scientific reports

Check for updates

OPEN

Irrelevant region preserving for counterfactual image manipulation

Yinuo Peng[1], Yuxuan Wang[1], Chenyue Wang[1], Xin Wang[2] & Jiatian Pi[1]✉

Image editing is one of the most significant and potential research topics in the field of multimodal learning. Several existing methods based on Contrastive-Language-Image-Pretraining (CLIP) have achieved high-resolution image editing recently, but the challenging problem of complex editing and attribute disentanglement has not been solved yet. In this paper, we propose an image editing method combining the powerful capability of complex editing with the accurate protection of the irrelevant attributes, simultaneously addressing above two challenging issues. To gain a more comprehensive semantic representation, we design a simple but effective structure with the cross-attention mechanism, allowing better fusion between text and image feature. In addition, a mask-controlled method is applied to keep the semantics of irrelevant regions unchanged after editing. We conduct extensive experiments and analysis to evaluate the generative capability of our method. The results demonstrate that our design successfully achieves semantic representation and accurate editing, and outperforms the compared methods in image quality.

Artificial intelligence (AI) has been experiencing rapid growth and evolution in recent years. As one of the classical networks in this field, regarding the mutual competition between two networks as a game, Generative Adversarial Networks (GANs)[1] consist of a generator network that produces synthetic data and a discriminator network distinguishes between the synthetic data and the true data without additional manual annotation costs, which has achieved tremendous success worldwide. Numbers of state-of-the-art methods based on GANs have been developed to improve the effectiveness and quality of the generative network. Deep Convolutional Generative Adversarial Networks (DCGANs)[2] utilized the feature extraction capability of convolutional neural network to improve the learning effectiveness of the generative network.

In order to generate higher-quality images step by step, Progressive GAN (PGGAN)[3] introduced a progressive structure, realizing the transition from low to high resolution. When it comes to StyleGAN[4], by style mixing, Karras et al. successfully found a disentangled latent space to control the process of image synthesis, making image manipulation in latent space possible.

Like other related methods of image editing, one of the major challenges of StyleGAN is the large amount of annotated data pairs and significant manual effort required for training, which comes at a high cost. Researchers have been exploring the possibility of incorporating zero-shot capabilities to reduce this burden. Therefore, for the first time, StyleCLIP[5] combines the powerful generative capability of StyleGAN and the joint representation of vision-language in CLIP[6], successfully realizing image generation with a given text prompt, and alleviating the need for manual effort. Then, more researchers have devoted much efforts in text-guided image manipulation such as StyleGAN-NADA[7], CLIPstyler[8],and TediGAN[9]. These methods achieved effective results based on the improvement of StyleCLIP, making image editing more diverse. However, they still struggle with performing complex edits. To address this, Yu et al. proposed CF-CLIP[10], a CLIP-based text-guided image manipulation network, to allow for accurate counterfactual editing. Through the innovative design of the Text Embedding Mapping (TEM) module and the CLIP-NCE, CF-CLIP significantly enhanced the ability to preserve semantic information, achieving more realistic and complex edits driven by target texts with various counterfactual concepts. However, like other state-of-the-art methods, CF-CLIP suffers from a common limitation: it often leads to undesirable edits of irrelevant regions or excessive changes. Fig. 1 illustrates an example of such a scenario. This issue is primarily due to insufficient disentanglement capabilities. The reasons for this limitation can be identified in two key points: first, these methods generally assume that there is perfect disentanglement between attributes. They perform image editing by directly manipulating the latent code and using the edited latent code to generate the image through the StyleGAN generator, without applying more precise constraints during the process. Second, the Text Embedding Mapping (TEM) module in CF-CLIP merely replicates the text embedding to match the dimensions of the image code and processes it with a simple linear layer, omitting the participation of image information. As a result, the obtained embedding lacks a strong correlation with the

[1]Chongqing Normal University, National Center of Applied Mathematics, Chongqing 401331, China. [2]Chongqing Changan Automobile Co., Ltd., Chongqing 400023, China. ✉email: pijiatian@cqnu.edu.cn

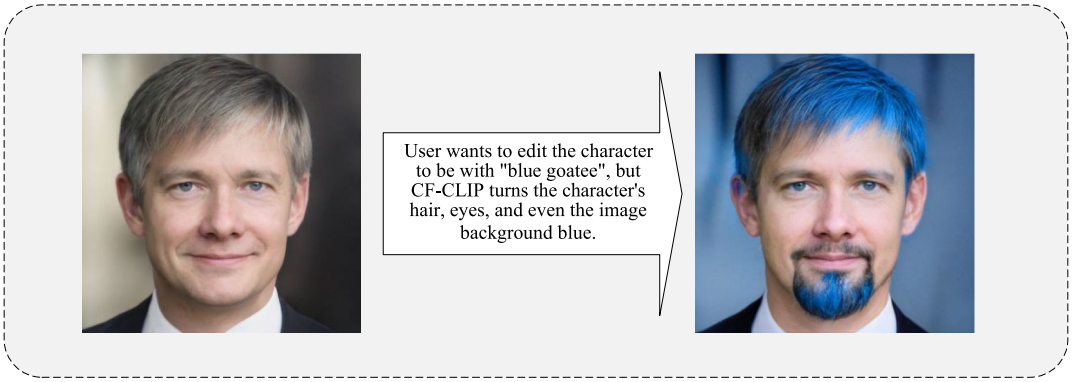

Fig. 1. Results generated by CF-CLIP[10] with the text prompt *blue goatee*. The method understands *blue*, but the attribute it acts on is inaccurate.

image, and the highlighted semantics cannot be effectively aligned with the corresponding regions of the image, leading to insufficient feature fusion and inaccurate editing of regions.

MCA-CLIP (i.e., a **CLIP**-based image editing network that combines the capability of **M**ask-Control and **C**ross-**A**ttention mechanism), the new structure proposed in this paper solves the problems above. It incorporates the functionalities of two key modules: the Feature Fusion module and the Precision Region Editing module. To achieve a more comprehensive semantic representation, we adopt cross-attention mechanisms that enable image information to participate in the text mapping network. This makes the fusion of text and images more explicit. More importantly, considering the accuracy of the editing area, we apply the mask-controlled method from CoralStyleCLIP[11] into our design to generate masks corresponding to textual meanings. It enables more precise control of the generation process with masks at each layer of the StyleGAN generator, rather than relying solely on the manipulation of latent codes to produce the target image. In addition, we also incorporate the background loss by ParseNet[12] into our loss function to further constrain changes in the background. By virtue of these strategies, irrelevant regions are effectively preserved and precise editing is ultimately achieved.

