CA-Edit: Causality-Aware Condition Adapter for High-Fidelity Local Facial Attribute Editing

Xiaole Xian[1*], Xilin He[1*], Zenghao Niu[1], Junliang Zhang[1], Weicheng Xie[1,3†], Siyang Song[4], Zitong Yu[5], Linlin Shen[1,2,3]

[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University,
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University,
[3]Guangdong Provincial Key Laboratory of Intelligent Information Processing,
[4]University of Exeter,
[5]Great Bay University

Abstract

For efficient and high-fidelity local facial attribute editing, most existing editing methods either require additional fine-tuning for different editing effects or tend to affect beyond the editing regions. Alternatively, inpainting methods can edit the target image region while preserving external areas. However, current inpainting methods still suffer from the generation misalignment with facial attributes description and the loss of facial skin details. To address these challenges, (i) a novel data utilization strategy is introduced to construct datasets consisting of attribute-text-image triples from a data-driven perspective, (ii) a Causality-Aware Condition Adapter is proposed to enhance the contextual causality modeling of specific details, which encodes the skin details from the original image while preventing conflicts between these cues and textual conditions. In addition, a Skin Transition Frequency Guidance technique is introduced for the local modeling of contextual causality via sampling guidance driven by low-frequency alignment. Extensive quantitative and qualitative experiments demonstrate the effectiveness of our method in boosting both fidelity and editability for localized attribute editing. Our codes will be made publicly available.

Introduction

Efficient and high-fidelity local facial attribute editing with textual description represents a challenging task in computer vision. GANs-based methods (Wang et al. 2022; Pernuš, Štruc, and Dobrišek 2023) have explored this task, which primarily optimize the original image within the latent space with a pre-trained StyleGAN model (Karras et al. 2020). However, these GANs-based methods require additional fine-tuning for different attributes. Subsequently, the prior diffusion-based image editing methods based on the text-to-image (T2I) diffusion models achieve image editing in various ways. These methods are either based on P2P (Hertz et al. 2022), utilizing the original image attention injection mechanism to preserve the layout, or based on DDIM Inversion (Song, Meng, and Ermon 2020), modifying the latent at the noise level. However, such methods may lead to inconsistencies beyond the editing target area. Regarding local facial attribute editing, image inpainting is a technique focused on local masked region painting, which also benefits from the recent advances in diffusion models (Avrahami, Lischinski, and Fried 2022; Yang et al. 2023; Yang, Chen, and Liao 2023). Besides, image inpainting has been also developed for local facial attribute editing, which focuses on the inpainting of local masked regions, based on advanced diffusion models (Avrahami, Lischinski, and Fried 2022; Yang et al. 2023; Yang, Chen, and Liao 2023). Text-guided image inpainting(Avrahami, Lischinski, and Fried 2022) allows prompt-driven content generation in specific areas without finetuning during inference, while maintaining consistency between the editing and unmasked regions, which is thus used in our method.

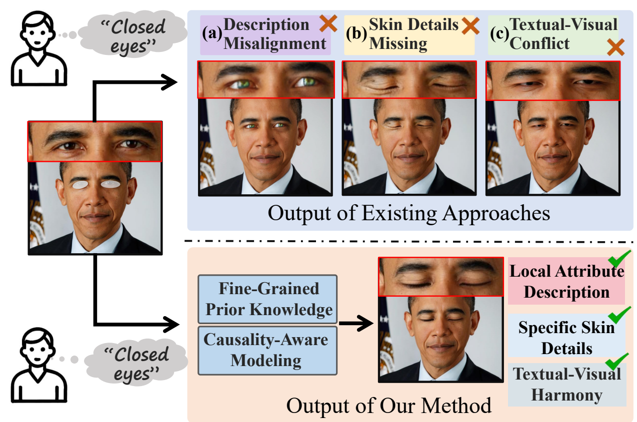

Figure 1: **(Top)** The existing text-guided inpainting pipeline for our local attribute editing task. **(Bottom)** Our method takes account of the causality of the the specific details from the original image, improving the editability and the fidelity.

However, existing methods for image inpainting may suffer from concerns in terms of editability and fidelity. **The first problem:** they (Zhang, Rao, and Agrawala 2023; Ju et al. 2024) struggle to understand the contextual relationship between unmasked facial regions and the textual de-

*These authors contributed equally.
†Corresponding author

scription, resulting in the neglect of the text prompt while creating a plain completion (Fig.1 (a)). For addressing this problem, Hd-Painter (Manukyan et al. 2023) can better align the inpainting generation with the text by modifying the latent, while it still fails for local facial text prompts. The root cause is that previous diffusion models are primarily trained on natural image-text pairs, lacking the fine-grained knowledge of human faces. **The second problem:** For facial inpainting, previous works (Rombach et al. 2022; Yang, Chen, and Liao 2023) do not take adequate consideration of the contextual causality between the masked region and the specific details (skin texture, skin tone, and other details) of the original image. The causality consideration is further constrained by the conflict between textual editing conditions and the preservation of these details in original image. In facial images, even slight differences in these details become visibly obvious, largely impairing the overall naturalness. (Fig.1 (b)) Therefore, the key to maintaining the skin details and mitigating the difference lies in the reasonably causality-aware modeling of these specific details from the original image.

