

PairEdit: Learning Semantic Variations for Exemplar-based Image Editing

Haoguang Lu¹ Jiacheng Chen¹ Zhenguo Yang² Aurele Tohokantche Gnanha³

Fu Lee Wang⁴ Qing Li⁵ Xudong Mao^{1*}

¹Sun Yat-sen University ²Guangdong University of Technology

³Huawei Noah’s Ark Laboratory ⁴Hong Kong Metropolitan University

⁵The Hong Kong Polytechnic University

Abstract

Recent advancements in text-guided image editing have achieved notable success by leveraging natural language prompts for fine-grained semantic control. However, certain editing semantics are challenging to specify precisely using textual descriptions alone. A practical alternative involves learning editing semantics from paired source-target examples. Existing exemplar-based editing methods still rely on text prompts describing the change within paired examples or learning implicit text-based editing instructions. In this paper, we introduce PairEdit, a novel visual editing method designed to effectively learn complex editing semantics from a limited number of image pairs or even a single image pair, without using any textual guidance. We propose a target noise prediction that explicitly models semantic variations within paired images through a guidance direction term. Moreover, we introduce a content-preserving noise schedule to facilitate more effective semantic learning. We also propose optimizing distinct LoRAs to disentangle the learning of semantic variations from content. Extensive qualitative and quantitative evaluations demonstrate that PairEdit successfully learns intricate semantics while significantly improving content consistency compared to baseline methods. Code will be available at <https://github.com/xudonmao/PairEdit>.



Figure 1: Editing results of PairEdit trained on three image pairs (1st-2nd rows) or a single image pair (3rd row). Our method effectively captures semantic variations between source and target images.

*Corresponding author (xudong.xdiao@gmail.com).

1 Introduction

Recent advancements in diffusion models [22, 45] have significantly improved the quality and diversity of visual outputs, particularly in text-to-image synthesis tasks. The versatility of diffusion-based frameworks has further expanded their applicability beyond image generation into sophisticated image editing domains. Notably, text-guided editing has emerged as a powerful method, enabling fine-grained control over semantic attributes through natural language prompts [21, 40]. Additionally, diffusion models have been effectively employed in image-guided editing tasks [64, 9], facilitating the transformation of visual inputs guided by reference images, and in instructional editing tasks [5, 18], allowing intuitive edits through explicit instructions.

Among these, text-guided image editing has achieved remarkable success, enabling precise and flexible editing. Nevertheless, certain editing semantics are challenging to specify clearly through textual descriptions alone. A practical alternative involves learning semantics directly from paired images—consisting of before-and-after editing examples. However, existing exemplar-based editing methods typically rely on large language models or manual efforts to provide text prompts describing the change from source to target images [20, 8], or require encoding the change into the latent space of pre-trained instructional editing models [39, 53]. Notably, Concept Slider [17] introduces a loss function designed to train a single LoRA [23] with opposing scaling factors (positive and negative) to capture semantic variations by compelling predictions of identical noise. However, as illustrated in Figure 3, this method still struggles with learning complex semantics and maintaining content consistency between original and edited images.

In this paper, we introduce PairEdit, a novel visual editing method capable of effectively learning complex semantics from a small set of image pairs or even from a single image pair, without using any textual guidance. We explore optimizing LoRA to capture semantic variations between source and target images. To this end, we introduce a guidance-based noise prediction for LoRA optimization, explicitly modeling semantic variations by converting paired images into a guidance direction (i.e., $\epsilon_{\text{target}} - \epsilon_{\text{source}}$). Furthermore, we propose a content-preserving noise schedule designed to align the guidance scale with the LoRA scaling factor, enabling more effective semantic learning.

To disentangle semantic variation from content within paired images, we propose separating their learning processes by jointly optimizing two distinct LoRA modules: a content LoRA and a semantic LoRA. This optimization strategy encourages the content LoRA to reconstruct the source image while guiding the semantic LoRA to capture semantic variations from source to target images.

Our approach facilitates visual image editing based on a limited number of paired examples, effectively learning various semantics such as appearance change, age progression, and stylistic transformation. Moreover, our approach enables continuous control over the semantics by adjusting the scaling factor of the learned semantic LoRA. We demonstrate the effectiveness of our method through comprehensive qualitative and quantitative evaluations against several state-of-the-art methods. The results show that PairEdit achieves superior performance in terms of both identity preservation and semantic fidelity compared to existing baselines.

2 Related Work

Text-to-Image Diffusion Models. Diffusion models [51, 54, 22] have emerged as a dominant paradigm in text-to-image synthesis, which progressively refines Gaussian noise into high-quality images. In particular, latent diffusion models [45] employ U-Net architectures [46] to efficiently denoise in compressed latent spaces, setting the stage for notable improvements in resolution and scalability. Recent developments [13, 31] have initiated a shift from U-Net to vision transformer-based architectures, known as Diffusion Transformers (DiTs) [43]. These models utilize global attention mechanisms and advanced positional encodings to enhance model capacity and performance. These DiT-based diffusion models, such as Flux [31] and Stable Diffusion 3 [14], have consistently demonstrated state-of-the-art generation quality, with performance scaling predictably with model size. Moreover, flow-matching objectives [32, 33] have further enhanced the generation quality of these DiT-based models. Leveraging these advancements, Flux has achieved remarkable success in various applications such as image editing [12, 47, 58], personalized generation [15, 28, 63], and reference image generation [24, 36].

Image Editing. Generative adversarial networks [19] have been extensively studied in the context of image editing by leveraging their expressive latent spaces [70, 49]. Recently, diffusion models, known for their superior capabilities in text-to-image generation, have attracted significant attention in image editing. Various input conditions have been investigated in diffusion-based image editing methods, as reviewed in [25]. Among these, text-guided image editing has achieved great success, offering an intuitive and flexible way for users to describe desired edits. This category includes approaches utilizing either descriptive texts for the edited image [21, 30, 29, 41, 57, 6, 42, 4] or explicit editing instructions [5, 18, 50, 69, 16, 26]. Additionally, some methods employ masks as input conditions to achieve precise control [66, 11, 60, 1, 2, 71], while others utilize reference images to guide the editing [67, 65, 55, 64, 9]. Another notable approach involves learning semantics directly from paired examples [3, 61, 17].

Exemplar-based Image Editing. Exemplar-based image editing has emerged as a powerful paradigm in image editing, effectively leveraging paired examples rather than relying on explicit textual instructions. MAE-VQGAN [3] first poses this problem as an image inpainting task, a framework that has since been adopted by several approaches [56, 61, 35, 20]. Alternative techniques utilize ControlNet-based architectures [67], as demonstrated by methods such as InstructGIE [38] and PromptDiffusion [62], which treat example images as spatial conditions. Other approaches build on the generalization capabilities of InstructPix2Pix [5] by inverting visual instructions into textual embeddings [39] or into LoRA weights [53]. Pair Customization [27] explicitly learns separate LoRAs for style and content within an image pair. Concept Slider [17] introduces a loss function that encourages a single LoRA with opposing scaling factors (positive and negative) to capture semantic variations by constraining them to predict the same noise. Despite these advancements, existing methods still face significant challenges in learning complex editing semantics from paired examples while maintaining content consistency between the original and edited images.