In summary, our contributions are as follows:

- We propose MCA-CLIP, a CLIP-based image manipulation framework that enables more effective protection of irrelevant areas, with powerful disentanglement capabilities for accurate editing.
- We design the Feature Fusion module based on cross-attention mechanisms, aiming to achieve a better fusion of features between text and image knowledge.
- We incorporate CORAL into the Precision Region Editing module, utilizing masks to constrain the changes in irrelevant regions.
- According to extensive experiments conducted, our method undoubtedly performs better than previous methods that realized complex editing in disentanglement.

Related works

Generative adversarial networks

Generative Adversarial Networks (GANs)[1] have demonstrated significant dominance in the field of image synthesis over the past few years. Numerous approaches have been developed to enhance the effectiveness of GANs from various aspects. In order to improve the quality of synthetic images, Karras et al. proposed a training methodology for GANs, known as the Progressive Growing GANs (PGGAN)[3]. The method begins with low-resolution images and then progressively increases the resolution by adding layers to the networks, which sufficiently stabilizes the training process and allows for the reliable synthesis of high-resolution images. However, PGGAN directly generates images level by level, with features that are uncontrollable and interconnected. To this end, Karras et al. introduced StyleGAN[4], a style-based generator that utilizes style variable y to control the layers of the synthesis network. The mapping network in StyleGAN achieves further disentanglement by mapping different attributes of the face into multiple dimensions, allowing independent manipulation of each attribute. One year later, Karras et al. found that there were characteristic artifacts in images generated by StyleGAN. They identified the causes for the water droplet-like artifacts and redesigned the generator's architecture, which was named StyleGAN2[13]. They eliminated redundant operations, relocated bias terms b and noise B outside of the style block, and replaced the AdaIN operation with a demodulation process. These effectively reduced the coupling relationship between blocks and successfully eliminated the artifacts, improved image quality, and made the generator easier to invert.

Similar to numerous image manipulation works[10,14–16], our method utilizes the powerful generative capability of StyleGAN2. This allows for the synthesis of high-resolution images while flexible controlling is being conducted.

Vision-language pre-training models

Vision and language are two essential components in multimodality research. Visual-language (V-L) tasks encompass image captioning[17], visual question answering (VQA)[18], image-text retrieval, visual grounding and

text-to-image generation, among others. Several works have been proposed to address these tasks[19–24]. In this paper, we address text and image alignment issues by employing one of the state-of-the-art methods for V-L tasks, Contrastive Language-Image Pre-training (CLIP)[6]. Trained on 400 million text-image pairs, CLIP learns a joint representation of vision and language through contrastive learning. It consists of a text encoder and an image encoder, which produce 512-dimensional embeddings for text and images, respectively. The similarity between these embeddings is calculated, and the text with the similarity score is selected as the prediction by CLIP. This architecture and the training approach have resulted in the development of powerful representations and zero-shot image classification capabilities.

Text-guided image manipulation

As StyleGANs[4,13,25] are always used as the backbone of most image synthesis methods[5,7,9,26,27], manipulating images in latent space $W+$ has become increasingly popular in recent years. Similarly, the powerful semantic representation capability of CLIP has sparked renewed interest in text-guided editing across a variety of attributes.

StyleCLIP[5] was the first to integrate the generative capability of StyleGAN with the robust vision-language joint representations learned by CLIP, where two techniques in this work are aimed at minimizing the CLIP-space distance between the target text and the edited image. The third technique used the CLIP text encoder to obtain a vector corresponding to the text prompt and mapped the vector into a manipulation direction for editing. However, these methods are limited to the domain of training data. Therefore, Rinon Gal et al. proposed StyleGAN-NADA[7] that enables out-of-domain generation. The directional CLIP loss in[7] aimed to keep the direction from source image to edited image the same as that from the source to target textual prompt in CLIP-space so that the operator can fine-tune the trainable generator in each iteration to produce expected images. This method successfully shifted the domain of the pre-trained model towards a new domain, accomplishing out-of-domain generation, but the massive costs of manual efforts cannot be dismissed. Thus, CF-CLIP[10] was proposed.

In their work[10], Yu et al. developed the CLIP-NCE loss, constructing positive and negative samples of the noise contrastive estimation loss using edited image, source image, target text and the corresponding latent encoding of the text prompt to compute infoNCE loss[28]. This approach thoroughly explored the CLIP semantic knowledge comprehensively. Yu et al. also designed a text embedding mapping (TEM) module to explicitly utilize the semantic knowledge of CLIP embeddings. By leveraging the zero-shot transfer and semantic representation capability of CLIP and CLIP-NCE, CF-CLIP managed to perform complex editing without additional manual annotation. The persisting challenge, however, is to prevent unintended semantic changes in irrelevant areas, which serves as the primary motivation for our work.

Method

In this section, we will introduce the major framework of our approach in this paper. As previously stated, we design a Feature Fusion module for highlighting target semantics and a mask-controlled Precision Region Editing module for preserving irrelevant regions, which are simple but effective. Fig. 2 shows an overview of our method. The text embeddings e^t, encoded by CLIP, along with the image latent code from e4e[29], are accepted by Feature Fusion (FF) module. Then the new embedding t_w consisting of semantics and structure of image, is obtained. Subsequently, the latent mapper M manipulates t_w to produce a residual Δw, which contains semantic information related to the target editing. Ultimately, the residual Δw and the original image latent code w are fixed using the Precision Region Editing (PRE) module. This final step occurs under the guidance of masks derived from the feature generation process of the original image in the StyleGAN2 generator.

Feature fusion module

The text embedding mapping (TEM) module described in[10] disregards the important role of image information in the fusion of the two modalities in the process. To enhance the integration of CLIP embeddings with image features and to emphasize semantic information for target editing, we design an 18-layer Feature Fusion module. Drawing inspiration from the Transformer architecture[19], each layer of the module comprises a multi-head cross-attention (CA) block coupled with a straightforward feed-forward computation block. This design enables the concurrent processing of information from both the image and text domains. Upon entry of the image latent code into the Feature Fusion module, it is split into multiple components w_1, w_2,..., w_n, where n corresponds to the number of StyleGAN layers. It is worth noting that in order to accentuate the semantic information that is guided by the structure of the image, in the initial layer of the Feature Fusion module, we assign the text embedding $e^t \in \mathbb{R}^{1\times 512}$ as both the Key (K) and Value (V), while the first segment of the image latent code w as the Query (Q). Subsequently, the output t of the previous layer is employed as K and V for the subsequent layer. In a parallel manner, the image latent code w_i is used as Q for the i-th layer. After these series of manipulations of the cross-attention module, each layer produces an output $I_{out} \in \mathbb{R}^{1\times 512}$. The second dimension of these outputs is concatenated together to form the output of FF, which is the result of the feature fusion between the original image feature and the text embedding. This outcome is more concentrated on the editing attributes corresponding to the input description, a result of the computations performed by the cross-attention mechanism. The whole structure of FF is shown in Fig. 3(a)

and the procedure for this module can be formulated as:

$$\begin{aligned} &w = (w_1, w_2, ..., w_n), \\ &Q = w_i, K = temp^{(i-1)}, V = temp^{(i-1)}, i = 1, 2, \ldots, n, \\ &temp^i = Attention(Q, K, V) + Q, \end{aligned} \tag{1}$$

