Zero-Shot Reference-Based Image Editing Without Training or Guidance

Kelvin Li
UC Berkeley
kelvin.li.jm@berkeley.edu

Jorge Diaz Chao
UC Berkeley
jdiazchao@berkeley.edu

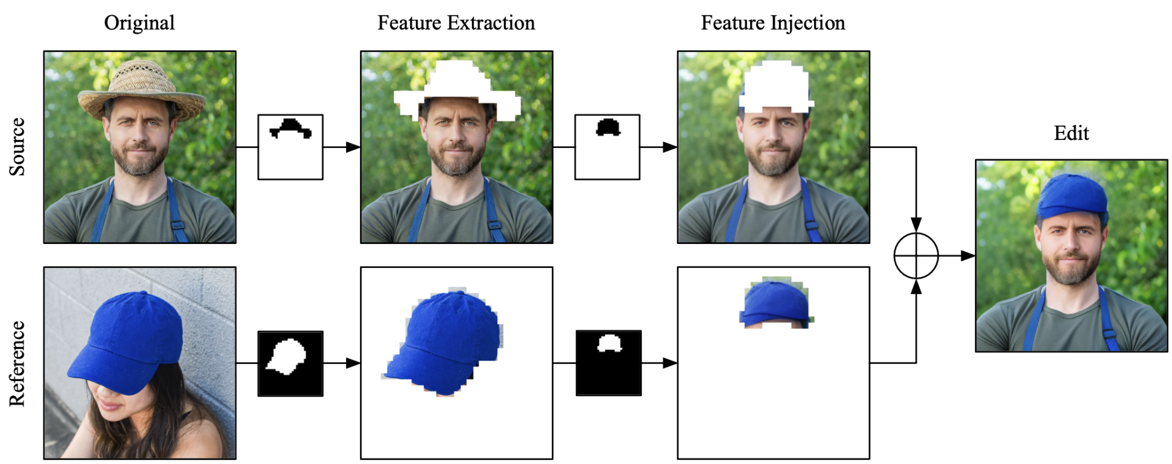

Figure 1. A visualization of our method. We extract desired information from the source and reference images and inject these information into the appropriate spatial regions of the edit image to generate the final output.

Abstract

Recent advancements in text-to-image diffusion models have significantly improved text-conditioned image generation and editing. However, expressing fine-grained modifications through text prompts alone remains inherently limited. To address this, reference-based editing methods have been explored, but they often require test-time fine-tuning, additional training, or expensive test-time computations (e.g., guidance), limiting their practicality and generalizability. We introduce a zero-shot, training-free and guidance-free reference-based image editing framework that allows for a wide range of precise edits by leveraging reference images. Our key innovation is the Hybrid Attention mechanism, which replaces standard self-attention in the diffusion U-Net with a novel masked cross-image attention approach. This mechanism enables controlled extraction and injection of spatial features from reference images into a target image during sampling. Additionally, we propose the use of automatic masks derived from the cross-attention maps of the diffusion UNet to define region selection, removing the need for manual input.

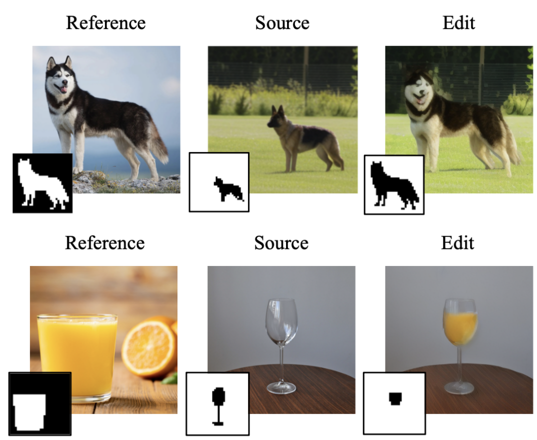

Figure 2. A preview of our results.

1. Introduction

The rapid advancement of text-to-image (T2I) generative models has enabled high-quality image generation and editing from natural language prompts. However, while these models produce impressive results, prompt-driven generation remains fundamentally limited in capturing a user's exact intent. Expressing nuanced edits through text alone is challenging, as natural language descriptions often fail to specify the fine details and localized modifications required for precise image editing.