For addressing this problem, existing approaches adapt the parallel attention with textual conditions(i.e. IP-Adapter (Ye et al. 2023)) to inject original image information and enhance contextual causality modeling. However, as shown in Fig. 1, this causality conflicts modeling with the text condition may lead to severe content leakage. (Fig.1 (c)) Meanwhile, from a localized contextual perspective, existing methods (Ju et al. 2024; Xu et al. 2024) lack explicit approaches for this fine-grained local context, causing disharmony in boundary regions of the primary editing regions, while the skin transitions are generally smooth.

To address these challenges, we proposed our CA-Edit from the local attribute data construction and causality-aware condition adapter. **For addressing the first problem**, training on detailed textual captions of local facial attributes would be crucial for editability. To this end, we introduce a data construction pipeline, leveraging Multimodal Large Language Models (MLLMs) (Chen et al. 2023; Li et al. 2023) for automatic local facial attribute captioning and the face parsing model for segmentation acquisition. **For addressing the second problem**, we introduce an additional adapter for original image condition, as well as a sampling guidance during inference, to fully explore original image cues. Specifically, (i) the Causality-Aware Condition Adapter (CA^2) is proposed to enhance the causality modeling while preventing the conflict with textual condition. (ii) a sampling guidance technique called Skin Transition Frequency Guidance (STFG) is proposed to mitigate the artifacts on the 'boundary regions' via enhancing the similarity between the generated image and the low-frequency components of the original image.

The main contributions of this work are summarized as:

- To address the limitations of existing datasets lacking local facial attribute captions, we propose LAMask-Caption, the first dataset with detailed local facial captions which contains 200,000 high-quality facial images and employs Large Multimodal Models (MLMMs) for automatic captioning of local facial regions.
- To jointly address the issues of fine-grained context modeling and content leakage, we propose the novel CA^2) that enhances contextual causality modeling in primary editing regions while regularizing the visual condition according to the textual condition and latent. Furthermore, we propose the novel STFG to preserve the skin details on the boundary regions by enhancing the low-frequency similarity with the original image during inference.
- Quantitative and qualitative experiments demonstrate that CA-Edit produces more harmonious and natural outcomes, showcasing the superiority of our method in local attribute editing.

Related Work

Generative Face Editing

The advancement of facial editing and manipulation has been promoted by the emergence of recent generative approaches. Early efforts in this area have explored the application of GANs-based models (Karras, Laine, and Aila 2019; Shen et al. 2020; Yang et al. 2021; Xia et al. 2021). MaskGAN (Lee et al. 2020) demonstrated the benefit of using spatially local face editing. InterFaceGAN (Shen et al. 2020) regularizes the latent code of an input image along a linear subspace. Recently, increasing researchers have resorted to diffusion models to enhance the generative capability for face editing. Methods like (Ding et al. 2023; Jia et al. 2023) both explored the use of 3D modalities as reference cues to make facial image editing more robust and controllable. Xu et al. (Xu et al. 2024) finetune a diffusion model for editing tasks tailored to the individual's facial characteristics. However, these approaches require extra conditions beyond text, limiting their suitability for our task due to user accessibility issues.

Text-driven image editing

Early works (Nitzan et al. 2022; Andonian et al. 2021; Xia et al. 2021) leveraging pretrained GAN generators (Karras, Laine, and Aila 2019) have explored the text-driven image synthesis. Among approaches for semantic image editing, text-guided image editing based on diffusion models has garnered growing attention. (Gal et al. 2022a; Ruiz et al. 2023; Rombach et al. 2022; Morelli et al. 2023; Mao, Wang, and Aizawa 2023; Zhong et al. 2023; Brooks, Holynski, and Efros 2023) have exploited diffusion models for text-driven image editing. Textual Inversion (Gal et al. 2022a) generates an image by learning a concept embedding vector combined with other text features. For better control of the original semantic cues, InstructPix2Pix (Brooks, Holynski, and Efros 2023) enables image editing based on textual instructions by leveraging a conditioned diffusion model trained on a dataset generated from the combined knowledge of a language model and a text-to-image model. DiffusionCLIP (Kim, Kwon, and Ye 2022) and Asyrp (Kwon, Jeong, and Uh 2022) draw inspiration from GAN-based methods (Gal et al. 2022b) that use CLIP, and use a local directional CLIP loss between image and text to manipulate images. However,

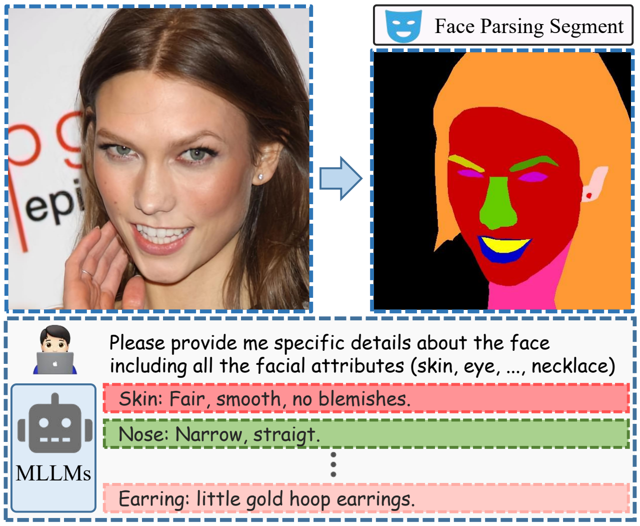

Figure 2: The pipeline of LAMask-Caption construction.

these methods either require additional finetuning or lead to changes outside target editing regions, which fail to meet the requirement of local editing.