3 Method

3.1 Preliminaries

Rectified-Flow Models. Our approach is based on the Flux model, a type of rectified-flow model [32, 34] for text-to-image generation. Rectified-flow models define a transition from a Gaussian noise distribution p_1 to the real data distribution p_0 . Given empirical observations from two distributions $x_0 \sim p_0$, $x_1 \sim p_1$, and $t \in [0, 1]$, the forward process of rectified-flow models is modeled as a continuous path:

$$x_t = (1 - t)x_0 + t\epsilon, \quad \epsilon \sim N(0, 1) \quad (1)$$

To reverse this process and recover data from noise, a velocity prediction network v_θ is trained to predict the velocity v of the flow. This network can serve as a noise prediction network ϵ_θ using the reparameterization technique introduced in [13].

Classifier-Free Guidance. Classifier-Free Guidance (CFG) is a technique introduced to improve the quality and controllability of samples generated by diffusion models without requiring an external classifier. CFG leverages the same prediction network ϵ_θ in both conditional and unconditional modes. During sampling, predictions from the conditional model and the unconditional model are combined using a guidance scale γ :

$$\hat{\epsilon}_\theta = \epsilon_\theta(x_t, \emptyset) + \gamma(\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t, \emptyset)). \quad (2)$$

The term $\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t, \emptyset)$ is often referred to as the guidance direction. Our approach aims to construct a guidance direction corresponding to the target semantic variation observed within the paired images.

3.2 Learning Semantic Variations with PairEdit

Our goal is to learn semantic variations from a small set of image pairs. The key challenge lies in extracting accurate semantics that generalize well to editing new images. We introduce a novel LoRA-based method enabling precise and continuous image editing using only a few image pairs or even a single pair. Our method is based on three main ideas. First, we propose a guidance-based

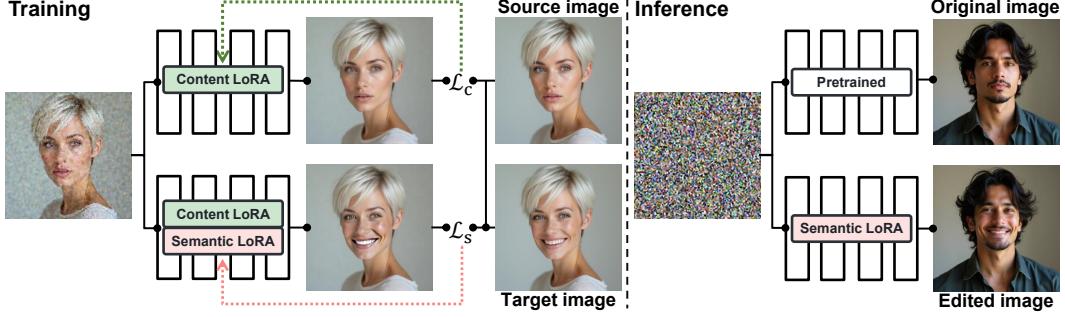


Figure 2: **Overview of PairEdit.** (Left) Given a pair of source and target images, we jointly train two LoRAs: a content LoRA, which reconstructs the source image using the standard diffusion loss (Eq. 3), and a semantic LoRA, which captures the semantic difference between the paired images using the proposed semantic loss (Eq. 10). (Right) During inference, when applying the learned semantic LoRA, the original image is edited towards the target semantic.

noise prediction that helps LoRA learn semantic variations from source to target images. Second, we introduce a content-preserving noise schedule for more effective semantic learning. Third, we propose separating semantic variation from content within image pairs by using two distinct LoRA adapters. An overview of the proposed PairEdit framework is depicted in Figure 2.

Separating Semantic Variation and Content. Our approach leverages the fact that paired images share the same content but differ only in target semantics. Inspired by recent studies in image stylization [27, 7], we jointly optimize two distinct LoRAs: a content LoRA, which reconstructs the source image, and a semantic LoRA, which captures semantic differences between source and target images. As illustrated in Figure 2, given the noised source image as input, the content LoRA aims to reconstruct the source image, while the semantic LoRA transforms the noised source image into the target image. Formally, we denote the content and semantic LoRA weights as θ_c and θ_s , respectively. For the content LoRA, we employ a standard diffusion loss for reconstruction:

$$\mathcal{L}_{\text{content}} = \mathbb{E}_{x_0^A, \epsilon_0, t} [\|\epsilon_0 - \epsilon_{\theta_c}(x_t^A, \emptyset)\|_2^2], \quad (3)$$

where x_0^A and x_t^A denote the original and noised source images, respectively. For the semantic LoRA, however, we cannot simply rely on the reconstruction loss, as it involves denoising the noised source image toward the target image. Therefore, we explicitly model the semantic variation using a guidance direction term.

Guidance-based Semantic Variation. As illustrated at the bottom of Figure 2, we jointly leverage the content and semantic LoRAs to denoise the noised source image toward the target image. To achieve this, the predicted noise by these two LoRAs should incorporate the semantic variation from source to target images. Inspired by CFG (Eq. 2), we propose encoding the semantic variation into the CFG guidance direction. Thus, the target prediction noise ϵ^* for content and semantic LoRAs is defined as:

$$\epsilon^* = \epsilon_{\theta_c^*}(x_t^A, \emptyset) + \gamma(\epsilon_{\theta_{c,s}^*}(x_t^A, \emptyset) - \epsilon_{\theta_c^*}(x_t^A, \emptyset)), \quad (4)$$

where θ^* denotes the “ground truth” weights for content and semantic LoRAs, and γ controls the strength of the guidance. In this equation, the first term $\epsilon_{\theta_c^*}$ corresponds to content reconstruction, while the guidance direction term $\epsilon_{\theta_{c,s}^*} - \epsilon_{\theta_c^*}$ corresponds to semantic variation. For simplicity, we denote $\epsilon_{\theta_c^*}(x_t^A, \emptyset)$ and $\epsilon_{\theta_{c,s}^*}(x_t^A, \emptyset)$ as ϵ_t^A and ϵ_t^B , respectively. Note that the first term $\epsilon_{\theta_c^*}$ can be replaced with the true noise ϵ_0 added to the source image. Thus, we reformulate Eq. 4 as:

$$\epsilon^* = \epsilon_0 + \frac{\gamma}{\Delta t}[(x_t^A - \Delta t \epsilon_t^A) - (x_t^A - \Delta t \epsilon_t^B)]. \quad (5)$$

Applying the denoising formula (i.e., $x_{t-\Delta t} = x_t - \Delta t \epsilon$), and considering that ϵ_t^B denoises the source image towards the target image, we derive:

$$\epsilon^* = \epsilon_0 + \frac{\gamma}{\Delta t}(x_{t-\Delta t}^A - x_{t-\Delta t}^B). \quad (6)$$

Utilizing Eq. 1 and applying identical noise to both source and target images, we obtain $x_{t-\Delta t}^A - x_{t-\Delta t}^B = (1-t+\Delta t)(x_0^A - x_0^B)$, yielding:

$$\epsilon^* = \epsilon_0 + \frac{\gamma}{\Delta t}(1-t+\Delta t)(x_0^A - x_0^B). \quad (7)$$

Here, the weight $\frac{\gamma}{\Delta t}(1-t+\Delta t)$ is time-dependent. However, in practice, it is beneficial to establish a fixed weight aligned with a constant scaling factor of LoRA during optimization. To address this issue, we introduce a new noise schedule designed to make the weight of $x_0^A - x_0^B$ time-independent.

