly improved performance over baseline models for the multi-instruction editing task, demonstrating the superiority of our method when faced with complex editing instructions.

In the CLIP direction similarity metric, our method scores lower than MagicBrush [66] and InsDiff [14] because they tend towards over-editing, making larger changes to the input image in the direction of the instructions, whereas our method focuses on necessary edits and minimizes effects on irrelevant areas. CLIP [43] itself has difficulty perceiving fine-grained changes [38], so this also proves that we perform more fine-grained editing. Over-editing would lead to reduced CLIP image similarity and Dinov2 image similarity, and increased CLIP direction similarity. Our method balances the preservation of details in the original image and execution of editing instructions.

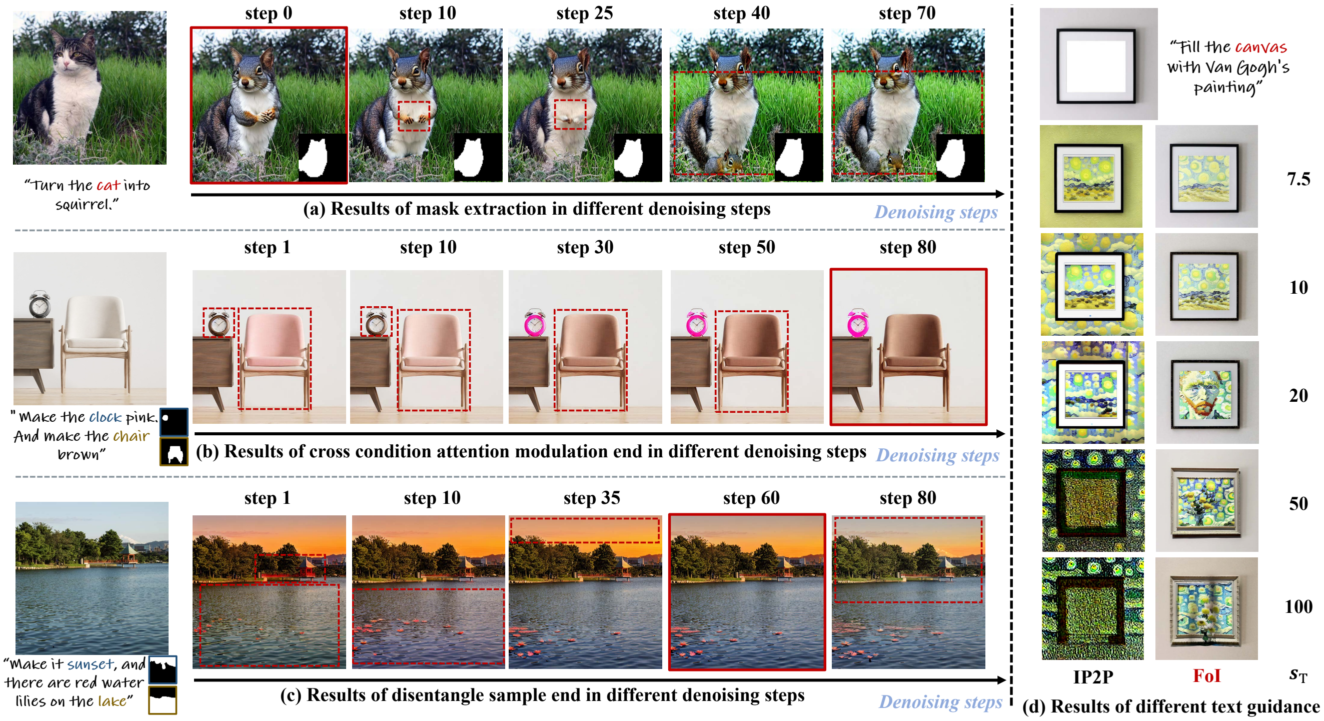
	Method	CLIP-I	Dino-I	CLIP-D	PickScore
Single-Instruction	Diffedit [10]	0.8627	0.7916	0.0844	0.0639
	NTI+P2P [33]	0.8522	0.7928	0.0981	0.0951
	IP2P [7]	0.8605	0.8264	0.1685	0.1353
	MagicBrush [66]	0.9178	0.8702	0.1934	0.1780
	InsDiff [14]	0.8755	0.8612	**0.2064**	0.1377
	FoI (ours)	**0.9402**	**0.9277**	0.1699	**0.3901**
Multi-Instruction	Diffedit [10]	0.8505	0.7529	0.0629	0.0616
	NTI+P2P [33]	0.8560	0.7526	0.0865	0.0332
	IP2P [7]	0.8769	0.8369	0.1605	0.1059
	MagicBrush [66]	0.8609	0.8291	**0.1807**	0.1591
	InsDiff [14]	0.8439	0.7938	0.1785	0.1325
	FoI (ours)	**0.9255**	**0.9159**	0.1685	**0.5077**

Table 1. **Quantitative comparisons.** We compare our model with baseline models in terms of CLIP image similarity, Dinov2 image similarity, CLIP direction similarity, and PickScore. Our method achieves state-of-the-art results in image similarity and PickScore. Because our method aims to minimize over-editing, the CLIP direction similarity is lower than MagicBrush [66] and InsDiff [14], which tend to over-edit.