Diffusion Models for Inpainting

Image inpainting is devoted to reconstructing or filling in the missing regions of an image in a visually coherent manner. Benefited from the pretrained T2I diffusion models, many prominent works (Avrahami, Lischinski, and Fried 2022; Yang et al. 2023; Ju et al. 2024; Lugmayr et al. 2022; Yang, Chen, and Liao 2023) that are zero-shot and do not affect the regions outside the edited area, were developed. Stable Diffusion Inpainting (Rombach et al. 2022) and ControlNet Inpainting (Zhang, Rao, and Agrawala 2023) both leverage large-scale pre-trained T2I models, fine-tune them to adapt models for this task. During inference, the method (Avrahami, Lischinski, and Fried 2022) removes noises in a weighted manner according to the mask at each time step, which can reduce the occurrence of unnatural artifacts. (Levin and Fried 2023) use a continuous mask rather than a binary mask, to enable fine-grained control over the diffusion of each pixel. Paint-by-example (Yang et al. 2023) uses image embedding to replace the original text embedding to improve image-to-image inpainting. However, due to the lack of image-text pairs of face attributes for training or adequate causality exploration in keeping the skin details, the inference stage of the aforementioned methods often results in artifacts.

Preliminaries

Diffusion Model. Diffusion models are a family of generative models that consist of the processes of diffusion and denoising. The diffusion process follows the Markov chain and gradually adds Gaussian noise to the data, transforming a data sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ into the noisy sample $\mathbf{x}_{1:T} = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ in T steps. The denoising process utilizes a learnable model to generate samples from this Gaussian noise distribution denoted as $p_\theta(\mathbf{x}_{0:T})$ at time step t based on the condition c, where θ denotes the learnable parameters. Eventually, the training of the model is formulated as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \boldsymbol{c}, t} \| \epsilon - \epsilon_\theta(\mathbf{x}_t, \boldsymbol{c}, t) \|_2^2, \tag{1}$$

where $\mathbf{x}_0$ denotes the original image, c, $t \in [0, T]$ represents the condition and the timestep of the diffusion process.

Reference Net for Diffusion Model. As introduced in BrushNet(Brooks, Holynski, and Efros 2023) and ControlNet(Zhang, Rao, and Agrawala 2023), a reference net is constructed by adding an additional branch dedicated to the spatial condition, which is well-suited for our task-specific mask generation. The additional condition is first encoded with the reference net, which is then added into the skipped connections of the Stable Diffusion (Rombach et al. 2022) UNet after being processed by zero convolutions. Eventually, the noise prediction of U-Net with the reference net is formulated as $\epsilon_\theta(\mathbf{x}_t, \boldsymbol{c}_{img}, \boldsymbol{c}_{txt}, t)$, where c_{img} and c_{txt} represent the image and text conditions, respectively.

Method

To enable local facial attributes inpainting, we first construct the dataset LAMask-Caption including the face images, textual descriptions of local facial attributes and the specific segmentation mask of the attributes (Fig. 2). To adapt the T2I model to our task, we trained a reference network copied from the U-Net. Based on this network, we introduced Causality-Aware Condition Adapter (CA^2) to enhance skin detail causality while balancing textual and visual cues for precise and seamless attribute editing. Additionally, to reduce the artifacts between generated content and the unmasked regions, our Skin Transition Frequency Guidance (STFG) technique further leverages the skin detail in the original image during inference, to avoid the effect of imprecise input masks.

LAMask-Caption Construction Pipeline

A key reason that current diffusion models encounter difficulties with local facial editing is the lack of precise textual captions describing local facial attributes in the training data, as mainstream diffusion models are primarily trained on large-scale natural image datasets such as Laion-2B (Schuhmann et al. 2022) or MS-COCO (Lin et al. 2014). Hence, a face dataset with local attributes-text pairs is essential for finetuning the pretrained diffusion model to adapt to facial local attribute editing. While the existing CelebA-dialog dataset (Jiang et al. 2021) and FaceCaption-15M (Dai et al. 2024) contain manually annotated textual captions for each image, it mainly focuses on overall attributes (i.e. age, skin) rather than local facial attributes. Therefore, their global captions would fail to meet the demand as training data of local facial attribute editing, which motivates us to develop a new dataset with complete local facial attribute captions.

Specifically, we introduce our LAMask-Caption, a dataset consisting the triples of detailed textual captions of local facial attributes, high-resolution images and attribute masks. The overview of our LAMask-Caption construction pipeline is shown in Fig. 2. Via this framework,