Content-Preserving Noise Schedule. To achieve a time-independent weight for $x_0^A - x_0^B$, we propose a new noise schedule defined as:

$$x_t = x_0 + t\beta\epsilon, \quad (8)$$

where β controls the strength of the noise. Compared to the standard noise schedule (Eq. 1), our method preserves content information when $t=1$; hence, we refer to it as the content-preserving noise schedule. Using this schedule, we derive $x_{t-\Delta t}^A - x_{t-\Delta t}^B = x_0^A - x_0^B$ when applying identical noise to x_0^A and x_0^B . Consequently, the target noise prediction becomes:

$$\epsilon^* = \beta\epsilon_0 + \eta(x_0^A - x_0^B), \quad (9)$$

where $\eta = \frac{\gamma}{\Delta t}$.

As illustrated at the bottom of Figure 2, our semantic variation loss encourages the predicted noise $\epsilon_{\theta_{c,s}}$ by content and semantic LoRAs towards the target noise ϵ^* , which is defined as:

$$\mathcal{L}_{\text{semantic}} = \mathbb{E}_{x_0^A, x_0^B, \epsilon_0, t} [\|\epsilon^* - \epsilon_{\theta_{c,s}}(x_t^A, \emptyset)\|_2^2]. \quad (10)$$

It is important to note that we optimize only the semantic LoRA weights with this loss, stopping gradient flow to the content LoRA weights. The benefits of our content-preserving noise schedule are two-fold. First, we set a fixed η aligned with a constant scaling factor of the semantic LoRA, which stabilizes the training process. Second, for large t values, our method preserves content information, resulting in meaningful semantic differences in $x_{t-\Delta t}^A - x_{t-\Delta t}^B$ (Eq. 6). In contrast, with the standard noise schedule, $x_{t-\Delta t}^A - x_{t-\Delta t}^B$ becomes meaningless as both $x_{t-\Delta t}^A$ and $x_{t-\Delta t}^B$ approach pure noise.

Although our noise schedule differs from the original one of the pretrained diffusion model, the pretrained model already has a general capability to handle and effectively denoise noisy inputs. Furthermore, LoRA can adapt the model’s existing knowledge to this new noising approach. During inference, we follow the approach of [37, 17] by disabling the semantic LoRA for the initial t steps to maintain content structure. Thus, the semantic LoRA does not need to learn denoising from purely noisy inputs.

Our full objective is:

$$\theta_s^* = \arg \min_{\theta_c, \theta_s} \mathcal{L}_{\text{content}} + \lambda \mathcal{L}_{\text{semantic}}, \quad (11)$$

where λ controls the strength of the semantic loss.

4 Experiments

4.1 Implementation and Evaluation Setup

Implementation Details. Our implementation is based on the publicly available FLUX.1-dev², with both model weights and text encoders frozen. The rank of LoRA weights is set to 16. The parameter β is set to 3 for global editing semantics (e.g., stylization) and 1 for local editing semantics (e.g., smile). For all experiments, η and λ are set to 4 and 1, respectively. We jointly train content and semantic LoRAs for 500 steps using a learning rate of 2×10^{-3} . The entire training process takes approximately 8 minutes on a single NVIDIA A100 80GB GPU. Following [37, 17], we set the LoRA scaling factor to 0 during the initial 14 steps to maintain the structure of the original image. Additional implementation details for our method and baseline methods are provided in Appendix A.

²<https://huggingface.co/black-forest-labs/FLUX.1-dev>

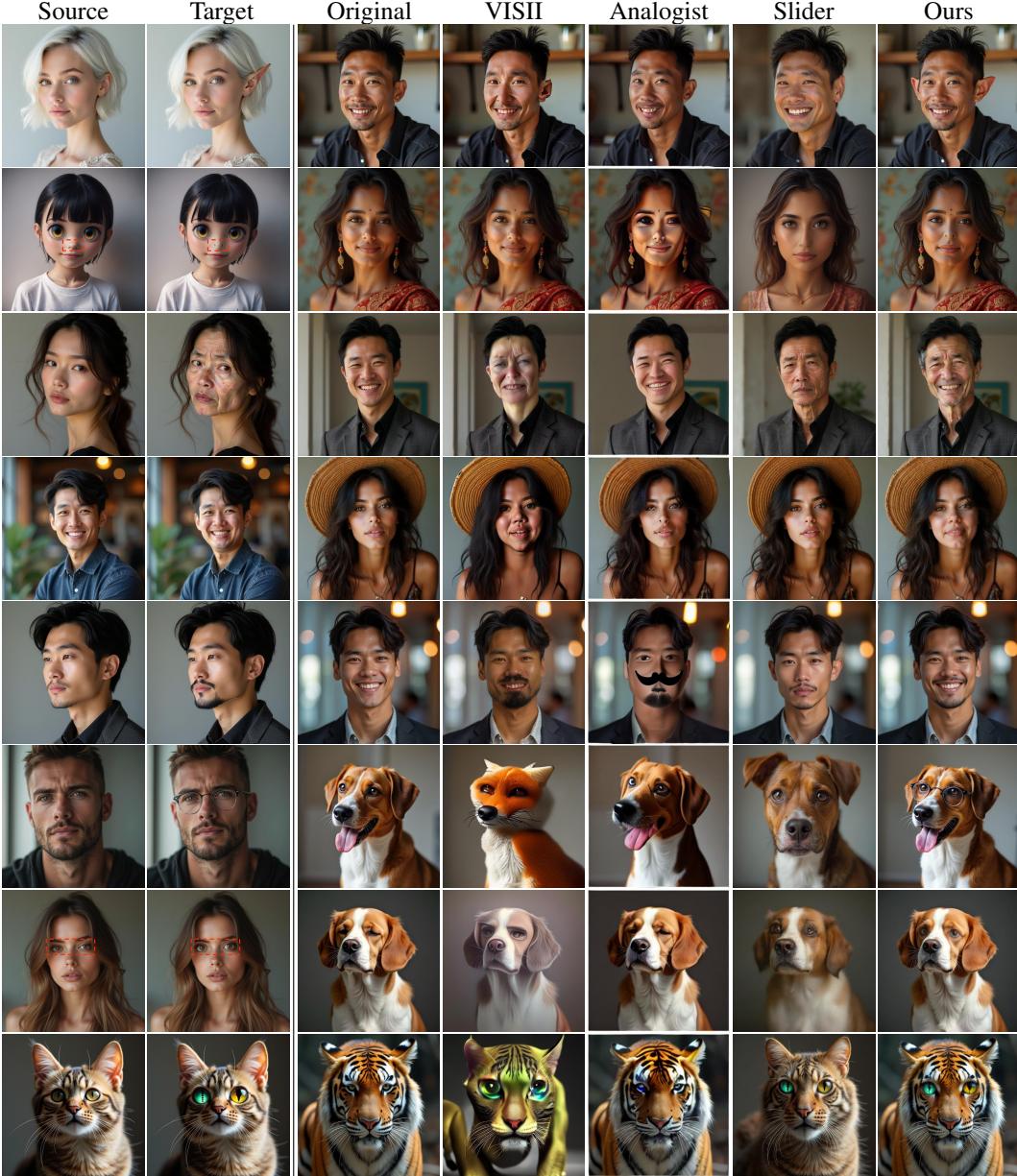


Figure 3: **Qualitative comparison.** We present exemplar-based image editing results of our method and three baseline methods, including VISII [39], Analogist [20], and Slider [17]. Our method demonstrates superior performance in accurately editing the original image while preserving its content.