To address this limitation, many reference-based image editing methods have emerged. Instead of relying solely on text, reference images provide additional visual guidance, allowing for more controlled and targeted edits. However, existing methods typically fall into one of the following categories:

- **Test-time fine-tuning** methods adapt the diffusion U-Net [12, 19] or an auxiliary model [20] to a given image by optimizing its parameters during inference. This approach allows the model to tailor directly to the specific image. However, these methods are computationally expensive at test-time and hence impractical for real-time applications.
- **Pretraining-based** methods [2, 3, 6, 8, 13, 24–27] train base diffusion models on large-scale datasets with paired images for various editing tasks. By leveraging extensive training, these models learn how to perform specific edits, enabling faster inference without requiring optimization at test time. However, this approach is constrained by the availability of paired training data, limiting the model's ability to generalize beyond predefined tasks such as object insertion or style transfer.
- **Guidance-based** methods [5, 16, 18] optimize the latent embedding throughout the diffusion process using specialized score functions, enabling more precise control compared to text-only inputs. While these approaches do not require additional training, they impose a substantial computational cost during inference, increasing both runtime and GPU memory consumption due to the need for backpropagation through the diffusion model.

Existing methods that are both training-free and guidance-free are limited in scope, often supporting only basic edits like object insertion [7, 14, 23] or global style/structure/appearance transfer [1, 4, 11, 15], failing to provide a generalizable framework for more diverse, complex editing tasks.

To this end, we introduce a novel training-free, zero-shot, and guidance-free framework for reference-based image editing that extends beyond basic edits, supporting a wide range of localized, and high-fidelity modifications in a fully automatic manner. Furthermore, our framework provides deeper insights into the role of attention layers in diffusion models, shedding light on their semantic reasoning capabilities.

2. Preliminaries

2.1. DDIM Inversion

Denoising Diffusion Probabilistic Models (DDPM) [10, 21] are trained to optimize the objective:

$$\min_{\theta} \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,I),t\sim\text{Uniform}(1,T)} \left\|\epsilon - \epsilon_\theta(z_t, t, C)\right\|_2^2 .$$

Here, the diffusion model learns to predict the noise component ϵ given a noisy latent z_t, timestep t, and conditioning information C. During sampling, the noise is gradually removed by sequentially predicting it across T timesteps, transforming a noisy latent z_T to a clean latent z_0.

Denoising Diffusion Implicit Models (DDIM) [22] extend the diffusion framework by introducing a deterministic sampling process. Unlike the stochastic nature of DDPMs, DDIM ensures consistent transformations between noisy latents and clean latents. This makes it particularly useful for latent reconstruction and editing, as it preserves structural and semantic consistency throughout the sampling trajectory. Given an initial noisy latent z_T, the DDIM sampling process iteratively refines it into a clean latent z_0. This is achieved through a deterministic update rule:

$$z_t = \frac{1}{\sqrt{\alpha_{t+1}}}\left(z_{t+1} - \frac{1-\alpha_{t+1}}{\sqrt{1-\bar{\alpha}_{t+1}}}\cdot \epsilon_\theta(z_{t+1}, t+1)\right).$$

DDIM inversion [17] maps a starting latent z_0 to a corresponding noisy latent z_T. The DDIM inversion follows a deterministic update rule:

$$z_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\hat{z}_0 + \sqrt{1-\bar{\alpha}_{t+1}}\cdot\epsilon_\theta(z_t, t),$$

where the estimated clean latent $\hat{z}_0$ is given by:

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\cdot\epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}}.$$

Through iterative application of the DDIM inversion update rule, an initial clean latent z_0 is transformed into progressively noisier latents z_t, with the final noisy latent z_T capturing the structural and semantic properties of the original latent while aligning with the learned diffusion process. This latent representation serves as the starting point for downstream tasks such as image reconstruction or image editing.

2.2. Diffusion U-Net Architecture

Many pretrained text-to-image (T2I) diffusion models are text-conditioned U-Nets, consisting of an encoder-decoder