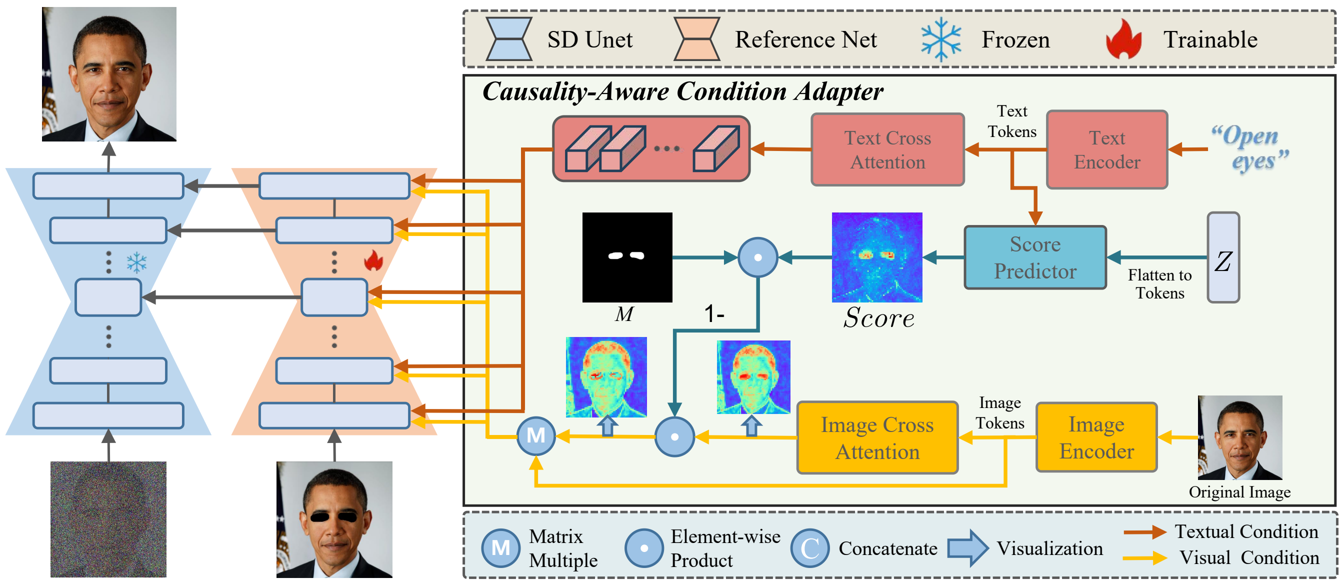

Figure 3: The training process of our method. The CA^2 in the Reference Net to inject specific skin details from the original image as image embedding via an additional attention mechanism.

we collect a high-quality facial image dataset comprising 200,000 high-quality images by combining filtered images from FaceCaption-15M with selections from FFHQ and CelebMask-HQ datasets.

We employ Multimodal Large Language Models (MLLMs) (Chen et al. 2023; Li et al. 2023) to generate local textual captions, encouraging diverse responses that describe the face images from various perspectives, including direct, indirect, and subjective perceptions. Additionally, we use a fine-tuned BiSeNet (Yu et al. 2018) to create segmentation masks for 19 facial attributes. Hereto, caption-mask pairs corresponding to local facial regions could be acquired, forming the core component of the proposed LAMask-Caption.

Causality-Aware Condition Adapter (CA^2)

One naive approach for injecting skin detail as a visual condition into a diffusion model is usually achieved through cross-attention, which requires parallel addition of cross-attention modules for the original image embedding, akin to IP-Adapter (Ye et al. 2023; Wang et al. 2024). However, we argue that the direct injection of visual cross-attention would lead to over-reliance on the visual condition during training while ignoring textual editing conditions (Jeong et al. 2024; Qi et al. 2024). To this end, we propose the novel Causality-Aware Condition Adapter (CA^2), as shown in Fig. 3, which injects specific skin details from the original image as image embedding through an additional attention mechanism, and adaptively adjusts the intensity of visual condition injection. The adjustment is conducted based on the influence of the textual prompt on the existing features, aiming to balance the impact of textual and visual conditions. The adapter encodes the contextual causality between the main editing region and specific skin details, while preventing visual-textual condition conflicts.

In our proposed CA^2, both the vision and text encoders of a pretrained CLIP are utilized for the feature extraction, formulated as:

$$\begin{cases} f_{txt} = CLIP_{txt}(txt) \in \mathbb{R}^{n_t \times c_t} \\ f_{vis} = CLIP_{vis}(x) \in \mathbb{R}^{n_v \times c_v} \end{cases} \tag{2}$$

where n_t, n_v denote the numbers of text and visual tokens, and c_t, c_v are the dimensions of text and vision tokens. $CLIP_{txt}, CLIP_{vis}$ are the CLIP text and vision encoders, respectively. x is the original image.

Subsequently, we intend to use the textual pooling token $f_{txt}^{pool} \in \mathbb{R}^{1\times c_t}$ along with the diffusion model's latent features $Z \in \mathbb{R}^{n_z \times c_z}$ to predict textual importance scores. We spatially replicate f_{txt}^{pool} to $f_{txt}^{s} \in \mathbb{R}^{nz\times c_t}$ to align the token numbers, where n_z is the token number of Z.

To obtain the score that is used to weight the importance of visual condition, a simple two-layer MLP with a softmax activation function is constructed as the score predictor. The score takes the concatenation of textual class token and diffusion latent features along the channel dimension as input and then predicted following:

$$Score = \mathcal{S}(\text{Concat}(Z, f_{txt}^{s})) \tag{3}$$

where $\mathcal{S}(\cdot)$ is the score predictor, $Score \in \mathbb{R}^{n_z}$ and then it will be reshaped to match the spatial dimension of the latent feature. Meanwhile, the visual cross-attention map is calculated as:

$$A_{vis} = \text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}_{vis})^{\top}}{\sqrt{d}}) \tag{4}$$

where $\mathbf{Q} = Z\,W^{Q}$, $\mathbf{K}_{vis} = f_{vis}W_{vis}^{K}$, are the query of latent feature Z and the key from vision feature f_{vis}, respectively. W^{Q} and W_{vis}^{K} are the corresponding weight matrices.