Datasets. We create paired source and target images as follows: First, we apply existing image editing techniques, such as SDEdit [37], to translate source images into preliminary target images. Next, we transfer edited regions from the preliminary target images onto the corresponding regions of source images, generating the final target images. Additionally, some image pairs are collected from the web or sourced from [27]. For semantic learning, PairEdit is trained using either three image pairs (e.g., age, chubbiness, and elf ears) or a single image pair (e.g., stylization, lipstick, and dragon eyes).

Evaluation Setup. We compare our method with four exemplar-based editing methods: VISII [39], Analogist [20], Edit Transfer [8], and Visual Concept Slider [17], as well as two text-based editing methods that support continuous editing: SDEdit [37] and Textual Concept Slider [17]. Note that

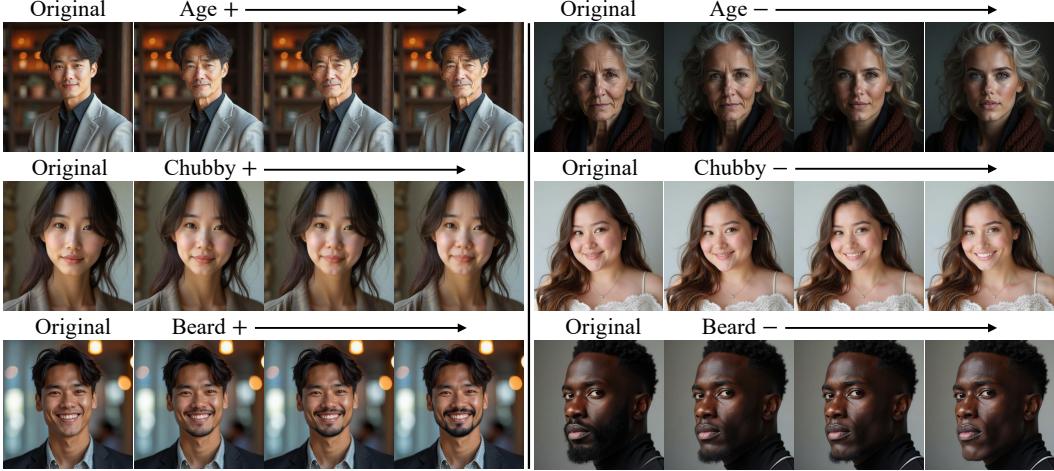


Figure 4: Examples of continuous editing by our method. By adjusting the scaling factor of the learned LoRA, our method enables a high-fidelity and fine-grained control over the semantic from exemplar images.

Table 1: **Quantitative comparison.** We evaluate each method by measuring identity preservation when performing similar editing magnitude. Identity preservation is measured using LPIPS distance, and editing magnitude is measured using the cosine similarity over CLIP embeddings.

Semantics	SDEdit		Textual Slider		Visual Slider		Ours	
	CLIP↑	LPIPS↓	CLIP↑	LPIPS↓	CLIP↑	LPIPS↓	CLIP↑	LPIPS↓
Age	0.2285	0.1956	0.2266	0.1631	0.2257	0.1716	0.2382	0.1359
Smile	0.2533	0.1419	0.2556	0.1749	0.2724	0.1380	0.2896	0.1120
Chubbiness	0.2347	0.2173	0.2332	0.1423	0.2329	0.1747	0.2420	0.0815
Glasses	0.2419	0.1370	0.2427	0.1602	0.2421	0.1706	0.2886	0.0911

Analogist and Edit Transfer require large language models or manual efforts to provide textual prompts describing the edits. For quantitative evaluation, we assess each method across four distinct semantics: age, smile, chubbiness, and glasses. For each semantic, we generate 500 pairs of original and edited images using the same random seed across all methods.

4.2 Results

Qualitative Evaluation. Figure 3 provides a visual comparison of editing results between our method and the baselines. We examine various editing tasks, including facial feature transformation, appearance alteration, and accessory addition. As shown, VISII [39] and Analogist [20] struggle to accurately capture semantic variations between source and target images, generating low-quality images. Concept Slider [17] captures some semantic variations but consistently fails to preserve the identity of the original image. It also struggles with complex semantics (e.g., elf ears and chubbiness) and exhibits limited generalization capability (e.g., adding glasses to dogs). Due to limited space, the comparison with Edit Transfer [8] is presented in Appendix B. Edit Transfer also fails to capture semantic variations and produces images nearly identical to the original. In contrast, PairEdit successfully performs all desired edits learned from paired examples. Moreover, our method achieves high-quality continuous editing by adjusting the scaling factor of the learned semantic LoRA, as illustrated in Figure 4. Additional qualitative evaluations are provided in Appendix B, and we also present a visual comparison of editing results using a single image pair in Appendix E.

Quantitative Evaluation. For quantitative assessment, we compare our method with three baselines that support continuous editing. We evaluate each method by measuring identity preservation while maintaining a similar editing magnitude. To ensure valid editing for the baselines, we employ a set of simple semantics for evaluation. Identity preservation is quantified using the LPIPS distance [68]

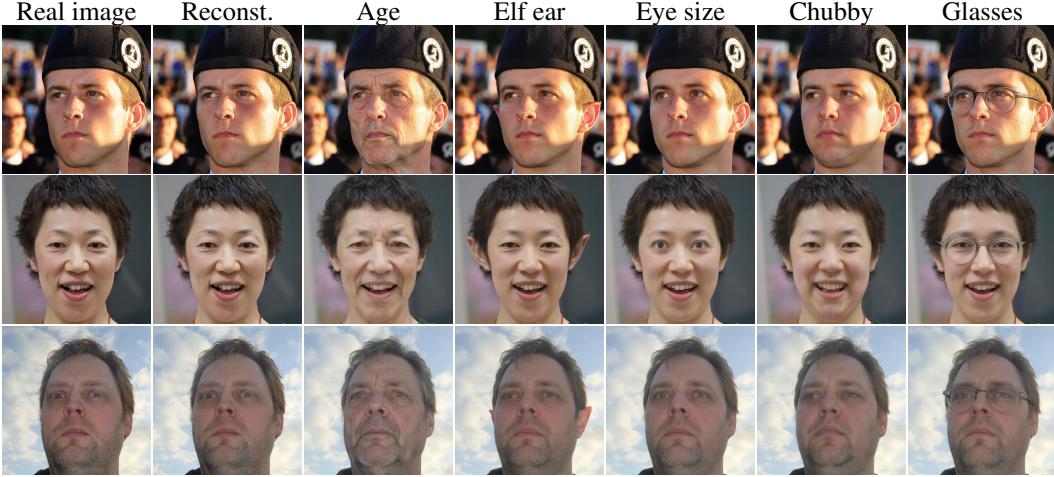


Figure 5: **Real image editing.** The reconstructed image is obtained by optimizing a LoRA over the real image. We apply the learned semantic LoRAs to the reconstructed image by merging the LoRAs during inference.

between the original and edited images, while editing magnitude is measured via cosine similarity between CLIP embeddings [44] of the edited images and their corresponding textual editing descriptions. As demonstrated in Table 1, our method achieves significantly lower LPIPS distances compared to the baselines when applying comparable editing magnitudes.