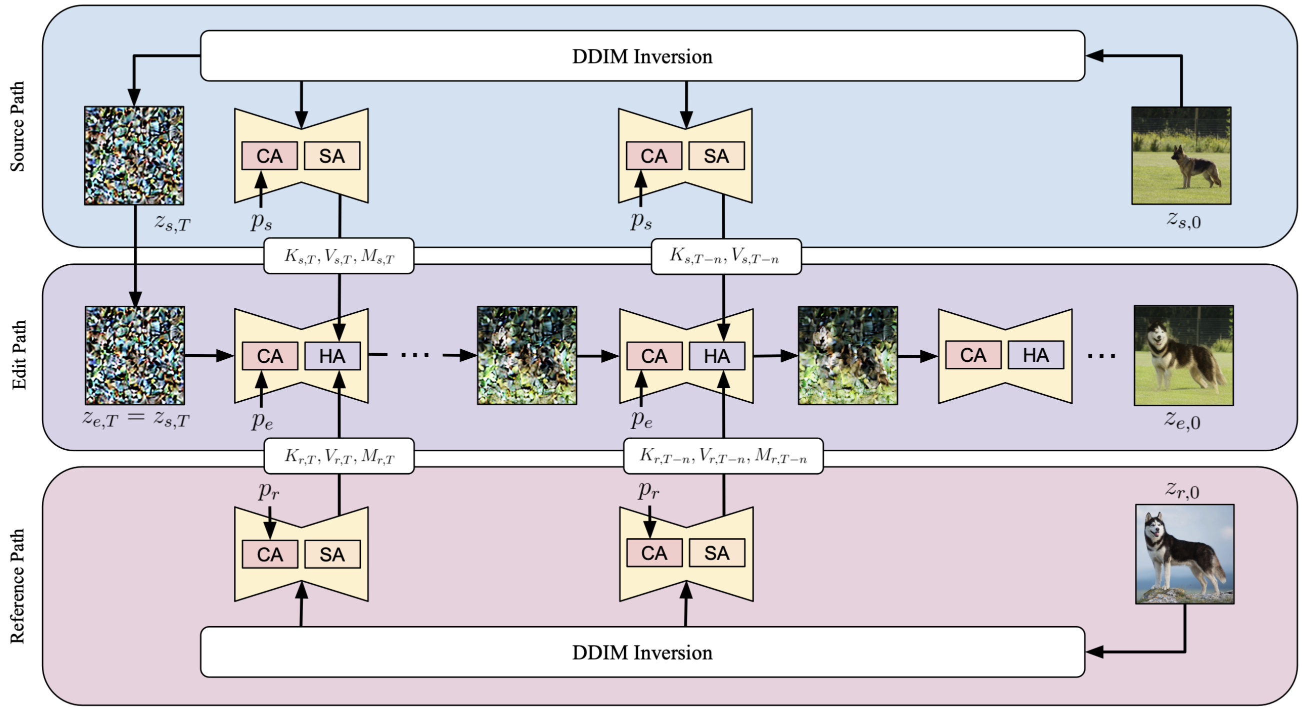

Figure 3. An overview of the architecture. We perform DDIM inversion on the source $z_{s,0}$ and reference $z_{r,0}$ images. During each inversion step, we store the queries (Q), keys (K), and values (V) from the self-attention and cross-attention layers. The final noised latent of the source image $z_{s,T}$ produced by the DDIM inversion is used as the starting latent for the edit path. In the edit path, we replace the self-attention blocks with the Hybrid Attention mechanism to extract and inject desired information via the stored queries (Q), keys (K), and values (V) of the self-attention layers. At the end of the edit path, we arrive at our edited latent $z_{e,0}$

structure. These models employ a combination of convolutional layers, self-attention layers, and cross-attention layers. Self-attention helps capture both local and global dependencies, while cross-attention allows for effective conditioning on textual prompts.

2.2.1 Self-Attention

Self-attention plays a crucial role in conveying both short-range and long-range dependencies across the entire latent. It ensures a coherent image construction by enabling each spatial position to attend to every other spatial position. The self-attention mechanism is expressed as:

$$\mathbf{Q} := \mathbf{h}_{l,t}\mathbf{W}_l^Q, \quad \mathbf{K} := \mathbf{h}_{l,t}\mathbf{W}_l^K, \quad \mathbf{V} := \mathbf{h}_{l,t}\mathbf{W}_l^V$$

where $\mathbf{h}_{l,t} \in \mathbb{R}^{(hw)\times d}$ represents the diffusion features at time step t, and $\mathbf{W}_l^Q, \mathbf{W}_l^K, \mathbf{W}_l^V \in \mathbb{R}^{d\times d}$ are linear projections for query, key, and value matrices. The attention scores are computed as:

$$\mathbf{A} := \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)$$

where $\mathbf{A} \in \mathbb{R}^{(hw)\times(hw)}$ represents the attention map, encoding how each pixel in $\mathbf{Q}$ corresponds to each pixel in $\mathbf{K}$, with higher scores indicating higher similarity. Each pixel's self-attention output is then computed as:

$$\mathbf{h}_{l,t} \leftarrow \mathbf{AV}$$

This operation ensures that each pixel aggregates information from other pixels, weighted by their relevance, resulting in spatially aware feature refinement.