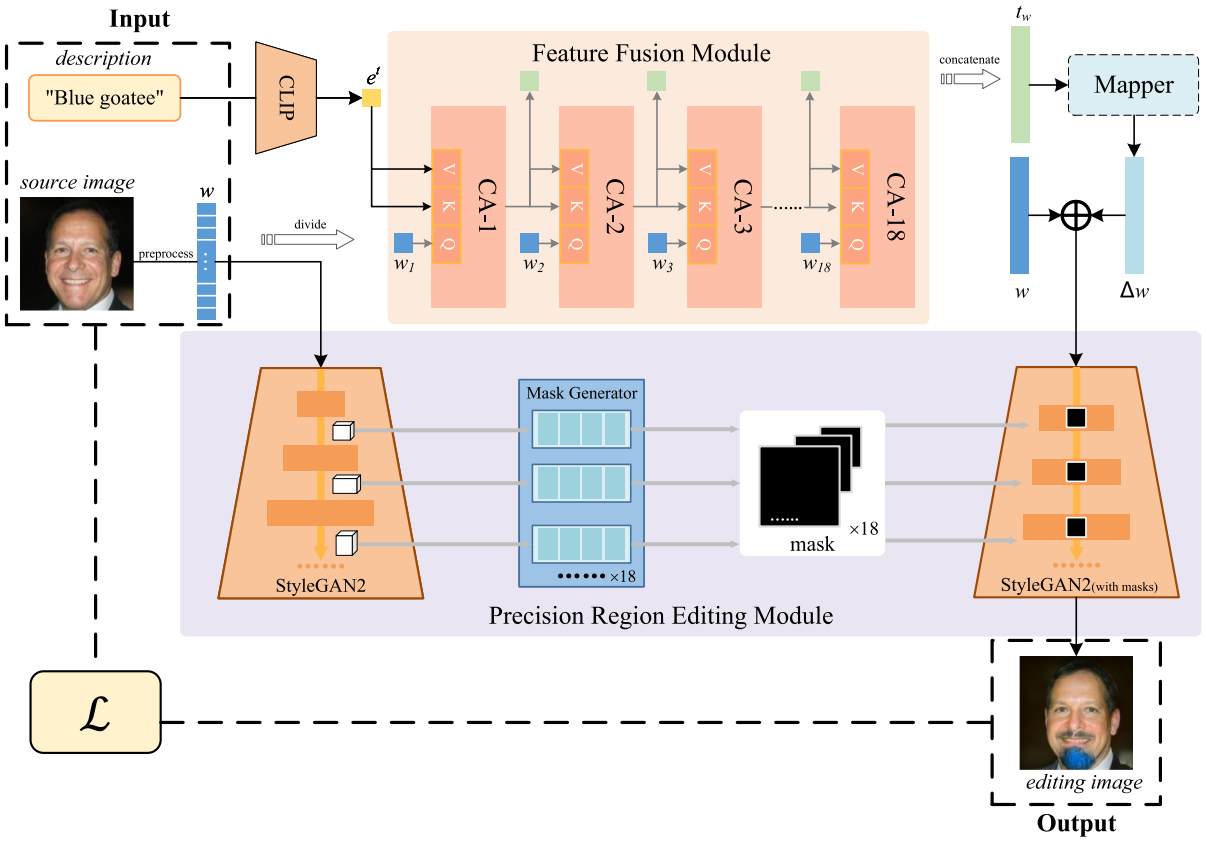

Fig. 2. The main structure of MCA-CLIP. Given the latent code w of the source image and the description, the Feature Fusion module computes t_w with strong semantics by 18 cross-attention blocks. The Precision Region Editing module learns to produce the needed masks for generator and finally generates the editing image with Δw, w and masks involved.

$$t_w = concat(temp^1, temp^2, \ldots, temp^n) \tag{2}$$

where $temp^i$ is the output of the i-th layer. When $i = 1$, which means at the first layer, e^t is used as K and V.

Precision region editing module

As explained above, the Feature Fusion module highlights the semantic information that is intended to be conveyed in the description from users. This focus makes the edits performed by residual networks more accurately to reflect the intended semantics targets. To further protect the regions irrelevant to target editing, and inspired by the generation process of StyleGAN2[13] and the highly effective method from CoralStyleCLIP[11], we draw upon CORAL, a co-optimized region and layer selection mechanism to design a structure named Precision Region Editing module.

In this module, we employ a mask that functions as a binary matrix, with values of 0 s and 1 s, at each layer of the generator module, serving to constrain the regions that are subject to editing operations. As noted in[11], an 18-layer convolutional attention network is designed to generate masks. It takes the feature $f^{(l)}$, extracted from the l-th layer of the StyleGAN2 generator module, as input and processes it to produce accurate mask predictions. Each layer consists of two convolutional layers, and the specific structure and process are illustrated in Fig. 3(b). Thus, with the mask at each layer involved, as depicted in Fig. 3(c), the generative process pays more attention to editing only the unmasked regions, thereby keeping the semantics of irrelevant attributes unchanged. The mask-controlled generation process can be formulated as:

$$masks = (m_1, m_2, \ldots, m_{18}) = G_M(f^{(1)}, f^{(2)}, \ldots, f^{(l)}) \tag{3}$$

$$f^{*(l)} = m_l \odot \widehat{f^{*(l)}} + (1 - m_l) \odot \widehat{f^{(l)}} \tag{4}$$

where $G_M(\cdot)$ represents the convolutional network responsible for generating the masks. $\widehat{f^{*(l)}}$ and $\widehat{f^{(l)}}$ refer to the original and edited features extracted by convolutional layers for $w^{(l)}$, $w'^{(l)}$ and $f^{*(l)}$, respectively, and $f^{*(l)}$ is the output of each blend layer generated by the variables mentioned above, which is illustrated in Fig. 3(c).