	Single-Instruction		**Multi-Instruction**	
	Instruction Align	**Image Align**	**Instruction Align**	**Image Align**
Diffedit [10]	9.42%	10.42%	0.75%	3.08%
NTI+P2P [33]	9.5%	16.58%	0.42%	4.25%
IP2P [7]	12.83%	10.92%	3.08%	3.92%
MagicBrush [66]	**23.17%**	22.83%	9.42%	4.75%
InsDiff [14]	21.92%	11.75%	5.50%	2.67%
FoI (ours)	**23.17%**	**27.5%**	**80.83%**	**81.33%**

Table 2. **Human preference study.** FoI outperforms baseline models in both instruction- and image-alignment, and achieves a huge advantage in multi-instruction measures.

Human Preference Study. For single and multi instruction edits, we conduct a human preference study using 20 images for each category, comparing our FoI with Diffedit, NTI+P2P, IP2P, MagicBrush, and InsDiff. The study includes 60 participants. For instruction alignment, participants are asked to choose the method that best matched the editing effect of the instruction. For image alignment, they select the method that best preserved the original image details (i.e., no changes occurred in unrelated areas). As indicated in Tab. 2, our FoI method is favored over the baseline methods for both single and multi instruction edits, with a significant preference gap observed in the multi-instruction editing scenarios, over 80% of participants per-


step 0
step 10
step 25
step 40
step 70
"Turn the cat into squirrel."
Denoising steps
step 1
step 10
step 30
step 50
step 80
"Make the clock pink. And make the chair brown"
Denoising steps
step 1
step 10
step 35
step 60
step 80
"Make it sunset, and there are red water lilies on the lake"
Denoising steps
"Fill the canvas with Van Gogh's painting"
7.5
10
20
50
100
IP2P
FoI
s_T

(a) Results of mask extraction in different denoising steps

(b) Results of cross condition attention modulation end in different denoising steps

(c) Results of disentangle sample end in different denoising steps

(d) Results of different text guidance

Figure 6. **Ablation study of different components.** The editing effect within the red box is poor. **(a)** Editing results and corresponding mask of mask extraction at different denoising steps. **(b)** Results of ending cross condition attention modulation at different denoising steps. **(c)** Results of ending disentangle sampling at different denoising steps. **(d)**. Robustness of our method compared to IP2P with fixed s_I across various s_T values. For (b) and (c), the mask extracted for the editing area are displayed to the right of each instruction.

ceive the editing quality and fidelity of FoI to surpass that of the baseline models. This underscores the superiority of FoI in precise and high-quality editing tasks.

More results are available in Appendix C.

5.3. Ablation Study

Mask Extraction Steps. As illustrated in Fig. 6 (a), searching for the mask over extended time steps does not enhance the outcomes; rather, it results in the inadvertent editing of unrelated areas in more denoising steps, and also diminishes the effectiveness within the intended editing regions.

Cross Condition Attention Modulation. As shown in Fig. 6 (b), halting cross-attention modulation at various denoising steps can have different effects. Early termination might cause instructions to affect unrelated areas, particularly in multi-instruction scenarios, where it can also disrupt other instructions. Notably, the effectiveness of sub-instructions increases with the number of steps conducted.

Disentangle Sample. As illustrated in Fig. 6 (c), even with the application of cross-attention modulation, the use of broad adjectives can inadvertently result in minor modifications to irrelevant areas. The Disentangle Sample method alleviates the issue of insufficient granularity in attention modulation, effectively separating the editing areas from irrelevant regions. However, using disentangle sampling for all steps can lead to suboptimal outcomes. For instance, at *step 80* in Fig. 6 (c), the *"Make it sunset."* instruction creates a more fragmented effect compared to the smoother result at *step 60*. Moreover, as demonstrated in Fig. 6 (d), where we set $s_I = 1.5$ and progressively increase s_T, unlike previous methods [7, 16, 21, 32] that require precise tuning of the balance between s_I and s_T, our approach maintains this balance with greater robustness.

Quantitative evaluation and analysis from our ablation study will be detailed in Appendix A.

6. Conclusion

We propose FoI, a tuning-free method that empowers the pretrained IP2P model to execute precise single-instruction edits as well as multi-instruction edits. We discover the IP2P model's implicit grounding capability and extract masks corresponding to each instruction. Furthermore, we utilize these masks for cross-condition attention modulation, which confines instructions within their respective masks while reducing interference between different instructions. Finally, we introduce disentangle sampling, designed to isolate editing areas from irrelevant regions and disentangle the directions of editing and preserving the original image. Our approach demonstrates exceptional performance in both qualitative and quantitative experiments.