User Study. We also conducted a user study to evaluate our method. In each question, participants were shown a pair of source and target images, an original image, and two edited images: one produced by our method and the other by a baseline method. Participants were asked to select the image exhibiting superior editing quality while preserving the original identity. A total of 720 responses were collected from 24 participants, as detailed in Table 2. The results clearly indicate a strong preference for our method.

Table 2: **User Study.** Participants were asked to select the image exhibiting superior editing quality while preserving the identity.

Baselines	Prefer Baseline	Prefer Ours
VISII [39]	6.2%	93.8%
Analogist [20]	1.3%	98.7%
Slider [17]	3.8%	96.2%

Real Image Editing. Editing real images typically involves finding an initial noise vector that reconstructs the input image using inversion techniques [52, 48]. However, we observe that directly applying existing inversion methods designed for Flux [48] with the learned semantic LoRA yields poor editing quality. This issue arises because these inversion methods fail to accurately map the input image back into Flux’s original latent space, a limitation also highlighted in [12]. Since inversion methods are not the primary focus of this paper, we adopt a simple reconstruction strategy by optimizing a LoRA over the input image. To apply learned edits to reconstructed images, we merge the two LoRAs during inference as follows:

$$\epsilon_{\theta_r}(x_t, \emptyset) + \gamma_{\text{real}}(\epsilon_{\theta_s}(x_t, \emptyset) - \epsilon_{\theta_r}(x_t, \emptyset)), \quad (12)$$

where θ_r and θ_s denote the reconstruction and semantic LoRA weights, respectively, and γ_{real} is set to 0.75. We empirically find that this merging strategy improves identity preservation compared to linear combination of LoRA weights, as illustrated in Appendix D. As shown in Figure 5, our approach achieves high-quality editing while effectively preserving the identity of real images.

Composing Sequential Edits. Our method supports combining multiple edits through the merging of several learned semantic LoRAs. We employ the same merging strategy during inference as in real image editing, resulting in better editing quality compared to a linear combination of LoRA weights. As illustrated in Figure 6, our method effectively composes multiple edits while preserving individual identities.

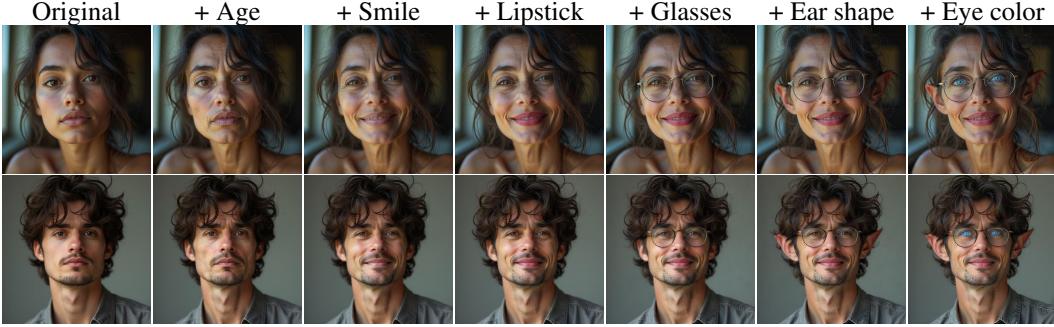


Figure 6: **Composing sequential edits.** Our method effectively composes different edits while preserving the original identity. Multiple semantic LoRAs are merged using the strategy illustrated in Eq. 12.

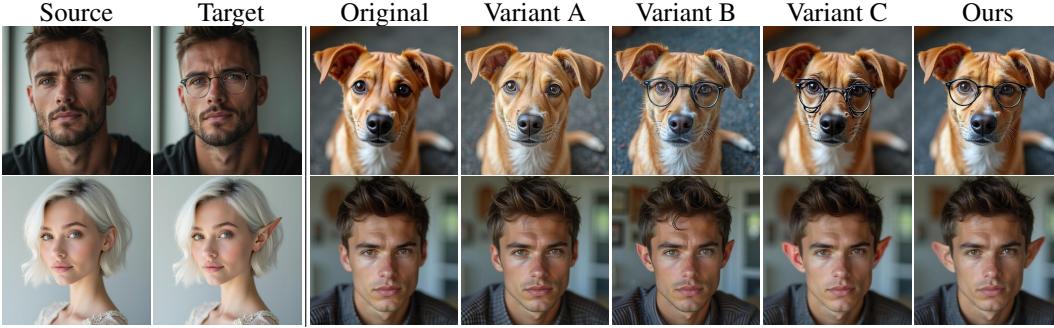


Figure 7: **Ablation study.** We evaluate three variants of our model: (A) replacing the semantic loss with the visual concept loss proposed in [17], (B) removing the content LoRA, and (C) replacing the content-preserving noise schedule with a standard noise schedule.

4.3 Ablation Study

In this section, we perform an ablation study to assess the effectiveness of individual components within our framework. Specifically, we evaluate three variants: (1) replacing the semantic loss with the visual concept loss proposed in [17], (2) removing the content LoRA, and (3) replacing the content-preserving noise schedule with a standard noise schedule. Figure 7 presents a visual comparison of editing results generated by each variant. The results demonstrate that all proposed components are crucial for achieving identity preservation and semantic fidelity. Omitting our semantic loss significantly reduces the model’s ability to capture complex editing semantics. Removing the content LoRA leads to inconsistent results, such as unintended fur color changes in the first row and hairstyle alterations in the second row. Employing a standard noise schedule negatively affects the generalization capability of the semantic LoRA, causing blurred glasses in the first row and inconsistent ear coloration in the second row. Additional ablation study results are provided in Appendix F.

5 Conclusions and Limitations

In this paper, we introduced PairEdit, a novel visual editing framework designed to effectively capture complex semantic variations from limited paired-image examples. Utilizing a guidance-based target denoising prediction term, our method explicitly transforms semantic differences between source and target images into a guidance direction. By separately optimizing two dedicated LoRAs for semantic variation and content reconstruction, PairEdit effectively disentangles semantic attributes from content information. However, one limitation of PairEdit is its reliance on paired images, which is not directly applicable to unpaired datasets. In future work, we aim to explore methods capable of extracting editing semantics from unpaired image sets, further enhancing the flexibility and practical applicability of our approach.