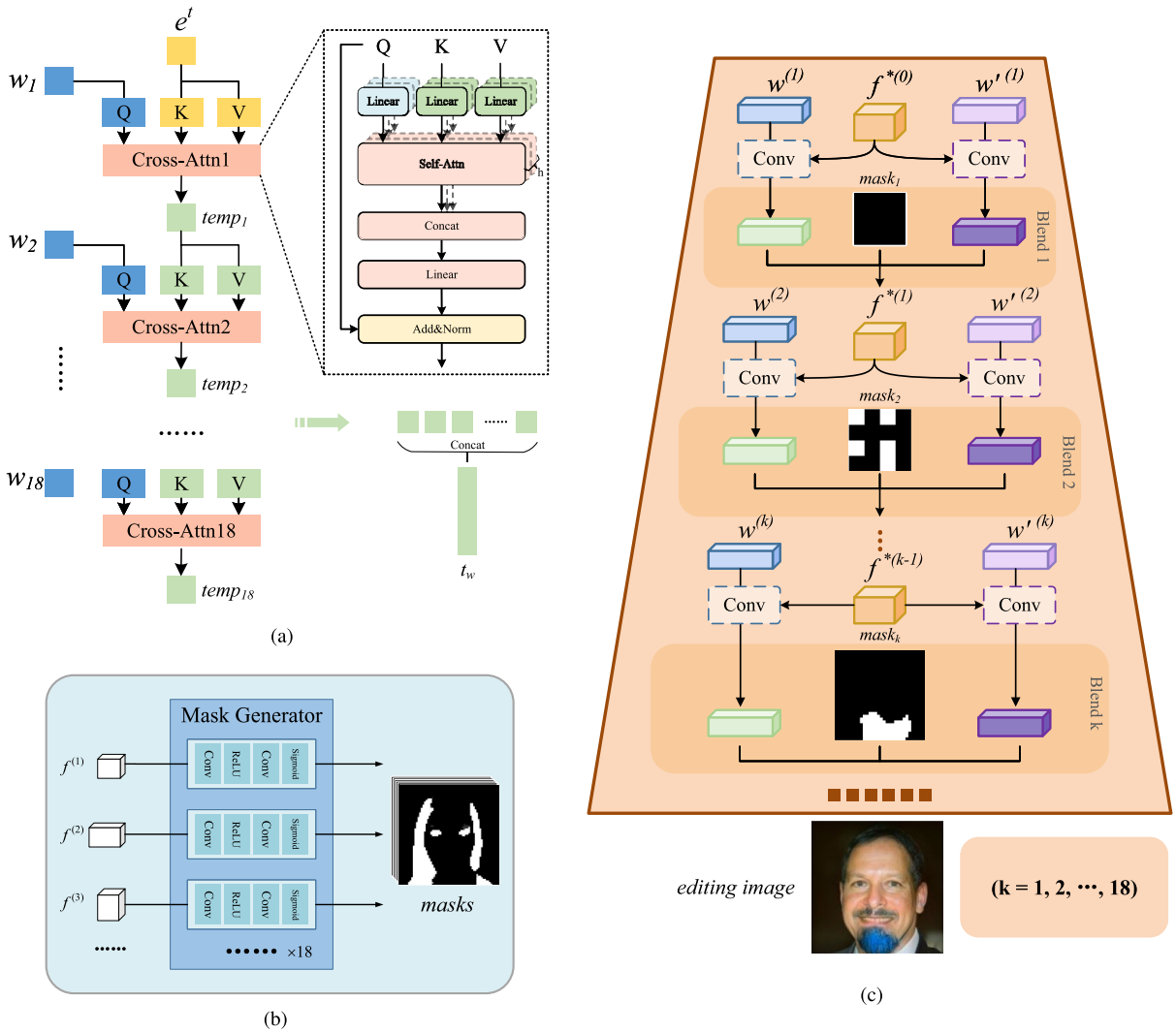

Fig. 3. (**a**) The structure of the Feature Fusion module. Every layer's output is an embedding $temp \in \mathbb{R}^{1\times512}$, and they are concatenated together to generate $t_w \in \mathbb{R}^{1\times18\times512}$. (**b**) Mask generation network. At each generator layer l, the feature $f^{(l)}$ is fed into a convolutional-attention module to predict a probability mask $m_l \in [0,1]^{H_l\times W_l}$. (c) The framework of Masked Generator. CORAL[11] performs multi-layer blending with custom edit regions per layer.

Loss functions

In this section, we will describe the training objectives and associated loss functions employed in our proposed method. Initially, we are provided a latent code w representing an original image and a text description t. The objective of our method is to generate an edited image based on the original image, aligning with the given description, while ensuring that the irrelevant regions remain unchanged.

To achieve this, and to keep the identity information of the person in the image the same after editing, we apply the identity loss[30] in the training process, defined as:

$$\mathcal{L}_{id} = 1 - \cos < R(x), R(\hat{x}) > \tag{5}$$

where x denotes the original image and $\hat{x}$ represents the edited image. $R(\cdot)$ refers to the pretrained ArcFace[31] network, which extracts identity features from the image and $Cosine(\cdot,\cdot)$ computes the similarity between two features vectors.

The residual Δw carries the editing information acting on the latent code w of the original image. The larger Δw is, the more changes will occur. Therefore, in latent space $\mathcal{W}+$, we adopt the L_2 distance to control the edits with smaller l_2 norms, which is defined as:

$$\mathcal{L}_{l_2} = ||\Delta w||_2. \tag{6}$$

Following the loss in CF-CLIP[10], we also use the proposed CLIP-NCE loss to guarantee semantic consistency between the edited image and the given description. It is formulated as:

$$\begin{aligned}\mathcal{L}_{NCE} = &- \log \frac{e^{(Q \cdot K_T^+ / \tau)}}{e^{(Q \cdot K_T^+ / \tau)} + \sum_{K^-} e^{(Q \cdot K^- / \tau)}} \\ &- \log \frac{e^{(Q \cdot K_I^+ / \tau)}}{e^{(Q \cdot K_I^+ / \tau)} + \sum_{K^-} e^{(Q \cdot K^- / \tau)}}.\end{aligned} \tag{7}$$

In this formulation, similar to a common contrastive loss, Q represents the query, K^- represents the negative samples, and the positive samples are composed of two components: K_T^+ and K_I^+. These values are defined as followed:

$$\begin{aligned}Q &= E_I(I_{aug}) - E_I(I_{src}), \\ K^- &= E_T(t_{src}) - E_I(I_{src}), \\ K_T^+ &= E_T(t_{tgt}) - E_T(t_{src}), \\ K_I^+ &= E_T(t_{tgt}) - E_I(I_{src}).\end{aligned}$$

I_{aug} represents the augmented images, and I_{src} represents the original image. Meanwhile, t_{src} represents the text prompt after prompt engineering[6], and t_{tgt} is the given description, which is our editing target. $E_I(\cdot)$ and $E_T(\cdot)$ are encoders for image and text respectively.