2.2.2 Cross-Attention

In cross-attention layers, the query ($\mathbf{Q}$) is derived from the spatial latents, while the key ($\mathbf{K}$) and value ($\mathbf{V}$) originate from the encoded textual prompt. This mechanism enables effective conditioning of the generated latents on text. The cross-attention map [9] in this case reflects how each pixel attends to different textual tokens, guiding the diffusion model in aligning textual attributes with specific spatial regions. Notably, the cross-attention map allows us to identify the spatial regions associated with particular objects described in the prompt.

3. Method

Our approach begins with DDIM inversion applied to both the source and reference images, denoted as $z_{s,0}$ and $z_{r,0}$, respectively. During each inversion step, we store the queries (Q), keys (K), and values (V) from the self-attention and cross-attention layers. The final noised latent of the source image $z_{s,T}$ produced by the DDIM inversion is used as the starting latent for the edit path. $z_{s,T}$ serves as a strong initialization point for image editing by encapsulating essential structural information about the image thus facilitating accurate reconstruction. In the edit path, we replace the self-attention blocks with a novel Hybrid Attention mechanism. This mechanism allows for fine-grained extraction and injection of information in the edit path via the stored queries (Q), keys (K), and values (V) of the self-attention layers. At the end of the edit path, we arrive at our edited latent $z_{e,0}$ This process is outlined in Figure 3.

3.1. Hybrid Attention

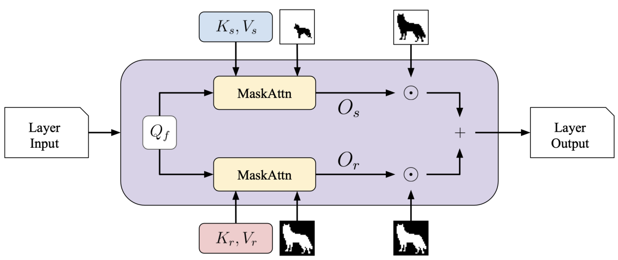

Figure 4. Hybrid Attention block.

As illustrated in Figure 1, our approach follows a two-step process: (1) extracting relevant information from the source and reference images, and (2) injecting these information into the appropriate spatial regions of the edit image to generate the final output.

To achieve information extraction, we apply a masked cross-image attention (MaskAttn) mechanism. Specifically, we use the query ($\mathbf{Q}_e$) from the edit image while utilizing the key ($\mathbf{K}_s, \mathbf{K}_r$) and value ($\mathbf{V}_s, \mathbf{V}_r$) representations from the source and reference images, respectively. The extraction process is guided by a binary extraction mask M_s or M_r, which defines the spatial regions to be extracted from each image. :

$$O_s = \text{MaskAttn}(\mathbf{Q}_e, \mathbf{K}_s, \mathbf{V}_s; M_s)$$

$$O_r = \text{MaskAttn}(\mathbf{Q}_e, \mathbf{K}_r, \mathbf{V}_r; M_r)$$

$$\text{MaskedAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}; M) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + M\right)\mathbf{V}$$

where M assigns large negative values (e.g., $-\infty$) to positions to be excluded. Once the relevant information has been extracted, we proceed to information injection, ensuring that the extracted features are transferred to the desired regions of the edit image. This is controlled by a binary injection mask M_e, which specifies where the injected information should be placed. The final Hybrid Attention output is computed as follows:

$$\text{HybridAttn} = O_s \odot (1 - M_e) + O_r \odot M_e$$

3.2. Automatic Masks

Masks play a crucial role in our pipeline by defining the regions for information extraction and injection. While users have the option to manually specify the source extraction, reference extraction and injection masks, automating this process is often preferable. To generate masks automatically, we utilize cross-attention maps extracted from the U-Net's cross-attention layers during the DDIM inversion process. These maps act as heatmaps, highlighting the relevance of each textual token to different regions of the image. By applying a thresholding operation to these attention maps, we obtain binary masks for objects of interest, which can then be used as the extraction or injection masks.

4. Experiment

We applied our method to the Stable Diffusion 2.1 model using publicly available pretrained weights. All editing experiments were conducted on real images, covering a variety of subjects including human portraits, animals and objects.

4.1. Types of Edits

Our method enable a variety of fine-grained edits, including reference-based object replacement, reference-based object insertion, object removal as well as prompt-based edits. The extraction and injection masks can either be manually provided by the user or taken from the cross-attention maps. Figure 5 depicts some of our results.