References

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022.
- [2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. In *SIGGRAPH*, 2023.
- [3] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. Visual prompting via image inpainting. In *NeurIPS*, 2022.
- [4] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledit++: Limitless image editing using text-to-image models. In *CVPR*, 2024.
- [5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.
- [6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinjiang Zheng. Masactr: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023.
- [7] Bolin Chen, Baoquan Zhao, Haoran Xie, Yi Cai, Qing Li, and Xudong Mao. Consislora: Enhancing content and style consistency for lora-based style transfer. *arXiv preprint arXiv:2503.10614*, 2025.
- [8] Lan Chen, Qi Mao, Yuchao Gu, and Mike Zheng Shou. Edit transfer: Learning image editing via vision in-context relations. *arXiv preprint arXiv:2503.13327*, 2025.
- [9] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024.
- [10] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, 2023.
- [11] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, 2022.
- [12] Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. Fluxspace: Disentangled semantic editing in rectified flow transformers. *arXiv preprint arXiv:2412.09611*, 2024.
- [13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lace, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.
- [14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lace, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.
- [15] Haoran Feng, Zehuan Huang, Lin Li, Hairong Lv, and Lu Sheng. Personalize anything for free with diffusion transformer. *arXiv preprint arXiv:2503.12590*, 2025.
- [16] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfel Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *ICLR*, 2024.
- [17] Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *ECCV*, 2024.
- [18] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. Instructdiffusion: A generalist modeling interface for vision tasks. In *CVPR*, 2024.

- [19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [20] Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context learning with image diffusion model. In *SIGGRAPH*, 2024.
- [21] 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.
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.
- [23] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.
- [24] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.
- [25] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024.
- [26] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, and Ying Shan. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*, 2024.
- [27] Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-to-image models with a single image pair. In *SIGGRAPH Asia*, 2024.
- [28] Hao Kang, Stathi Fotiadis, Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Min Jin Chong, and Xin Lu. Flux already knows – activating subject-driven image generation without training. *arXiv preprint arXiv:2504.11478*, 2025.
- [29] 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 *CVPR*, 2023.
- [30] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022.
- [31] Black Forest Labs. Flux, 2024.
- [32] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.
- [33] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.
- [34] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.
- [35] Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering. In *ICML*, 2024.
- [36] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025.
- [37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.

- [38] Zichong Meng, Changdi Yang, Jun Liu, Hao Tang, Pu Zhao, and Yanzhi Wang. Instructgie: Towards generalizable image editing. In *ECCV*, 2024.
- [39] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via visual prompting. In *NeurIPS*, 2023.
- [40] 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.
- [41] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH*, 2023.
- [42] Or Patashnik, Daniel Garabi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *ICCV*, 2023.
- [43] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- [46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.
- [47] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *ICLR*, 2025.
- [48] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *ICLR*, 2025.
- [49] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.
- [50] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *CVPR*, 2024.
- [51] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.
- [52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [53] Xue Song, Jiequan Cui, Hanwang Zhang, Jiaxin Shi, Jingjing Chen, Chi Zhang, and Yu-Gang Jiang. Lora of change: Learning to generate lora for the editing instruction from a single before-after image pair. *arXiv preprint arXiv:2411.19156*, 2024.
- [54] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.
- [55] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Generative object compositing. In *CVPR*, 2023.
- [56] Yasheng Sun, Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, and Hideki Koike. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. In *NeurIPS*, 2023.
- [57] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.

- [58] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.
- [59] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020.
- [60] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*, 2023.
- [61] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, 2023.
- [62] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. *arXiv preprint arXiv:2305.01115*, 2023.
- [63] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025.
- [64] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, 2023.
- [65] 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 *CVPR*, 2023.
- [66] Tao Yu, Runseng Feng, Ruoyu Feng, Jimming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.
- [67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.
- [68] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- [69] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. Hive: Harnessing human feedback for instructional visual editing. In *CVPR*, 2024.
- [70] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.
- [71] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *ECCV*, 2024.

A Implementation Details.

Our method leverages FLUX.1-dev, with both model weights and text encoders fixed. We employ the Adam optimizer to tune the LoRA weights, setting the rank to 16. The content and semantic LoRAs are jointly trained for 500 steps using a learning rate of 2×10^{-3} . For all experiments, we perform image generation with 28 inference steps. To preserve the structure of the original image, we follow the approach described in [37, 17], setting the LoRA scaling factor to 0 for the initial 14 steps. For Textual Concept Slider [17], we utilize their official Flux implementation. For Visual Concept Slider [17], due to the unavailability of the official Flux implementation, we implement the Flux-based model following their SDXL implementation. For other baseline methods, including VISII [39], Analogist [20], and Edit Transfer [8], we utilize their official implementations and follow the hyperparameters described in their papers. For SDEdit [37], we use the diffusers Flux implementation. When using GPT-4o, the editing prompt is: “The first and second images represent a ‘before and after’ editing pair. Please analyze the changes made between them and apply the same edit to the third image.”

B Additional Qualitative Results

In Figure 8, we present additional qualitative comparisons against three baseline methods: Edit Transfer [8], GPT-4o, and Visual Concept Slider [17]. Our method demonstrates superior performance in terms of identity preservation and semantic fidelity compared to the baselines. Edit Transfer struggles to accurately capture the semantic variations between source and target images. GPT-4o shows poor identity preservation, and Visual Concept Slider also fails to preserve the original identity while struggling with complex semantic edits.

C Additional Real Image Editing Results

In Figure 9, we provide additional examples of real image editing. Reconstructed images are obtained by optimizing LoRAs directly over real images. We apply the learned semantic LoRAs to these reconstructed images using guidance-based LoRA fusion as described in Eq. 12. The results demonstrate the effectiveness of our approach in editing real images across various semantic attributes.

D Comparison of LoRA Fusion Methods

In Figure 10, we compare two LoRA fusion methods: (1) linear combination of LoRA weights and (2) guidance-based LoRA fusion (Eq. 12). The linear combination approach tends to produce blurry outputs for certain semantics, whereas the guidance-based LoRA fusion provides better identity preservation and image quality.

E Learning with a Single Image Pair

In this section, we evaluate PairEdit when trained using only a single image pair. As shown in Figure 11, our method outperforms baseline methods in both identity preservation and semantic fidelity. However, we observe that providing multiple image pairs further helps the model learn complex semantics and enhances its generalization capability (e.g., adding glasses to dogs).

F Additional Ablation Study

As discussed in Section 4.3, we evaluate three variants of our model: (1) replacing the semantic loss with the visual concept loss from [17], (2) removing the content LoRA, and (3) substituting the content-preserving noise schedule with a standard noise schedule. Additional results of the ablation study are presented in Figure 12.

G User Study

As described in Section 4.2, we conducted a user study to evaluate our method against the baselines. Figure 13 shows an example question from the user study. Given a pair of source and target images, an original image, and two edited images: one produced by our method and the other by a baseline method. Participants were asked to select the image exhibiting superior editing quality while preserving the original identity. The results are presented in Table 2.

H Societal Impact

Similar to existing image editing techniques, our approach enables users to effectively edit images by optimizing LoRA weights of large-scale pre-trained diffusion models. By allowing individuals to manipulate images using their own data, this method supports a wide range of applications, such as novel content generation and artistic creation. Despite these positive outcomes, the use of generative models also introduces risks, including the creation of misleading or false information. To address these concerns, it is essential to advance reliable detection methods for distinguishing real images from synthetic ones [59, 10].