In addition to optimizing at the semantic representation level, we also take the scale of editing regions into account. In this case, we adopt the **Minimal Edit-area Constraint** and **Smoothness loss** introduced in[11] to regulate the sizes of masked and unmasked regions:

$$\mathcal{L}_{area} = \sum_l n_l (\sum_{i,j} m_{i,j}^{(l)}), \tag{8}$$

$$\mathcal{L}_{tv} = \sum_{i,j,l} ||m_{i,j}^{(l)} - m_{i+1,j}^{(l)}||_2^2 + \sum_{i,j,l} ||m_{i,j}^{(l)} - m_{i,j+1}^{(l)}||_2^2 \tag{9}$$

where m refers to the matrix of masks, and n_l is a normalizing constant related to the growing feature dimensions during the StyleGAN2 generator module.

Finally, to further ensure the preservation of irrelevant regions, especially the background, we adopt the background loss to keep the background unchanged before and after editing. It is expressed as:

$$\mathcal{L}_{bg} = MSE(x \circ m_{bg}, \hat{x} \circ m_{bg}) \tag{10}$$

where m_{bg} is the background of the image extracted by ParseNet[12].

Therefore, the overall loss function of our method is expressed as a weighted combination of the losses above.

$$\begin{aligned}\mathcal{L} = &\lambda_{id}\mathcal{L}_{id} + \lambda_{l_2}\mathcal{L}_{l_2} + \lambda_{NCE}\mathcal{L}_{NCE} \\ &+ \lambda_{area}\mathcal{L}_{area} + \lambda_{tv}\mathcal{L}_{tv} + \lambda_{bg}\mathcal{L}_{bg}.\end{aligned} \tag{11}$$

Experiment

This section first evaluates MCA-CLIP in the context of faces, demonstrating a more accurate and higher-quality editing compared to other methods. Then we conduct the ablation study to analyze the irreplaceable contributions of the modules we design.

Settings

The experiments are conducted on CelebA-HQ[3] and FFHQ[4], two commonly used high-quality human face image datasets. CelebA-HQ is a high-quality version of the CelebA[32] and consists of 30,000 human face images with a resolution of 1024 × 1024, and FFHQ includes 70,000 high-resolution face images. For input of the network, we preprocessed the datasets with e4e[29] model, encoding images into inverted latent codes in $W+$ space.

In this paper, we choose StyleCLIP[5] and CF-CLIP[10] as compared methods. The former is the first method that combines the powerful generative capability of StyleGAN with the joint representation of CLIP, leading to the development of CLIP downstream tasks. With the proposal of CLIP-NCE, the latter demonstrates the unique capability of counterfactual editing, which outperforms many recent works in text-guided image manipulation.

The values of λ in the loss function are set at 0.2, 0.8, 0.3, 0.00002, 0.00003 and 0.2 for $\mathcal{L}_{id}$, $\mathcal{L}_{l_2}$, $\mathcal{L}_{NCE}$, $\mathcal{L}_{area}$, $\mathcal{L}_{tv}$ and $\mathcal{L}_{bg}$, respectively. Our experimental environment consists of a 12th Gen Intel Core i9-12900HX CPU, an NVIDIA GeForce RTX 3080 Ti GPU (16GB), with PyTorch 2.1.0, Python 3.9, CUDA 11.8, and cuDNN 8.7.

Qualitative results

In this subsection, we compare our method with chosen state-of-the-art editing methods from the perspective of image quality, taking into account the semantic consistency of the edited image and text, and the disentanglement between attributes.

Fig. 4 shows the results over CelebA-HQ[3]. We mainly select 7 phrases as description prompts, three of which are complex texts with counterfactual concepts and four are commonly used for applications. As shown in Fig. 4, StyleCLIP[5] can perform the simple edits but not the complex edits. For instance, although it attempts to alter the specified region in accordance with the description, it only manages to alter it within the bounds of conventional cognition and fails to perform the semantics of the counterfactual concept. And CF-CLIP[10], while it enables complex edits, does not consider the protection of irrelevant regions at all. For example, the clothes and background in the original image have been completely changed, which reveals the poor disentangled capability of this work. By contrast, MCA-CLIP solves the problems, enabling counterfactual editing while maintaining the semantics of irrelevant regions for precise editing.

The bottom row of Fig. 4 shows a more detailed background comparison between the input image and images generated by each method. The background of the image generated by CF-CLIP[10] changes dramatically from the origin in both simple and complex descriptions. For example, in the description *Blue Goatee*, there is a big change in its color, as well as many extraneous patterns. We only employ StyleCLIP[5] as a comparison in simple text groups because it fails to perform complex edits. StyleCLIP generates images with minimal backdrop modification, which is similar to our method. Furthermore, our approach demonstrates a better-disentangled capacity to keep the background unchanged in both simple and complex text editing.

In addition to the results shown in Fig. 4, we further evaluate our method on facial images with different angles. Shown in Fig. 5, the results demonstrate that even under variations in pose, the proposed method can successfully alter the features indicated by text prompts while preserving other facial attributes and general identity.

Quantitative results

In this section, we evaluate the generative ability of these methods using several metrics as follows.