The query matrix of the vision feature is the same as that of text cross-attention. Pixels with higher textual importance scores should have their vision attention suppressed, as this indicates stronger textual editing. Conversely, pixels with lower scores should receive higher vision attention to enhance dependence on the original image. Therefore, we intend to suppress the vision attention values within the mask region according to the obtained Score as:

$$\begin{aligned} A^s_{vis} &= A_{vis} \odot (1 - Score \odot M) \\ F^s_{vis} &= A^s_{vis} \mathbf{V}_{vis} \end{aligned} \tag{5}$$

where $\odot$ denotes the Element-wise product, M is the input mask that has been downsampled to the same spatial resolution as the $Score$ prior to the flattened representation. $\mathbf{V}_{vis} = f_{vis} W^V_{vis}$ denotes the value of vision feature in cross-attention. Eventually, the latent feature processed by our CA^2 can be computed as:

$$\begin{aligned} F_{txt} &= \mathrm{Softmax}(\tfrac{\mathbf{Q}(\mathbf{K}_{txt})^\top}{\sqrt{d}}) \mathbf{V}_{txt} \\ Z_s &= F_{txt} + F^s_{vis} \end{aligned} \tag{6}$$

where $\mathbf{K}_{txt}$ and $\mathbf{V}_{txt}$ denote the key and value of f_{txt} in Eq. (2), respectively.

Skin Transition Frequency Guidance (STFG)

While CA^2 preserves skin details in the main editing areas, real-world facial editing often uses imprecise masks, leading to unnatural transitions in 'boundary regions'. These smooth skin areas are sensitive to low-frequency changes. To address this, we introduce a sampling guidance technique for low-frequency components during denoising, to produce natural transitions in these regions.

Specifically, given the localization and semantic representation capabilities of textual cross-attention maps in diffusion models to identify 'boundary regions'. The mean of attention maps, i.e., $\overline{A}_{txt}$ is computed across all text tokens and attention layers. We identify the 'boundary regions' as regions within the mask M where the attention values on $\overline{A}_{txt}$ are below a threshold $\gamma(\overline{A}_{txt}, M)$. The indexes Idx of the pixels belonging to 'boundary region' is represented as:

$$\begin{aligned} Idx &= \{(i,j) | \overline{A}_{txt}(i,j) \le \gamma(\overline{A}_{txt}, M)\} \\ \gamma(\overline{A}_{txt}, M) &= \mu(\overline{A}_{txt} \circ M) - \sigma(\overline{A}_{txt} \circ M) \end{aligned} \tag{7}$$

where $\overline{A}_{txt} \circ M$ represents the elements of $\overline{A}_{txt}$ within the mask M, $\mu(\cdot)$ and $\sigma(\cdot)$ denote the operators of mean and standard deviation.

We further employ frequency guidance in the Fourier domain to selectively enhance low-frequency similarity on the estimated latent, i.e., designing a guidance function to pixel-wisely align the low-frequencies between the original noisy latent z_t and the predicted latent $\widehat{z}_t$ on each timestep t. Since the frequency components should be calculated on the clean latent, we estimate the one-step prediction $\widehat{z}_{t\to 0}$ from $\widehat{z}_t$ as:

$$\widehat{z}_{t\to 0} = \frac{\widehat{z}_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\widehat{z}_t, t)}{\sqrt{\bar{\alpha}_t}} \tag{8}$$

where $\bar{\alpha}_t$ is the hyperparameter of noise schedule parameter. Subsequently, we only keep the low-frequency components ($\frac{H}{2} < h < \frac{3H}{4}$ and $\frac{W}{2} < w < \frac{3W}{4}$ in FFT shifted image) of $\widehat{z}_{t\to 0}$ and z_0 in the frequency domain to obtain $\widehat{z}'_{t\to 0}$ and z'_0, respectively. Consequently, the guidance function used to align these two can be defined as follows:

$$g(z'_0, \widehat{z}'_{t\to 0}) = \frac{1}{|Idx|} \sum_{(i,j)\in Idx} \left\| \widehat{z}'_{t\to 0}(i,j) - z'_0(i,j) \right\|_2^2 \tag{9}$$

where $|Idx|$ is the cardinality of the set Idx. We follow the score-based guidance (Song et al. 2020), and use $g(z'_0, \widehat{z}'_{t\to 0})$ to steer the diffusion process. Eventually, we can update the direction of $\hat{\epsilon}_t$ as follows:

$$\hat{\epsilon}_t = \epsilon_\theta(z_t, t, txt, x) - \lambda \rho_t \nabla_{z_t} g(z'_0, \widehat{z}'_{t\to 0}) \tag{10}$$

where λ is a hyperparameter of the guidance strength and ρ_t denotes the noise schedule parameter of timestep t.

Experiment

Evaluation Metric

Objective Metrics. To comprehensively evaluate the performance of different methods on the task of local facial attributes editing, we utilize **FID / Local-FID** (Heusel et al. 2017), **LPIPS** (Zhang et al. 2018), **identity similarity (ID)**, **MPS** (Zhang et al. 2024) and **HPSv2** (Wu et al. 2023) as evaluation metrics. **FID** and **LPIPS** are used to provide an estimate of image fidelity. It's important to note that in this specific task, unlike general image generation, lower LPIPS values indicate higher fidelity. **MPS** and **HPSv2** are more effective and comprehensive zero-shot objective evaluation metrics on text-image alignment and human aesthetics preferences. ID evaluates the face identity between the results and the original images.

