ITERDIFF: TRAINING-FREE ITERATIVE FACE EDITING VIA EFFICIENT CLIP-GUIDED MEMORY BANK

Chun-Yao Chiu[†] *Feng-Kai Huang*[‡] *Teng-Fang Hsiao*[†] *Hong-Han Shuai*[†] *Wen-Huang Cheng*[‡]

[†]Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University
[‡]Department of Computer Science and Information Engineering, National Taiwan University

Fig. 1: Comparisons of iterative editing methods: The baseline method, ip2p [1] (top row), introduces noticeable structural distortions during complex edits such as "Put on a scarf" and "Add a beard to the face." In contrast, our approach, IterDiff (bottom row), maintains natural and consistent transformations across all depicted changes, demonstrating enhanced coherence and realism in the edits.

ABSTRACT

The rise of generative models has transformed image generation and editing, enabling high-quality, user-guided outputs. Iterative face editing, essential for applications like virtual makeup and entertainment, allows users to refine images progressively. However, this process often leads to artifact accumulation, semantic inconsistency, and quality degradation over multiple edits. Existing methods, while effective in single-step modifications, struggle with sequential edits. To robustly maintain fidelity and consistency in iterative face editing across multiple sessions, we propose *IterDiff*, a training-free framework leveraging diffusion models with a novel Training-Free Feature Preservation (TF^2P) approach to tackle these challenges by storing and retrieving key-value (KV) pairs from self-attention layers. Additionally, we further improve its efficiency and feasibility by Efficient CLIP-guided Memory Bank (ECMB). Experiments on the proposed benchmark show that IterDiff excels in prompt alignment, content consistency, and image quality, providing a robust solution for iterative facial attribute editing. Code, dataset and supplementary materials are available at `https://github.com/david20571015/IterDiff`.

***Index Terms*—** Diffusion Model, Image Editing

1. INTRODUCTION

Image editing is a pivotal component of modern digital workflows, enabling users to manipulate and refine images for diverse applications. While recent breakthroughs in generative models, particularly diffusion models [2, 3], have significantly advanced single-step editing, real-world scenarios often require iterative editing. In this paradigm, images undergo sequential refinements—akin to professional tools such as Photoshop-should still achieve high fidelity. Face editing exemplifies the practical importance of this approach: it is broadly useful in areas like digital entertainment and virtual try-ons, yet remains highly sensitive to errors, as even subtle inconsistencies can adversely affect identity, structural integrity, and overall realism. For instance, consider Fig. 1 where iterative editing can distort the image structure or modify attributes that are unrelated to the targeted changes.

To facilitate the iterative editing, a recent line of studies

has been proposed. [4] first predicts a structured semantic panel for synthesizing the final image, containing object-level attributes such as position, size, and color. This design enables iterative editing by modifying the semantic panel incrementally that users can add or adjust specific objects using prompts while preserving the rest of the scene. However, it requires extensive fine-tuning of diffusion models, leading to significant computational overhead. Another approach employs mask-based editing in the latent space [5], offering spatially targeted modifications through multi-granular control. However, mask-based editing often becomes ineffective for face editing since the target frequently involves the entire face, making the mask redundant.

To maintain essential information across iterative edits, we draw inspiration from recent studies [6, 7], which demonstrate that key-value (KV) pairs extracted from the self-attention layers of diffusion models can capture detailed semantic and spatial features. While prior works have improved facial detail preservation in face editing [8, 9], they focus on single-step modifications and overlook the challenges of distribution shift and information loss in iterative edits. In response, we propose the *Training-Free Feature Preservation (TF^{2}P)* approach, utilizing a memory bank to store and retrieve KV pairs. This enables the reuse of pertinent features in subsequent edits without compromising new modifications. Additionally, to reduce memory overhead and mitigate diminishing effects of edits, we introduce the the Efficient CLIP-guided Memory Bank (ECMB) strategy. This method employs CLIP similarity scores to selectively preserve crucial information, minimizing interference with ongoing edits. Moreover, recognizing the lack of public datasets for evaluating iterative face editing, we present *IterEditBench*, a dedicated benchmark specifically designed for this purpose. Overall, our main contributions can be summarized as follows.