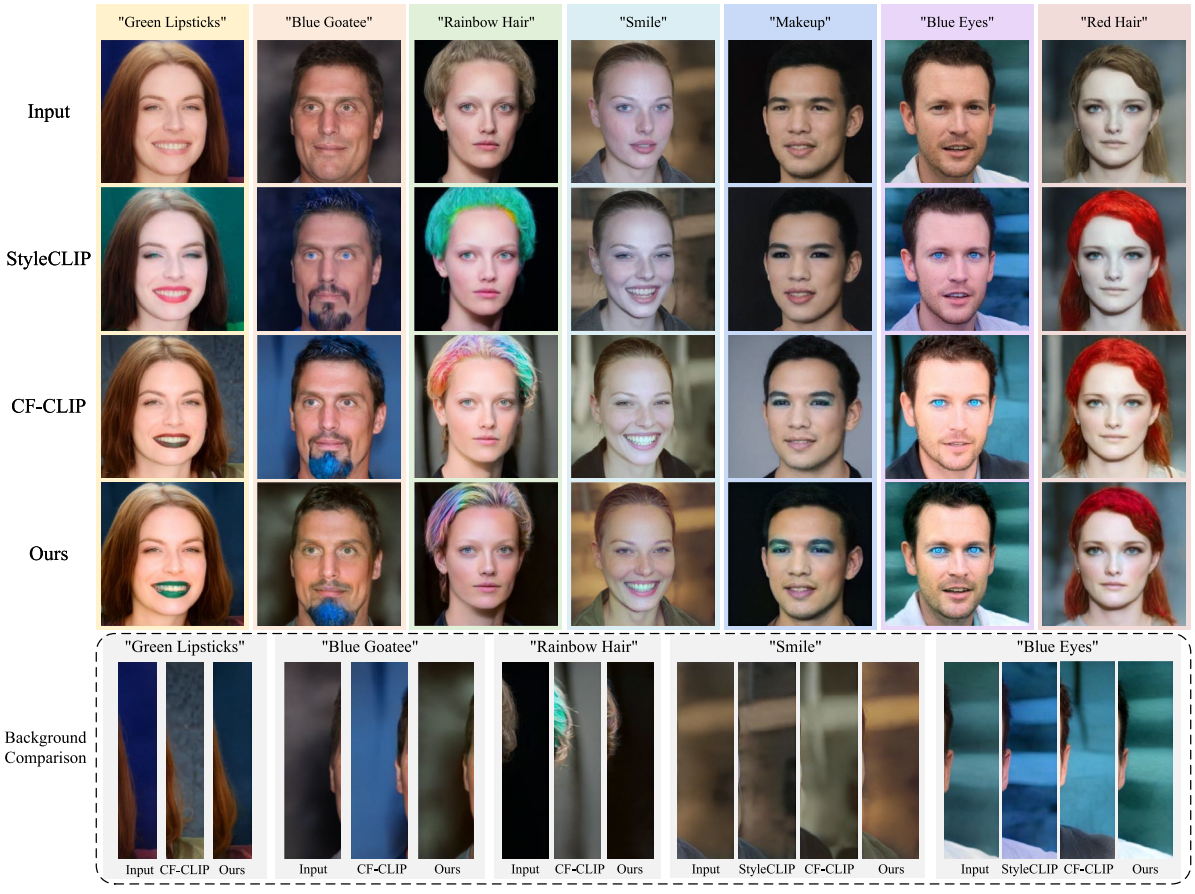

Fig. 4. The image results of the comparison experiment. The first row shows the chosen text prompt for manipulation and the second row shows the original image for input of models. The three rows following demonstrate the edited results of two compared methods (StyleCLIP[5], CF-CLIP[10]) and one proposes in this paper. At the bottom is a detailed display of the image background.

Fig. 5. Results on facial images with different angles, including frontal views, profile views, and faces with occlusions.

	LPIPS↓		SSIM↑		BLIP-Pmatch↑	
	Ours	**CF-CLIP**	**Ours**	**CF-CLIP**	**Ours**	**CF-CLIP**
GreenLipsticks	**0.3164**	0.348	**0.7557**	0.7048	**0.8457**	0.8281
RainbowHair	0.3978	**0.3769**	**0.677**	0.6355	**0.8045**	0.5972
BlueGoatee	**0.3676**	0.4028	**0.7304**	0.6434	**0.8116**	0.7484
BlueEyes	**0.3378**	0.3911	**0.6992**	0.6161	**0.8865**	0.8715
Makeup	**0.3025**	0.3468	**0.7588**	0.6624	0.1147	**0.1221**
Smile	**0.3522**	0.3688	**0.7343**	0.7188	0.5427	**0.6374**
RedHair	**0.3293**	0.3798	**0.7193**	0.6779	0.7685	**0.7921**

Table 1. The Quantitative Results of the Comparison Experiment. Lower Values of LPIPS[33] Indicate Higher Similarity Between Two Images, and the Values Closer to 1 of SSIM[34] Indicate Higher Similarity Instead. Higher BLIP-Pmatch[35] Indicates Better Matching of Image and Text, Which Means that the Image Manipulation Method is More Capable of Semantic Representation. These Results are Calculated as an Average of 200 Edited Images for Each Text Prompt, Respectively.

- Learned Perceptual Image Patch Similarity (LPIPS)[33]: A metric for measuring perceptual similarity between two images. It is learned through deep learning methods to obtain a better simulation of human visual perception.
- Structure Similarity Index Measure (SSIM)[34]: A metric used to measure the similarity between two images, which takes into account not only brightness and contrast, but also structural information.
- Bootstrapping Language-Image Pre-training (BLIP)[35]: A multimodal framework proposed by Salesforce in 2022, which transfers flexibly to both vision-language understanding and generation tasks. We utilize the Image-Text Matching Loss (ITM) to compute the matching possibility between edited image and the target text.

These metrics evaluate the quality of image editing methods from three dimensions: the naturalness of the edited image, the consistency of irrelevant regions before and after editing, and the degree of semantic matching between the edited image and the text. These evaluations are conducted under the condition that the corresponding semantic edits can be successfully achieved. The quantitative results are shown in Table 1. As shown in the table, we can find that the LPIPS and SSIM of MCA-CLIP outperform the compared method, demonstrating the higher visual quality of generated results. Through a joint analysis of BLIP-PMatch numerical and visualization effects, our method and CF-CLIP[10] both match the semantics of the text prompts, successfully realizing target editing with counterfactual concepts.

User study

To compare the generative quality of the proposed method and the chosen methods from a more realistic perspective, we perform subjective user studies including questions about edit accuracy, visual realism and semantic consistency. For each evaluation perspective, we choose an equal number of simple and complex phrases with counterfactual concepts for the questionnaire, and randomly select one of the images generated by each method to ask the user about the quality of the image. The order of images generated by each method is shuffled when presenting to each participant. In each question, we ask participants to rank the images, with the

Methods	Visual realism↑		Edit accuracy↑		Semantic consistency↑		Score↑
	Simple	Complex	Simple	Complex	Simple	Complex	
StyleCLIP	1.84	1.40	**2.14**	1.365	**2.65**	1.34	10.735
CF-CLIP	2.05	1.93	1.975	2.10	2.13	2.305	12.49
MCA-CLIP (Ours)	**2.115**	**2.68**	1.89	**2.54**	1.23	**2.365**	**12.82**

Table 2. The User Study Results. The Visual Realism Refers to Whether the Generated Images Look Real and Natural, Edit Accuracy Denotes the Preservation of Editing-Irrelevant Regions by the Methods during the Process, and Semantic Consistency Represents whether the Generated Images Represent the Semantics of the Target Text Prompt, Respectively.