Experimental Setup

Benchmark. As this work focuses on text-guided local facial attribute editing, we introduce *FFLEBench*, i.e., a pioneering benchmark evaluation dataset motivated by the lack of existing benchmarks. *FFLEBench* includes 15,000 samples from FFHQ, along with local masks and corresponding textual captions. Note that the samples used to construct *FFLEBench* are independent of those used for training. The masks are the convex hull or the dilation of the segmentation masks, aiming to imitate the rough mask input.

Quantitative Experiment Results

We quantitatively evaluate our method on FLEBench, compared with baseline models using both objective metrics and user study. As shown in Tab. 1, the proposed method surpasses the compared methods except for the FID of Stable Diffusion Inpainting. Particularly, better performance on FID and LPIPS indicates that our method can edit facial attributes with higher fidelity. Due to its tendency to neglect the text prompt to maintain high fidelity, SD inpainting exhibits the lowest FID score. Our approach outperforms on the MPS and HPS v2 metrics, indicating our edits align with human aesthetics and maintain textual consistency. All the

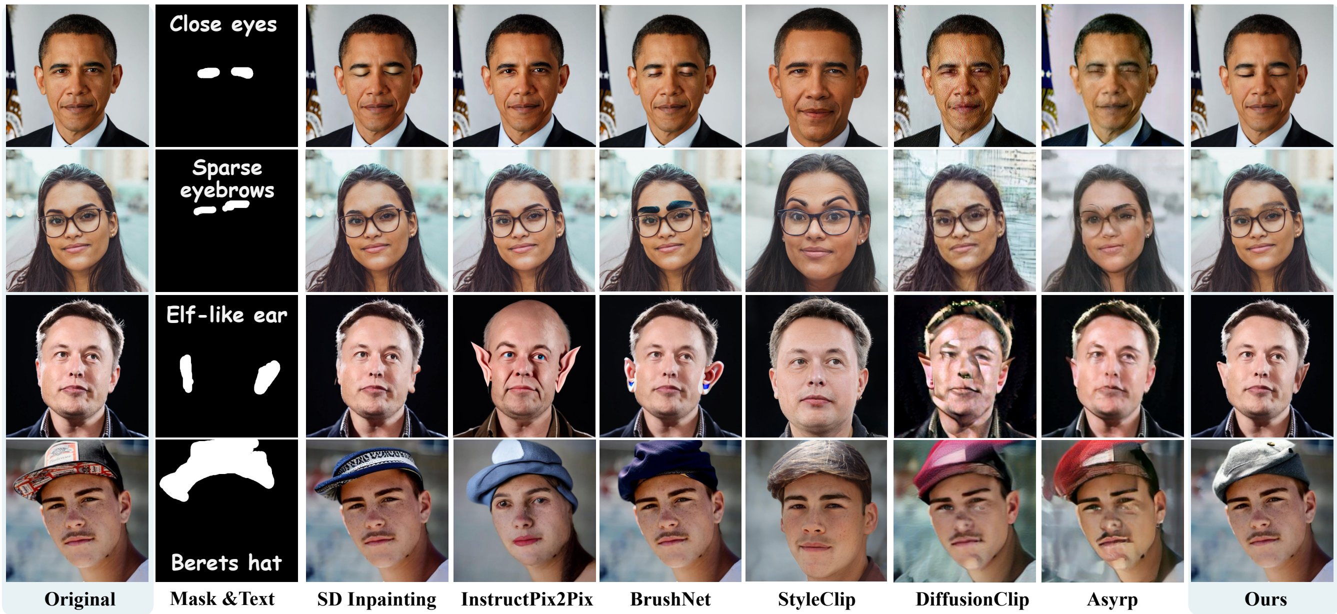

Figure 4: Qualitative comparison on local facial attributes editing. Compared with zero-shot methods (i.e. SD inpainting (Wang et al. 2022), InstructPix2Pix (Brooks, Holynski, and Efros 2023), BrushNet (Ju et al. 2024)) and the facial editing methods (StyleClip (Patashnik et al. 2021), Diffusionclip (Kim, Kwon, and Ye 2022), Asyrp (Kwon, Jeong, and Uh 2023)), our approach not only aligns the edited parts with the text prompts, but also better preserves the information from the original image.

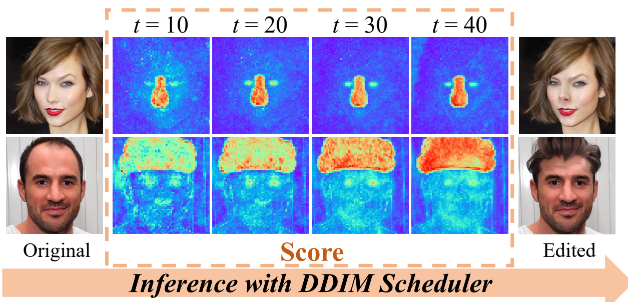

Figure 5: The visualization of the score in CA^2 during inference. The lighter regions indicate the higher values in the maps. The DDIM scheduler with $t = 50$ timesteps is used.

observation highlights the strength of our approach in preserving visual coherence and effectively capturing the textual guidance during the editing. In addition, our approach achieves better local attribute editing results without requiring the extra fine-tuning time for different attributes, which is needed by other facial editing methods (fine-tuning time shown in brackets after the method names).

In the user study, the percentages represent the proportion of users who prefer our method over others. As shown in Tab. 1, our method attains the top rank compared to the other inpainting and face editing methods.