- We propose a novel diffusion-based *Training-Free Feature Preservation (TF^{2}P)* that enhances the consistency and quality of iterative face editing while maintaining editing capabilities.
- Moreover, we introduce the *Efficient CLIP-guided Memory Bank (ECMB)* that stores and employs the most important features in each editing based on CLIP similarity to resolve the challenge of memory cost.
- We build a new benchmark, *IterEditBench*, of iterative face editing and demonstrate the effectiveness of our method, achieving state-of-the-art performance.

2. PRELIMINARY

2.1. Related Work

Recent advancements in text-to-image (T2I) models [3, 10, 11] have catalyzed the development of innovative image editing techniques across style transfer [12, 13], shape modification [7, 14], and attribute alteration [1, 15, 16, 17]. In the context of iterative image editing, traditional single-step tools such as InstructPix2Pix [1] struggle with consistency and artifact accumulation over successive edits. While approaches like EMILIE [5] and Ranni [4] offer improvements using latent iteration and semantic panels, they face challenges in granularity and efficiency. In contrast, the proposed TF^{2}P approach and ECMB strategy to dynamically manage semantic features and maintain edit quality.

2.2. Problem Formulation

Iterative face editing requires a model to perform sequential modifications to an image based on step-by-step instructions. The goal is to ensure that each modification aligns with the user's input while preserving the semantic and structural consistency of the image across edits. Formally, given an initial image I^0, a sequence of editing instructions $\{c^1_{text}, c^2_{text}, \ldots, c^n_{text}\}$, and a pre-trained diffusion model, the task is to iteratively generate edited images $\{I^1, I^2, \ldots, I^n\}$, where I^i is the output of applying c^i_{text} to I^{i-1}. Each edit must maintain consistency with prior outputs while accurately reflecting the current instruction.

3. ITERDIFF FRAMEWORK

Fig. 2(a) presents the overview of the proposed IterDiff framework, which builds upon the pre-trained InstructPix2Pix model [1] and leverages a carefully designed *Training-Free Feature Preservation (TF^{2}P)* approach to dynamically store and retrieve semantic features. Additionally, the proposed Efficient CLIP-guided Memory Bank (ECMB) strategy prioritizes relevant features, ensuring consistency and scalability across editing iterations.

3.1. Training-Free Feature Preservation

InstructPix2Pix [1] augments the LDM framework by extending the input layer to accept an additional image condition c_{image} that is encoded from the input image I via $\mathcal{E}$, allowing it to perform text-guided edits on existing images. Together with the editing instruction c_{text}, this additional condition guides the generation of the edited latent z_0, which is decoded into the final edited image by the decoder $\mathcal{D}$.

Build upon InstructPix2Pix, to address the challenges of iterative face editing due to the accumulation of artifacts and loss of content fidelity over multiple iterations, we introduce *Training-Free Feature Preservation (TF^{2}P)* approach. Previous work shown that key-value (KV) pairs extracted from the self-attention layers of diffusion models can capture detailed semantic and spatial features. Therefore, we propose to store the KV pairs in a Memory Bank $\mathcal{M}$ as follows:

$$\mathcal{M}^i \leftarrow \mathcal{M}^{i-1} \cup \{e^i_t \mid t \in \mathcal{T}\}, \tag{1}$$