I Licenses for Pre-trained Models and Datasets

Our implementation is based on the publicly available FLUX.1-dev, which is licensed under the FLUX.1-dev Non-Commercial License. Most of the images used for evaluation are created using FLUX.1-dev and SDEdit [37]. Some image pairs are collected from the web or sourced from [27]. The license information for these images is not available online.

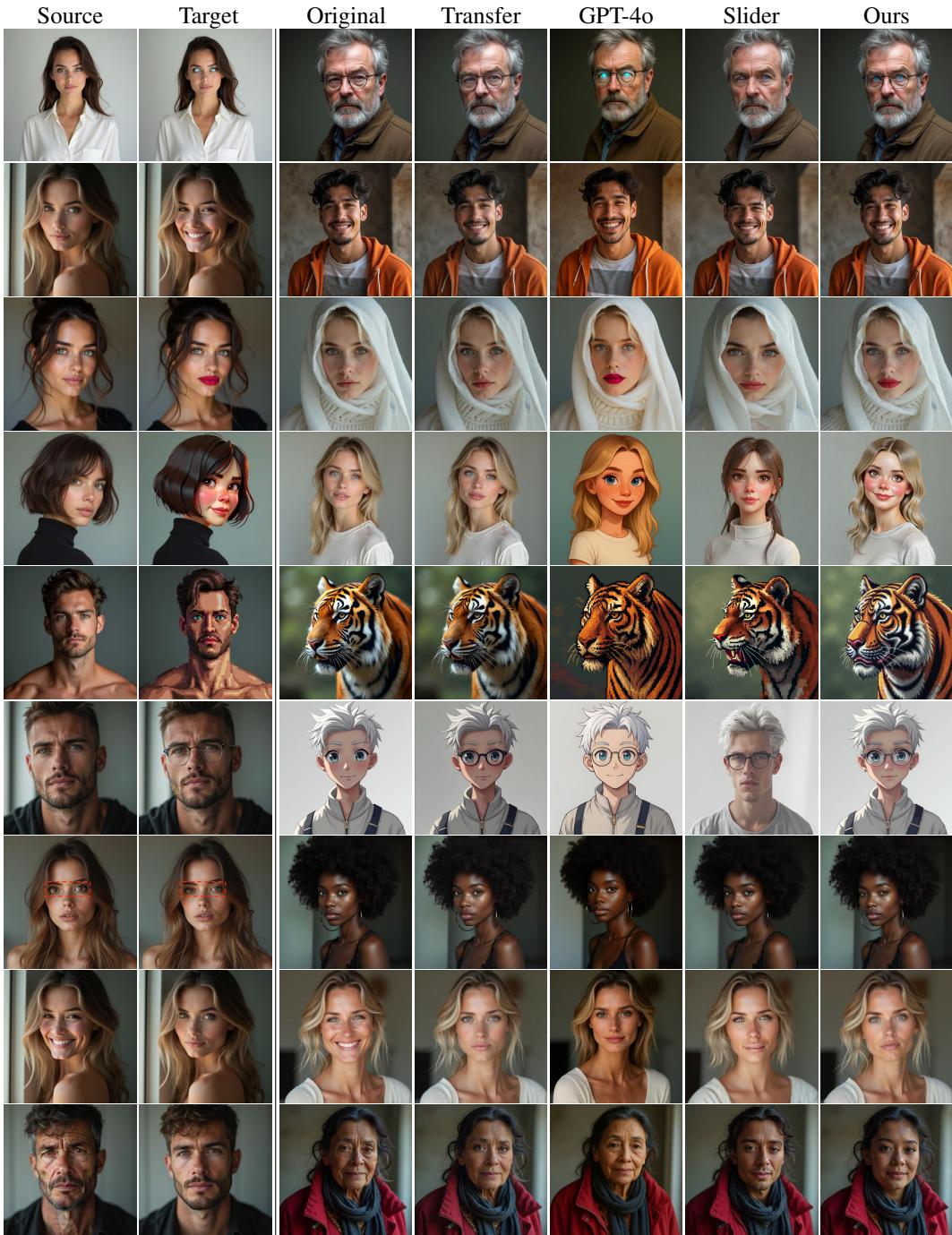


Figure 8: **Additional qualitative comparison.** We present exemplar-based image editing results from our method and three baseline methods: Edit Transfer [8], GPT-4o, and Slider [17]. Our method demonstrates superior performance in accurately editing the original image while preserving its content.

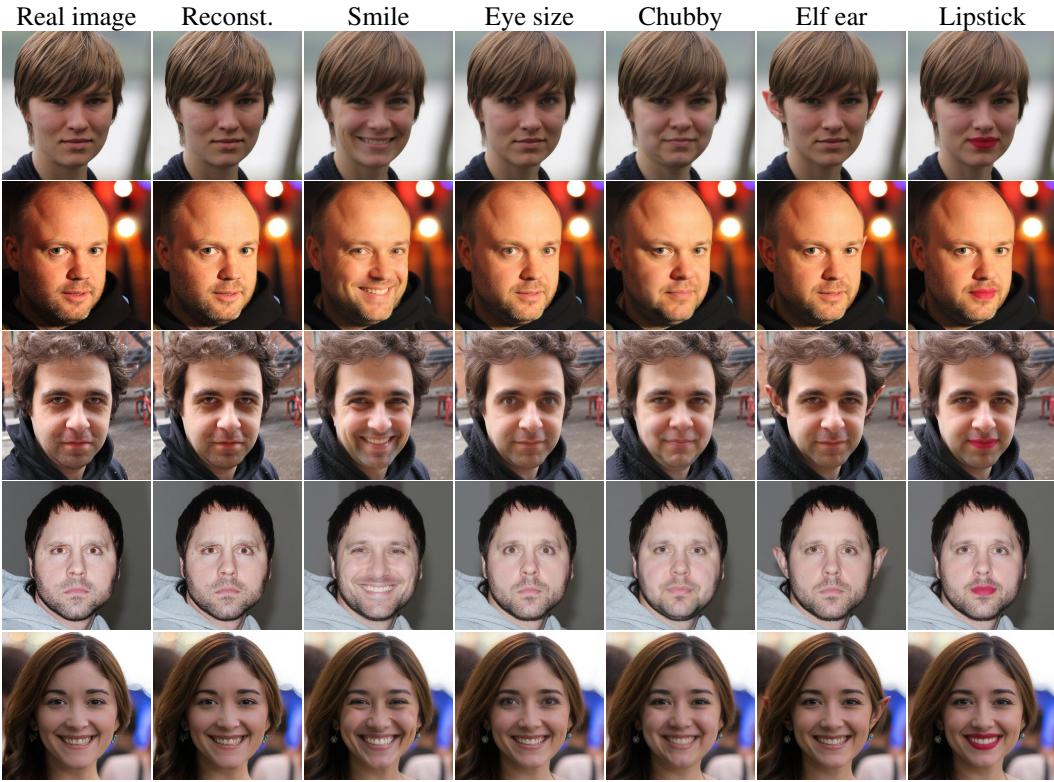


Figure 9: **Additional real image editing results.** The reconstructed image is obtained by optimizing a LoRA on the real image. We apply the learned semantic LoRAs to the reconstructed image by merging the LoRAs during inference.