Qualitative Experiment Results

Comparison with the SOTAs. In Fig.4, our method is qualitatively compared with the state-of-the-art (SOTA) methods across the local facial attributes, such as eyes, ears, and accessories. Other manipulation results are in the supplementary materials. Prompt neglect is an issue for other methods that sometimes struggle to modify local attributes according to textual descriptions, as evident in the "sparse eyebrows" example (second row). While they can capture text semantics in some cases, they miss the original images' specific skin details, compromising overall fidelity.

In addition, InstructPix2Pix, and the facial editing methods (i.e. StyleClip, DiffusionClip and Asyrp) exhibit undesirable content leakage into adjacent regions, resulting in effects beyond the intended target area. In contrast to prior limitations, our method enhances consistency between edited regions and text prompts, while preserving original skin details by understanding the contextual causality between generation and source image information.

Analysis of the Score in CA^2. We visualize the score in Eq.(3) to explore how our CA^2 dynamically prevents the conflict between visual and textual condition. The lighter region of a score map corresponds to higher values, which indicates less injection of image features in those regions. In Fig. 5, our model initially focuses more on image prompts to maintain skin tones. As inference progresses, it relies less on the original image and generates content based on text. The score map shows lower values at edges and areas with minimal editing, indicating these regions depend more on the original image. This aligns with the score's role in CA^2, allowing spatial control over the sensitivity to image prompts.

Analysis of the frequency guidance of STFG. To study the capacity of STFG in enhancing the low-frequency similarity during sampling, we calculate the cumulative amplitude difference in the Fourier domain between the sampled

Method	Objective Metrics					User study (Ours vs.)		
	FID/L-FID (↓)	LPIPS (↓)	HPSv2(↑)	ID (↑)	MPS (↑)	FF (↑)	TAC (↑)	HP (↑)
SD Inpainting	**3.11/1.61**	0.175	0.248	0.63	1.03	86.05%	79.32%	77.88%
BrushNet	5.45/2.30	0.285	0.254	0.59	1.34	86.05%	83.17%	82.69%
IntructPix2Pix	8.36/5.36	0.160	0.263	0.67	1.03	87.98%	83.65%	85.09%
DiffusionClip (310s)	8.19/5.68	0.301	0.257	**0.73**	1.13	93.56%	68.29%	92.31 %
Asyrp (408s)	8.11/6.32	0.260	0.240	0.62	1.80	86.05%	63.29%	84.28 %
StyleClip (40s)	6.38/4.83	0.249	0.263	0.63	1.09	93.68%	83.17%	68.38%
Ours	4.81/1.99	**0.085**	**0.264**	0.72	/	/	/	/

Table 1: Quantitative comparisons between the state-of-the-art methods and ours. "Ours vs." indicates the proportion of users who prefer our proposed method over a comparative approach. The proportion in user study exceeding 50% indicates that our method outperforms the counterpart. MPS exceeding 1.00 indicates that our method outperforms the counterpart. Number in "( )" is the time required for single attribute fine-tuning of facial editing methods.

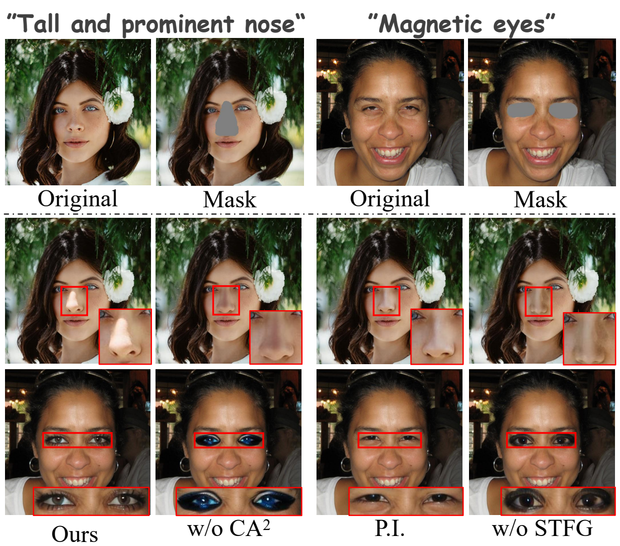

Figure 6: Ablation study of our modules. "Parallel Injection (P.I.)" removes the score in Eq. (3), "w/o CA^2" removes the CA^2 but preserves textual condition.

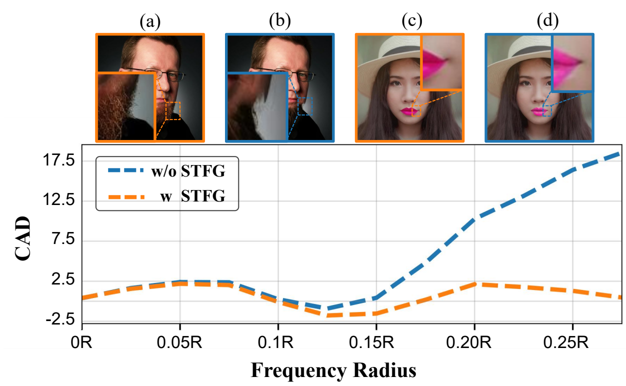

Figure 7: Cumulative amplitude difference (CAD) in the Fourier domain between the sampled and the original images is calculated within the mask region, specific to the FFT-shifted image with a radius representing by the x-axis. (a) and (c) are the sampled images with ('w') STFG. (b) and (d) are those without ('w/o') STFG.

and the original images, varying as the frequency radius within the mask region. The amplitude difference specific to our STFG (the orange line) fluctuates around zero, indicating that STFG can effectively promote low-frequency similarity between the sampled and original images, which is helpful for the skin detail preservation. The images shown above the graph demonstrate that the artifacts in the edge region have been effectively eliminated by STFG.