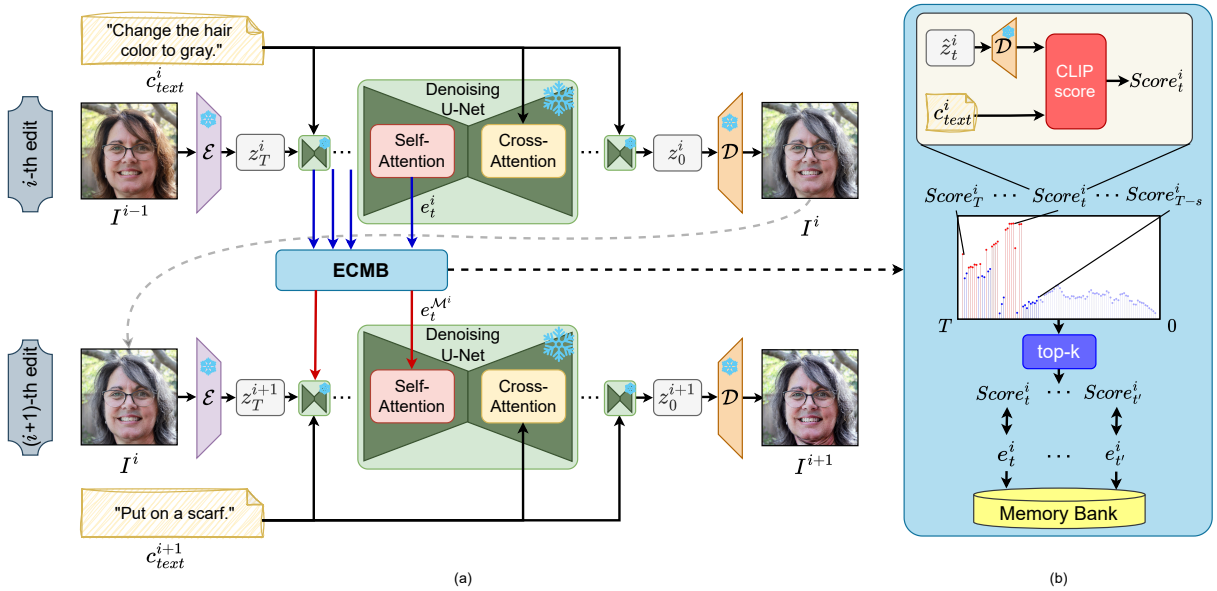


Fig. 2: IterDiff framework. (a) An overview of the iterative editing process, highlighting how the Memory Bank is integrated into the framework and where saving (blue arrows) and reading (red arrows) are applied. (b) Efficient CLIP-Guided Memory Bank mechanism, where key-value pairs are prioritized based on CLIP similarity before being stored in the Memory Bank.

where $\mathcal{M}^0$ is initialized as a empty set, $i \geq 1$ and denotes the i-th edit, $e^i_t = (K^i_t, V^i_t)$ represents the entry consisting of the key K^i_t and value V^i_t at timestep t, and $\mathcal{T}$ denotes all timesteps $\{0, 1, \ldots, T\}$.

Equipped with the memory bank, the fidelity and consistency of iterative face editing can be enhanced. By dynamically utilizing the Memory Bank $\mathcal{M}$ to recall and apply relevant KV pairs for each edit, we reduce the introduction of unwanted artifacts and the degradation of content integrity that typically occur with repeated modifications.

3.2. Efficient CLIP-Guided Memory Bank

Although TF^{2}P can better maintain the image fidelity, it results in excessive memory consumption and introduces redundant information. To this end, IterDiff integrates the *Efficient CLIP-Guided Memory Bank* to retain only the most relevant pairs, ensuring scalability and efficiency without compromising semantic consistency. Specifically, Fig. 2(b) illustrates the proposed ECMB, which contains two complementary operations: *Saving* and *Reading*. *Saving* focuses on updating the Memory Bank $\mathcal{M}^i$ during the i-th edit while *Reading* retrieves relevant key-value pairs during the $(i+1)$-th edit. This structure ensures that semantic features are effectively retained and reused across iterations. Below, we detail the *Saving* and *Reading* processes.

Saving. Fig. 2(b) illustrates how the Memory Bank $\mathcal{M}$ is updated during the *saving phase*. ECMB prioritizes KV pairs based on their semantic relevance to the current editing instruction, measured by using CLIP similarity [18]. Specifically, at timestep t, the importance score of the KV pair of the i-th editing, denoted by $score^i_t$, is computed by:

$$score^i_t = \text{CLIP}(\mathcal{D}(\hat{z}^i_t), c^i_{text}), \tag{2}$$

where $\hat{z}^i_t$ represents the estimation of the final noise-free latent z^i_0 at timestep t. During the denoising process, the diffusion model predicts this latent directly from the current input z^i_t. Afterward, the noise scheduler uses $\hat{z}^i_t$ to compute z^i_{t-1}, which serves as the input for the next step.