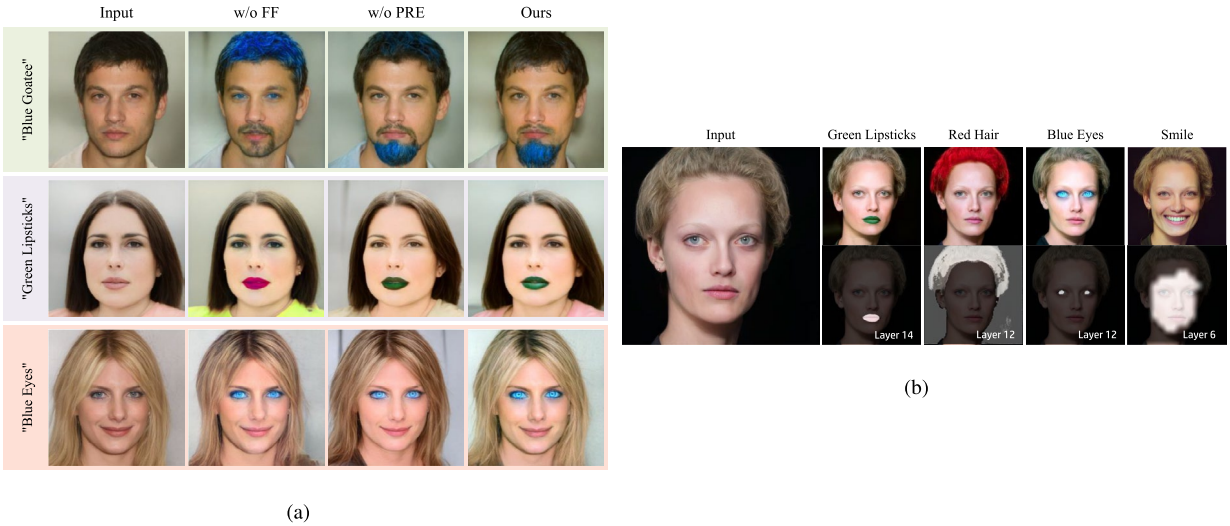

Fig. 6. The results of the ablation study. (**a**) Qualitative ablation studies of MCA-CLIP. The first column shows the original image as input. The second column shows the effectiveness of the Feature Fusion (FF) module, and the third column shows the effectiveness of the Precise Regions Editing (PRE) module. The last column shows the generated images by MCA-CLIP (Ours). (**b**) Display of the mask mechanism. The top column shows the generated images of different descriptions and the bottom one shows the corresponding mask of the target editing area.

first being the best and the third being the worst. Then we assign points to each method based on ranking results, with the first being assigned 3 points, the second 2 points and the third gaining 1 point. Finally, we calculate the average of the scores for each method. The results are shown in Table 2. It is noted that in complex editing, our method performs better in all three dimensions than the other two compared methods, while in simple editing, each method accomplishes a tiny difference in value with minimal backgrounds change. The combined score of our method is also higher than other two methods, demonstrating a better performance in image manipulation.

Ablation study

In this section, we conduct ablation studies to demonstrate the necessity and reasonability of our design, and the results are shown in Fig. 6.

The model *w/o FF* that replaces Feature Fusion (FF) module with Text Embedding Mapping (TEM)[10] loses the ability of complex editing for counterfactual description. As illustrated in the Fig. 6(a), it tends to manipulate the color of hair and eyes instead of the target goatee in the description *Blue Goatee*. The model without the Precision Region Editing (PRE) module enables the changes in the color of the goatee, but because of the lack of mask constraints, it ends up over-editing irrelevant regions, e.g. the color of the background and hair has been changed after editing. In addition, Fig. 6(b) displays the mask generated in this method. The unmasked regions match the semantics of the text, which effectively constraints the editing regions and preserves the irrelevant regions.

These ablation studies above show that the FF module and the PRE module are both indispensable for semantic representation and accurate editing in image manipulation.

Discussion

In recent years, diffusion-based models have gained significant popularity in the field of image generation and editing, due to their impressive generative capabilities. Methods such as Stable Diffusion[36] and DiffusionCLIP[37]

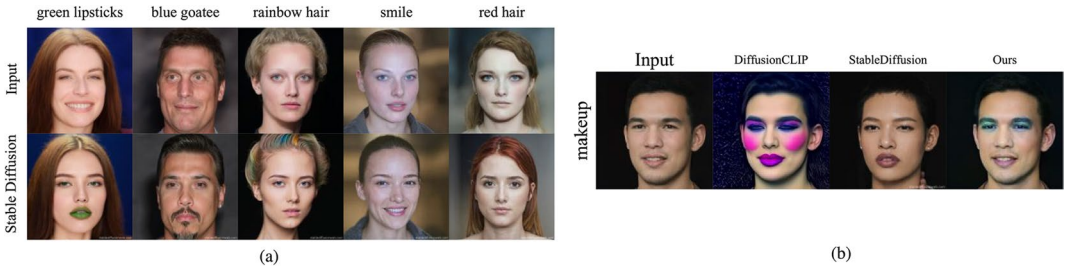

Fig. 7. (**a**) The results of Stable Diffusion[36]. It can successfully achieve some of semantics (failed to edit *blue goatee*), maintains the background well, but the appearance of the person has changed significantly. (**b**) One of results generated by DiffusionCLIP[37]. The edits made by DiffusionCLIP are somewhat exaggerated, and the style of the image has also changed.

have achieved impressive results in text-to-image generation and editing, inspiring a growing body of follow-up research.

In this subsection, we provide a brief qualitative comparison between our method and these diffusion-based approaches, focusing on two aspects: editing effectiveness and computational efficiency.

As illustrated in Fig. 7(a), although Stable Diffusion can follow most of the given instructions and preserve the background, it fails to maintain people's identity, which has changed considerably. In Fig. 7(b), when given the prompt *makeup*, the output of DiffusionCLIP shows an overly exaggerated makeup and noticeable changes in the overall style, deviating from our objective of preserving irrelevant regions.

Furthermore, several studies shows that diffusion-based methods typically demand significantly higher computational resources compared to GAN-based approaches[38,39].This is mainly because the diffusion generation process requires iterative denoising steps, which greatly increase both the memory consumption and the inference time. As a result, under the same constraints, diffusion-based methods are often restricted to operating at relatively low resolutions (e.g., 256×256), whereas GAN-based models such as StyleGAN can directly generate high-quality images at larger resolutions.

After evaluating the trade-offs between editing precision and computational efficiency, as discussed above, we selected GAN-based methods rather than diffusion-based methods as the primary baseline for our study.