Ablation Study

We demonstrate our module's effectiveness through qualitative and quantitative metrics (appendix). As seen in Fig. 6, removing CA^2 results in generated content that mismatches the original skin tone. The model without the score in Eq. (3) shows significant content leakage and fails to follow the text description accurately. This indicates that the score encourages the model to prioritize textual editing. Without our STFG, noticeable artifacts appear around attribute boundaries.

Conclusion

This paper introduces a novel inpainting technique for local facial attribute editing that overcomes the long-lasting issues in current models, i.e. the hardness of following the local facial attribute description and the lack of contextual causality modeling on mask regions. We present a new data strategy and a Causality-Aware Condition Adapter to effectively incorporate original image skin details for causality modeling while preventing conflict between visual and textual condition. Moreover, a Skin Transition Frequency Guidance is introduced to improve the coherence of generated content around the boundaries. Extensive experiments show the superior performance of our method over current SOTA ones.

Acknowledgments

The work was supported by the National Natural Science Foundation of China under grants no. 62276170, 82261138629, the Guangdong Basic and Applied Basic Research Foundation under grants no. 2023A1515011549, 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20220531101412030, and Guangdong Provincial Key Laboratory under grant no. 2023B1212060076.

References

Andonian, A.; Osmany, S.; Cui, A.; Park, Y.; Jahanian, A.; Torralba, A.; and Bau, D. 2021. Paint by word. *arXiv preprint arXiv:2103.10951*.

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

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Dai, D.; Li, Y.; Liu, Y.; Jia, M.; YuanHui, Z.; and Wang, G. 2024. 15M Multimodal Facial Image-Text Dataset. *arXiv preprint arXiv:2407.08515*.

Ding, Z.; Zhang, X.; Xia, Z.; Jebe, L.; Tu, Z.; and Zhang, X. 2023. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12736–12746.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022a. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022b. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

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

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jeong, J.; Kim, J.; Choi, Y.; Lee, G.; and Uh, Y. 2024. Visual Style Prompting with Swapping Self-Attention. *arXiv preprint arXiv:2402.12974*.

Jia, H.; Li, Y.; Cui, H.; Xu, D.; Yang, C.; Wang, Y.; and Yu, T. 2023. DisControlFace: Disentangled Control for Personalized Facial Image Editing. *arXiv preprint arXiv:2312.06193*.

Jiang, Y.; Huang, Z.; Pan, X.; Loy, C. C.; and Liu, Z. 2021. Talk-to-Edit: Fine-Grained Facial Editing via Dialog. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*.

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

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

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2426–2435.

Kwon, M.; Jeong, J.; and Uh, Y. 2022. Diffusion models already have a semantic latent space. *2210.10960*.

Kwon, M.; Jeong, J.; and Uh, Y. 2023. Diffusion Models Already Have A Semantic Latent Space. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Levin, E.; and Fried, O. 2023. Differential diffusion: Giving each pixel its strength. *arXiv preprint arXiv:2306.00950*.

Li, Y.; Zhang, Y.; Wang, C.; Zhong, Z.; Chen, Y.; Chu, R.; Liu, S.; and Jia, J. 2023. Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models. *arXiv:2403.18814*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Manukyan, H.; Sargsyan, A.; Atanyan, B.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. HD-Painter: High-Resolution and Prompt-Faithful Text-Guided Image Inpainting with Diffusion Models. *arXiv preprint arXiv:2312.14091*.

Mao, J.; Wang, X.; and Aizawa, K. 2023. Guided image synthesis via initial image editing in diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5321–5329.

Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. LaDI-VTON: latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8580–8589.

Nitzan, Y.; Aberman, K.; He, Q.; Liba, O.; Yarom, M.; Gandelsman, Y.; Mosseri, I.; Pritch, Y.; and Cohen-Or, D. 2022. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6): 1–10.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2085–2094.

Pernuš, M.; Štruc, V.; and Dobrišek, S. 2023. Maskfacegan: High resolution face editing with masked gan latent code optimization. *IEEE Transactions on Image Processing*.

Qi, T.; Fang, S.; Wu, Y.; Xie, H.; Liu, J.; Chen, L.; He, Q.; and Zhang, Y. 2024. DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8693–8702.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Shen, Y.; Yang, C.; Tang, X.; and Zhou, B. 2020. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4): 2004–2018.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; and Chen, A. 2024. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*.

Wang, T.; Zhang, Y.; Fan, Y.; Wang, J.; and Chen, Q. 2022. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11379–11388.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

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

Xu, J.; Motamed, S.; Vaddamanu, P.; Wu, C. H.; Haene, C.; Bazin, J.-C.; and De la Torre, F. 2024. Personalized face inpainting with diffusion models by parallel visual attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5432–5442.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Yang, H.; Chai, L.; Wen, Q.; Zhao, S.; Sun, Z.; and He, S. 2021. Discovering interpretable latent space directions of gans beyond binary attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12177–12185.

Yang, S.; Chen, X.; and Liao, J. 2023. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3190–3199.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

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

Zhang, S.; Wang, B.; Wu, J.; Li, Y.; Gao, T.; Zhang, D.; and Wang, Z. 2024. Learning Multi-Dimensional Human Preference for Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8018–8027.

Zhong, S.; Huang, Z.; Wen, W.; Qin, J.; and Lin, L. 2023. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, 567–578.