4.2. Optimizations

4.2.1 Layer Configuration for Hybrid Attention

We explored different configurations of layers for integrating our proposed Hybrid Attention mechanism in place of self-attention within the edit path. The Stable Diffusion 2.1 model consists of three main resolution blocks, each capturing different levels of abstraction.

- Deeper layers in the network encode global structure and high-level semantics. Modifying these layers can introduce unintended global changes, disrupting the overall composition of the image.
- Shallower (higher) layers focus on fine-grained textures and localized details, making them more suitable for precise edits while preserving the broader structure.

Figure 5. Our results prove our method to work across a wide range of editing domains, including reference-based object replacement, reference-based object insertion, object removal and prompt-based edits. For some of the prompt-based edits, we utilized an adaptive (A) cross-attention map as the insertion mask.

By strategically selecting which layers to modify, we ensure that edits remain localized and do not disrupt the structural integrity of the original image.

4.2.2 Post-Processing for Mask Refinement

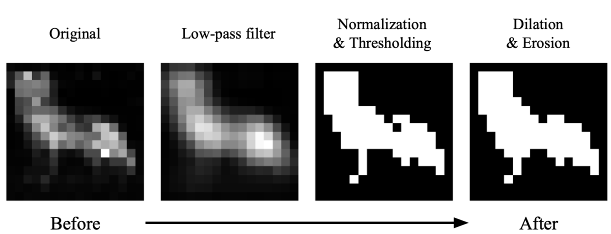

Figure 6. Automated thresholding for masks derived from cross-attention maps.

Cross-attention maps, particularly during early diffusion steps, can be noisy, making direct binary thresholding ineffective. To address this, we apply a Gaussian filter to reduce noise, followed by a normalization step to rescale attention values to a consistent range. Next, we apply thresholding, where pixels exceeding a selected threshold are retained as part of the mask. To further refine the mask and improve smoothness, we perform morphological operations, including dilation to expand regions and erosion to remove small artifacts. This results in cleaner, more robust masks suitable for use in Hybrid Attention.

5. Conclusion

This research project is a work in progress. Our proposed method has shown promising results in fine-grained, reference-based image editing without the need for additional training or guidance. This approach enables a wide range of precise edits, making it suitable for various downstream tasks, particularly real-time applications where computational efficiency is crucial. Additionally, our work provides deeper insights into the semantic reasoning of the diffusion U-Net, particularly how different layers and attention mechanisms contribute to spatial understanding, feature abstraction, and localized image editing.

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2

[2] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation, 2024. 2

[3] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6593–6602, 2024. 2

[4] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8795–8805, 2024. 2

[5] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. 2023. 2

[6] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair diffusion: A comprehensive multimodal object-level image editor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8609–8618, 2024. 2

[7] Roy Hachnochi, Mingrui Zhao, Nadav Orzech, Rinon Gal, Ali Mahdavi-Amiri, Daniel Cohen-Or, and Amit Haim Bermano. Cross-domain compositing with pretrained diffusion models, 2023. 2

[8] Runze He, Kai Ma, Linjiang Huang, Shaofei Huang, Jialin Gao, Xiaoming Wei, Jiao Dai, Jizhong Han, and Si Liu. Freeedit: Mask-free reference-based image editing with multi-modal instruction, 2024. 2

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. 3

[10] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 2

[11] Ying Hu, Chenyi Zhuang, and Pan Gao. Diffusest: Unleashing the capability of the diffusion model for style transfer. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, New York, NY, USA, 2024. Association for Computing Machinery. 2

[12] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941, 2023. 2

[13] DONGXU LI, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*, pages 30146–30166. Curran Associates, Inc., 2023. 2

[14] Pengzhi Li, Qiang Nie, Ying Chen, Xi Jiang, Kai Wu, Yuhuan Lin, Yong Liu, Jinlong Peng, Chengjie Wang, and Feng Zheng. Tuning-free image customization with image and text guidance. In *Computer Vision – ECCV 2024*, pages 233–250, Cham, 2025. Springer Nature Switzerland. 2

[15] Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance, 2024. 2

[16] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7465–7475, 2024. 2

[17] 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 (CVPR)*, pages 6038–6047, 2023. 2

[18] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8488–8497, 2024. 2

[19] 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 (CVPR)*, pages 22500–22510, 2023. 2

[20] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8839–8849, 2024. 2

[21] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 2

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. 2

[23] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models, 2024. 2

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18381–18391, 2023. 2

[25] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

[26] Yu Zeng, Vishal M. Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6786–6795, 2024.

[27] 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 (ICCV)*, pages 3836–3847, 2023. 2