After calculating scores of $\hat{z}^i_t$ for all timesteps $t \in [T - s, T]$, where s is a hyper-parameter for determining the range where *ECMB* is applied[1], the top-k pairs are selected:

$$\hat{\mathcal{T}}^i = \text{argtopK}(\{score^i_{T-s}, \ldots, score^i_T\}, k). \tag{3}$$

Here, argtopK$(\mathcal{X}, k)$ is the function returning the indices of the k largest elements in $\mathcal{X}$. The corresponding KV pairs for these indices are then either stored or updated in the memory bank $\mathcal{M}$. If $\mathcal{M}^{i-1}$ already contains indices from any timestep included in $\hat{\mathcal{T}}^i$, these are replaced by the new ones, i.e.,

$$\mathcal{M}^i \leftarrow \{e^{i-1}_t \mid e^{i-1}_t \in \mathcal{M}^{i-1}, t \notin \hat{\mathcal{T}}^i\} \cup \{e^i_t \mid t \in \hat{\mathcal{T}}^i\}. \tag{4}$$

Additionally, we update the set of the timesteps in the memory bank, denoted by $\mathcal{O}^i$, by letting $\mathcal{O}^i \leftarrow \mathcal{O}^{i-1} \cup \hat{\mathcal{T}}^i$, i.e., recording all the timesteps stored for $\mathcal{M}^i$. In other words, we can use the timestep $t \in \mathcal{O}^i$ as the index to retrieve the KV

[1]The value of s is usually smaller than T because in diffusion models, the final steps mainly add details [19], making it unnecessary to store KV pairs from earlier steps.

pair in the memory bank $\mathcal{M}^i$ by $\mathcal{M}^i(t)$. By maintaining $\mathcal{O}^i$, we can retrieve the stored KV pairs for the *reading phase* to guide future iterations.

Reading. During the $(i+1)$-th edit, at each timestep t, we retrieve the KV pair stored in $\mathcal{M}^i$ to replace the current KV pair in the U-Net if applicable:

$$(K_t^{i+1}, V_t^{i+1}) \leftarrow \begin{cases} \mathcal{M}^i(t), & \text{if } t \in \mathcal{O}^i \\ (K_t^{i+1}, V_t^{i+1}), & \text{otherwise.} \end{cases} \quad (5)$$

Additionally, to counteract the potential loss of editing strength caused by KV pair replacements, we multiply the guidance scale by a factor $g^i = T/(T - |\mathcal{M}^i|)$. This adjustment is applied during steps where the KV pairs are not replaced, ensuring editing performance across iterations.

4. EXPERIMENTS

4.1. IterEditBench Dataset

As no existing benchmarks focus on iterative face editing, we introduce a self-constructed benchmark dataset named **IterEditBench**, specifically designed for real-world iterative face editing tasks. The dataset consists of 1000 samples and each sample constructed by two components: (1) **Base Image**: A high-resolution facial image randomly selected from the FFHQ dataset [20], which provides diverse and high-quality facial images suitable for editing tasks. (2) **Editing Instructions**: A sequence of five instructions randomly sampled from a predefined prompt set. Here, we generate the prompt set by OpenAI's ChatGPT to include a variety of tasks (see Appendix 1). In each sample, the editing process begins with the base image. Instructions are applied sequentially, where each instruction modifies the output of the previous step. This setup simulates a realistic iterative workflow, ensuring that the model is tested for both consistency and adaptability across diverse editing scenarios.

4.2. Experimental Setup

Implementation Details. We adopt the official pre-trained InstructPix2Pix model [1] as our backbone, keeping all hyperparameters at their default values. To further improve the locality of facial edits, we incorporate S-CFG [21] into the InstructPix2Pix framework. For the proposed IterDiff, we empirically set $s = 40$ and $k = 20$ to strike a balance between memory efficiency and editing performance.[2] All images are resized to 512×512 pixels. We focus on attention maps of sizes 32×32, 16×16, and 8×8, which are from deeper layers of the model, to capture high-level semantic structures. This ensures that edits remain semantically meaningful. Additional experiments examining the redundancy of TF^{2}P and the impact of guidance factors are presented in Appendix 2.

[2]The sensitivity test of the hyperparameters can be found in Appendix 2

Baselines. For comparative analysis, we include state-of-the-art training-free methods such as InstructPix2Pix [1], InstructPix2Pix with S-CFG [21][3] and EMILIE [5][4] as baselines in our evaluations. S-CFG dynamically adjusts the guidance scale by leveraging self-attention and cross-attention maps to identify key regions for different tokens, improving editing precision. EMILIE, on the other hand, performs iterative editing directly in the latent space without decoding through the VAE, enabling more efficient and consistent updates across editing steps.

Metrics. To evaluate the performance of IterDiff, we employ the following metrics to measure content consistency, i.e., how well the edited images retain the content and identity of the input images, and overall image quality, i.e., the realism and coherence of the edited images.