References

[1] Chatgpt can now see, hear, and speak. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak, 2023. 5

[2] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. arXiv preprint arXiv:2306.14544, 2023. 2, 3, 4

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18208–18218, 2022. 3

[4] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. ACM Transactions on Graphics (TOG), 42 (4):1–11, 2023. 3

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324, 2022. 1, 3

[6] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In European conference on computer vision, pages 707–723. Springer, 2022. 3

[7] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In arXiv, 2023. 1, 2, 3, 5, 6, 7, 8

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020. 3

[9] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG), 42(4):1–10, 2023. 2, 3, 4, 5

[10] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427, 2022. 2, 3, 5, 6, 7

[11] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. arXiv preprint arXiv:2309.17102, 2023. 3

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022. 1

[13] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. ACM Transactions on Graphics (TOG), 41(4):1–13, 2022. 2, 5

[14] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, et al. Instructdiffusion: A generalist modeling interface for vision tasks. arXiv preprint arXiv:2309.03895, 2023. 3, 5, 6, 7

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020. 2

[16] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 3, 8

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022. 2, 3, 4, 6

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022. 3

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In arXiv, 2020. 3

[20] Xianxu Hou, Linlin Shen, Or Patashnik, Daniel Cohen-Or, and Hui Huang. Feat: Face editing with attention. arXiv preprint arXiv:2202.02713, 2022. 3

[21] KJ Joseph, Prateksha Udhayanan, Tripti Shukla, Aishwarya Agarwal, Srikrishna Karanam, Koustava Goswami, and Balaji Vasan Srinivasan. Iterative multi-granular image editing using diffusion models. arXiv preprint arXiv:2309.00613, 2023. 2, 3, 8

[22] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. arXiv preprint arXiv:2310.01506, 2023. 12

[23] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. arXiv preprint arXiv:2306.09316, 2023. 4

[24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems, 35:26565–26577, 2022. 6

[25] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6007–6017, 2023. 2, 3

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023. 3

[27] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. arXiv preprint arXiv:2305.01569, 2023. 5

[28] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with

composable diffusion models. In European Conference on Computer Vision, pages 423–439. Springer, 2022. 3

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499, 2023. 3

[30] Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. Deltaedit: Exploring text-free training for text-driven image manipulation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 2

[31] 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. arXiv preprint arXiv:2108.01073, 2021. 2, 3

[32] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G Derpanis, and Igor Gilitschenski. Watch your steps: Local image and scene editing by text instructions. arXiv preprint arXiv:2308.08947, 2023. 2, 3, 6, 7, 8

[33] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6038–6047, 2023. 2, 3, 5, 6, 7

[34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021. 1, 2, 3

[35] OpenAI. Gpt-4 technical report, 2023. 5

[36] OpenAI. Gpt-4v(ision) system card. 2023. 5

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023. 5

[38] Roni Paiss, Hila Chefer, and Lior Wolf. No token left behind: Explainability-aided image classification and generation. In European Conference on Computer Vision, pages 334–350. Springer, 2022. 7

[39] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In ACM SIGGRAPH 2023 Conference Proceedings, pages 1–11, 2023. 2, 3

[40] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 2085–2094, 2021. 3

[41] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. arXiv preprint arXiv:2303.11306, 2023. 3

[42] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023. 1, 3

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In ICML, 2021. 3, 5, 7

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1 (2):3, 2022. 1, 3

[45] Ambareesh Revanur, Debraj Basu, Shradha Agrawal, Dhwanit Agarwal, and Deepak Pai. Coralstyleclip: Co-optimized region and layer selection for image editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12695–12704, 2023. 3

[46] 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, 2022. 1, 3

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22500–22510, 2023.

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35:36479–36494, 2022. 1, 3

[49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, 35:25278–25294, 2022. 1

[50] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning, pages 2256–2265. PMLR, 2015. 3

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020. 3

[52] Raphael Tang, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. arXiv preprint arXiv:2210.04885, 2022. 2

[53] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. arXiv preprint arXiv:2308.12469, 2023. 4

[54] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1921–1930, 2023. 2, 3

[55] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. Advances in neural information processing systems, 30, 2017. 3

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017. 3, 5

[57] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22532–22541, 2023. 2, 3

[58] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. arXiv preprint arXiv:2309.15664, 2023. 3

[59] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. arXiv preprint arXiv:2305.18047, 2023. 3

[60] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. arXiv preprint arXiv:2308.06160, 2023. 4

[61] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. arXiv preprint arXiv:2303.11681, 2023. 4

[62] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7452–7461, 2023. 4, 5

[63] Zipeng Xu, Tianwei Lin, Hao Tang, Fu Li, Dongliang He, Nicu Sebe, Radu Timofte, Luc Van Gool, and Errui Ding. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18229–18238, 2022. 3

[64] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. arXiv preprint arXiv:2305.18295, 2023. 1, 3

[65] Zhen Yang, Dinggang Gui, Wen Wang, Hao Chen, Bohan Zhuang, and Chunhua Shen. Object-aware inversion and reassembly for image editing, 2023. 3

[66] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. arXiv preprint arXiv:2306.10012, 2023. 3, 5, 6, 7

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023. 2

[68] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. arXiv preprint arXiv:2303.09618, 2023. 3