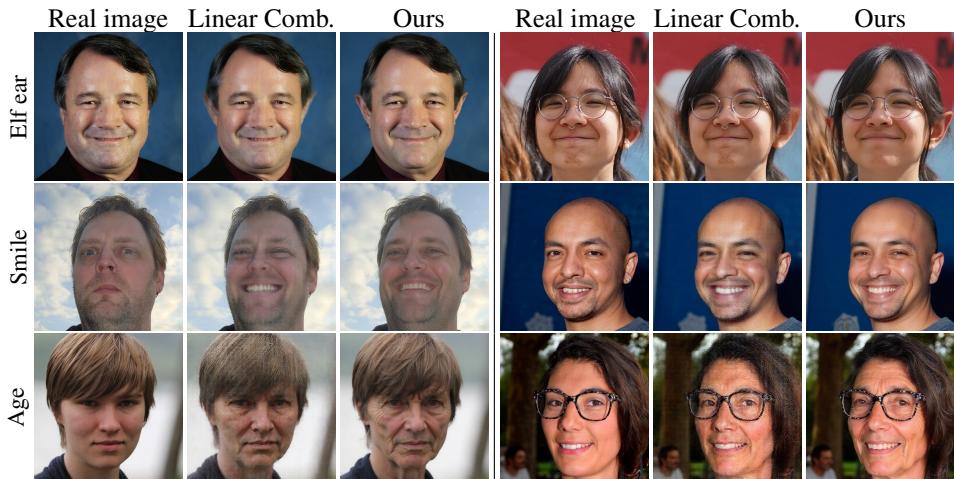


Figure 10: Comparison of two LoRA fusion methods: (1) linear combination of LoRA weights and (2) guidance-based LoRA fusion (Eq. 12). Guidance-based LoRA fusion achieves better identity preservation, whereas linear combination of LoRA weights tends to generate blurry images for certain semantics.

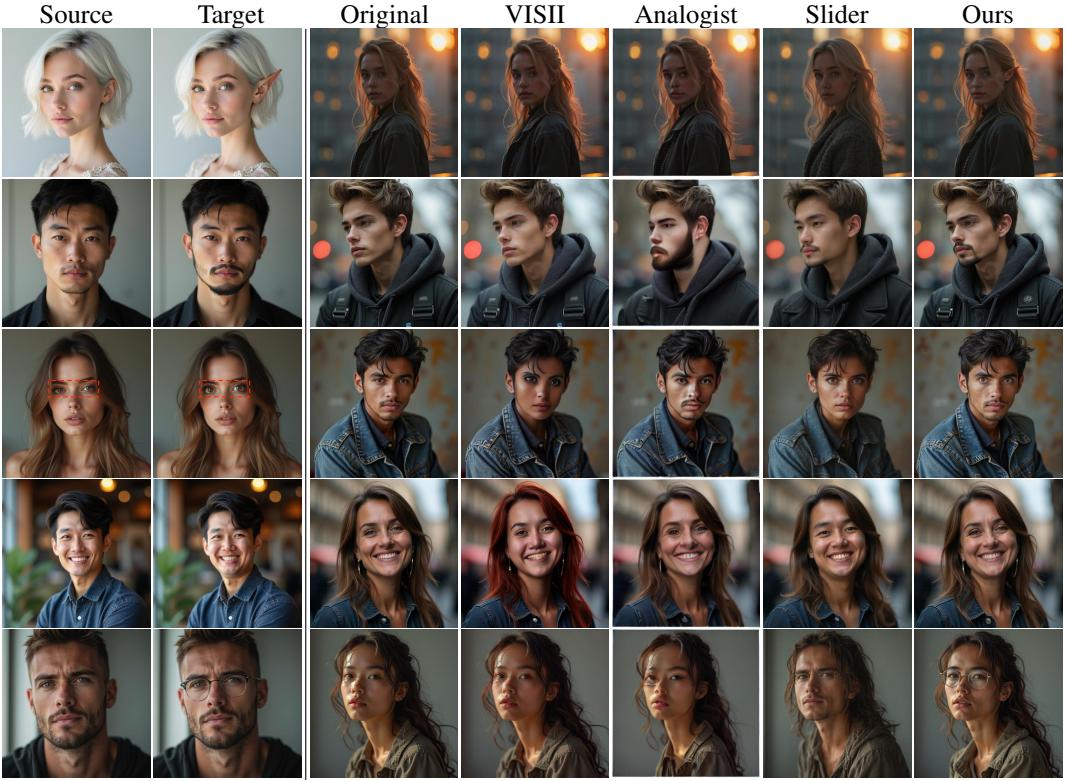
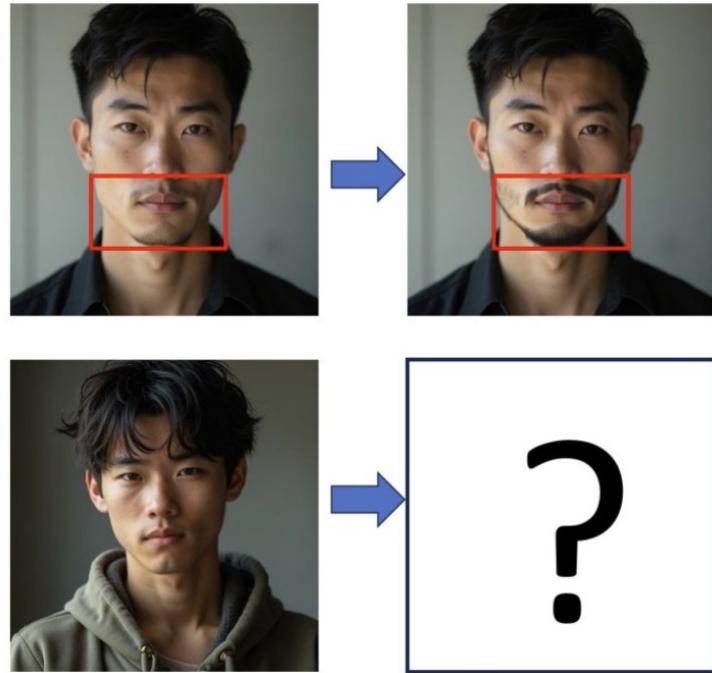


Figure 11: Comparison of PairEdit with three baseline methods under a single-image-pair training setting. Our method demonstrates superior performance in both identity preservation and semantic fidelity.



Figure 12: **Additional ablation study results.** We evaluate three variants of our model: (A) replacing the semantic loss with the visual concept loss proposed in [17], (B) removing the content LoRA, and (C) replacing the content-preserving noise schedule with a standard noise schedule.



First Row of Images: Observe the editing effects from the left image to the right image (highlighted in the red box).

Our goal is to apply the editing effect from the first row to the image in the second row.

Please choose the image below that achieves superior editing quality while preserving the original identity.



Figure 13: An example question from the user study. Given a pair of source and target images, along with an original image and two edited images, participants were asked to select the image that demonstrated superior editing quality while preserving the original identity.