- **CLIP-I** [18] evaluates semantic similarity in the CLIP embedding space , with higher values indicating better consistency.
- **LPIPS** [22] assesses perceptual similarity, where lower scores reflect better preservation of visual details.
- **Image Reward** [23] is a learned scoring model trained to align with human preferences for text-to-image generation. Higher scores indicate better alignment with textual prompts, greater visual realism, and improved overall aesthetic quality.

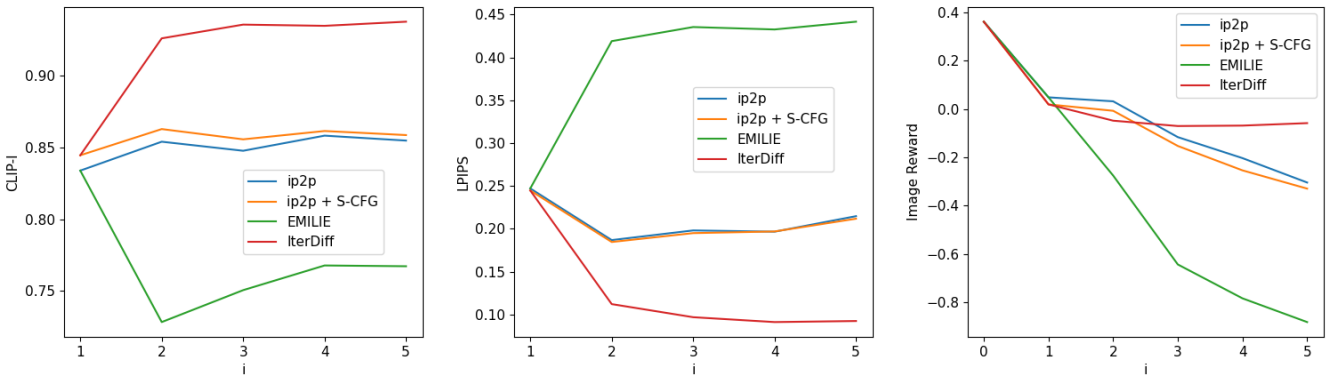
For metrics requiring image pairs (CLIP-I, LPIPS), comparisons are made between I^{i-1} and I^i. For metrics requiring prompts (Image Reward), we use *"A human face"* for I^0 and *"A human face edited with prompt: "*$\{c_{text}^i\}$*""* for I^i.

4.3. Quantitative Evaluation

The evaluation results summarized in Fig. 3, highlight the relative strengths of the compared methods across various metrics. In terms of content consistency (CLIP-I in Fig. 3(a) and LPIPS in Fig. 3(b)), IterDiff outperforms all other methods in terms of CLIP-I and achieves the lowest LPIPS score, highlighting its ability to effectively preserve the content and structure of input images during iterative edits. Notably, while other methods such as ip2p and EMILIE show a clear degradation in CLIP-I and LPIPS as the number of edits increases, IterDiff maintains stable performance. Furthermore, IterDiff clearly surpasses both InstructPix2Pix and its S-CFG variant in preserving structural and semantic consistency across edits. This stability can be attributed to the design of our memory bank. As the distribution shifts with each iterative edit, the memory bank retains earlier KV pairs that are less affected

[3]S-CFG is originally built on SD [3]; we have modified it to adapt to InstructPix2Pix.

[4]Since the official implementation is not publicly available, we have implemented the method based on the details provided in the paper.



(a) (b) (c)

Fig. 3: Quantitative curves in the i-th editing of (a) CLIP-I, (b) LPIPS, and (c) Image Reward for InstructPix2Pix (ip2p), InstructPix2Pix + S-CFG, EMILIE, and IterDiff.

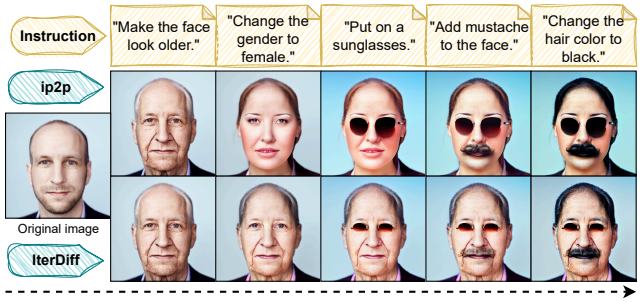


Fig. 4: Qualitative comparisons of ip2p and IterDiff

by these shifts. This mechanism mitigates the cumulative distribution drift by anchoring the editing process to these more stable reference points. In terms of image quality, IterDiff achieves the highest ImageReward score, indicating its superior perceptual quality relative to other methods.

4.4. Qualitative Evaluation

The qualitative evaluation (Fig. 1 and Fig. 4) of the two methods, IterDiff and InstructPix2Pix, reveals significant differences in their ability to perform iterative face editing tasks based on textual instructions. IterDiff consistently demonstrates a more refined and realistic approach to edits, preserving the original identity and structure of the person while seamlessly applying modifications. For example, as shown in Fig. 4, when instructed to *"Change the gender to female,"* IterDiff effectively transitions the facial features and overall appearance to accurately reflect a female version of the person, while preserving the original identity and ensuring the transformation appears natural and proportional. In contrast, InstructPix2Pix, despite capturing the intent of the instruction, not only changes the gender but also inadvertently makes the individual appear significantly younger. This age alteration often results in a less realistic outcome, as it introduces disproportionate changes that stray from the subject's original features. Likewise, when given the instruction *"Change the hair color to black,"* InstructPix2Pix darkens not only the hair but also the color of the sunglasses, which is an unintended modification. Meanwhile, IterDiff successfully modifies only the hair color without affecting unrelated elements in the image, demonstrating better localized editing control.

Acknowledgments

This work is partially supported by the National Science and Technology Council, Taiwan under Grants NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3.

5. CONCLUSION

In this work, we propose *IterDiff*, a training-free framework for iterative face editing using diffusion models. To preserve identity across edits, we introduce *Training-Free Feature Preservation* (TF^{2}P), which avoids additional training or fine-tuning. We also design an *Efficient CLIP-guided Memory Bank (ECMB)* to select the most relevant key-value (KV) pairs, enabling semantic consistency and reducing memory overhead. IterDiff achieves high-quality editing with minimal artifacts across multiple iterations. Experimental results on the *IterEditBench* dataset demonstrate that IterDiff outperforms existing methods in both identity preservation and image fidelity. With its lightweight design and robust generalization, IterDiff provides a solution for scalable, iterative editing. Future work may explore extending this framework to general object and scene editing, as well as dynamically managing KV pairs to further enhance efficiency.

6. REFERENCES

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[4] Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou, "Ranni: Taming text-to-image diffusion for accurate instruction following," in *CVPR*, 2024.

[5] KJ Joseph, Prateksha Udhayanan, Tripti Shukla, Aishwarya Agarwal, Srikrishna Karanam, Koustava Goswami, and Balaji Vasan Srinivasan, "Iterative multi-granular image editing using diffusion models," in *WACV*, 2024.

[6] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon, "Training-free consistent text-to-image generation," *TOG*, 2024.

[7] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng, "Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *ICCV*, 2023.

[8] Savas Ozkan and Mete Ozay, "Towards better control of latent spaces for face editing," in *ICIP*, 2024.

[9] Wendi Liang, Yihan Wen, Zewei Wang, Jianuo Jiang, Tat-Ming Lok, and Guanchong Niu, "Enhanced facial restoration with misinformation-filtered guide-denoising diffusion probabilistic models," in *ICIP*, 2024.

[10] 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," in *ICLR*, 2024.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al., "Scaling rectified flow transformers for high-resolution image synthesis," in *ICML*, 2024.

[12] 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 *CVPR*, 2024.

[13] Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong, "Z*: Zero-shot style transfer via attention reweighting," in *CVPR*, 2024.

[14] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai, "Dragdiffusion: Harnessing diffusion models for interactive point-based image editing," in *CVPR*, 2024.

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or, "Prompt-to-prompt image editing with cross-attention control," in *ICLR*, 2023.

[16] 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.

[17] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos, "Ledits++: Limitless image editing using text-to-image models," in *CVPR*, 2024.

[18] 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.

[19] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu, "Freeu: Free lunch in diffusion u-net," in *CVPR*, 2024.

[20] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[21] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu, "Rethinking the spatial inconsistency in classifier-free diffusion guidance," in *CVPR*, 2024.

[22] 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.

[23] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *NeurIPS*, 2024.