Conclusion

In this paper, we introduce an image manipulation method called MCA-CLIP, which is based on the StyleGAN2 and CLIP models, to simultaneously achieve complex editing and irrelevant region protection. To enhance semantic representation, we design the Feature Fusion module to facilitate a more comprehensive fusion of image-text knowledge. Additionally, we apply the valuable mechanism from CoralStyleCLIP to design the Precision Region Editing module, ensuring that irrelevant regions are not undesirably altered. We conduct comparison studies and quantitative evaluation to demonstrate the superiority of our method and perform an ablation study to highlight the indispensable contribution of each module in the framework.

Although the method demonstrates strong performance in typical image editing tasks, there are certain challenges and limitations that need further exploration. First, the method is currently applied to facial attribute editing, and its performance may be degraded when dealing with more significant changes in image content. This is partly due to the inherent difficulty of large-scale edits, as modifying a substantial portion of the image is more challenging. Another issue arises when mask control is required outside the facial region, as the generation process becomes harder to control and the mask may not be accurate. Second, the core of the method relies on editing in the latent space of StyleGANs, which requires pre-trained StyleGANs weights for generating the final image. Different datasets necessitate different StyleGANs weights, and the pre-trained weights used for facial generation in this paper may lead to the loss of background details. As a result, the effectiveness of this approach when dealing with complex backgrounds has not yet been fully explored.

As we look forward to advancing our method and expanding its capabilities, it is essential to also consider the ethical implications of counterfactual image manipulation, as it introduces several critical concerns. The ability to edit images with increasing precision and control raises significant ethical questions, particularly about its potential misuse and the erosion of trust in digital media.

In the context of facial attribute editing, issues like identity theft, deepfakes, and the spread of misinformation are major concerns. Furthermore, such technology could be misused in legal, political, or social contexts, potentially leading to injustice or harm. Unintended consequences could also arise, such as altering public perceptions of individuals or groups in ways that are misleading, reinforcing negative stereotypes, or setting unrealistic standards of beauty and behavior in the media. Moreover, the ethical implications of editing images of real people without their consent are profound, particularly when these alterations affect an individual's identity or appearance in ways they deem unacceptable. It is imperative to establish clear guidelines to ensure responsible usage of this technology, safeguarding individual privacy and preventing harmful manipulations.

Moving forward, we plan to address the technical limitations previously mentioned by tackling more challenging image editing tasks that require broader contextual understanding and a deeper semantic fusion between images and text. Alongside these technical advancements, we will prioritize the ethical considerations surrounding our method, ensuring its responsible use and minimizing potential misuse. Moreover, an

important future direction lies in exploring the integration of hierarchical VAEs with CLIP to incorporate causal relationships into the editing process. Such a framework would make it possible to reason about interventional and counterfactual queries, providing a more principled way to achieve controllable and interpretable edits. With careful attention to both technical improvements and ethical implications, we aim to enhance the robustness of the model, simplify its training process, and expand its generative capabilities to unlock new possibilities for Artificial Intelligence Generated Content.

Data availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Received: 27 August 2024; Accepted: 31 October 2025
Published online: 28 November 2025

References

1. Goodfellow, I. et al. Generative adversarial nets. *Advances in neural information processing systems* **27** (2014).
2. Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR: arxiv:1511.06434 (2015).
3. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. ArXiv:1710.10196 (2017).
4. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 4396–4405 (2018).
5. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D. & Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. *2021 IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)* 2065–2074 (2021).
6. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (2021).
7. Gal, R. et al. Stylegan-nada. *ACM Transactions on Graphics (TOG)* **41**, 1–13 (2021).
8. Kwon, G. & Ye, J.-C. Clipstyler: Image style transfer with a single text condition. *2022 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 18041–18050 (2021).
9. Xia, W., Yang, Y., Xue, J. & Wu, B. Tedigan: Text-guided diverse face image generation and manipulation. *2021 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 2256–2265 (2021).
10. Yu, Y. et al. Towards counterfactual image manipulation via clip. *Proc. 30th ACM Int. Conf. on Multimed.* (2022).
11. Revanur, A., Basu, D., Agrawal, S., Agarwal, D. & Pai, D. Coralstyleclip: Co-optimized region and layer selection for image editing. *2023 IEEE/CVF CConf. on Comput. Vis. Pattern Recognit. (CVPR)* 12695–12704 (2023).
12. Liu, W., Rabinovich, A. & Berg, A. C. Parsenet: Looking wider to see better. ArXiv:1506.04579 (2015).
13. Karras, T. et al. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 8107–8116 (2019).
14. Liu, H., Song, Y. & Chen, Q. Delving stylegan inversion for image editing: A foundation latent space viewpoint. *2023 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 10072–10082 (2022).
15. Zhu, Y.-C. et al. One model to edit them all: Free-form text-driven image manipulation with semantic modulations. ArXiv:2210.07883 (2022).
16. Hou, X., Zhang, X., Li, Y. & Shen, L. Textface: Text-to-style mapping based face generation and manipulation. *IEEE Transactions on Multimed.* **25**, 3409–3419 (2023).
17. Lin, T.-Y. et al. Microsoft coco: Common objects in context. In *European Conference on Computer Vision* (2014).
18. Agrawal, A. et al. Vqa: Visual question answering. *Int. J. Comput. Vis.* **123**, 4–31 (2015).
19. Vaswani, A. et al. Attention is all you need. In *Neural Information Processing Systems* (2017).
20. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics* (2019).
21. Li, J. et al. Align before fuse: Vision and language representation learning with momentum distillation. In *Neural Information Processing Systems* (2021).
22. Lee, K.-H., Chen, X., Hua, G., Hu, H. & He, X. Stacked cross attention for image-text matching. ArXiv:1803.08024 (2018).
23. Yu, L. et al. Mattnet: Modular attention network for referring expression comprehension. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1307–1315 (2018).
24. Reed, S. et al. Generative adversarial text to image synthesis. In *International conference on machine learning*, 1060–1069 (Pmlr, 2016).
25. Karras, T. et al. Alias-free generative adversarial networks. In *Neural Information Processing Systems* (2021).
26. Abdal, R., Zhu, P., Mitra, N. J. & Wonka, P. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)* **40**, 1–21 (2020).
27. Li, Z., Min, M. R., Li, K. & Xu, C. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. *2022 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 18176–18186 (2022).
28. van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. ArXiv:1807.03748 (2018).
29. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O. & Cohen-Or, D. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)* **40**, 1–14 (2021).
30. Richardson, E. et al. Encoding in style: a stylegan encoder for image-to-image translation. *2021 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 2287–2296 (2020).
31. Deng, J., Guo, J. & Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 4685–4694 (2018).
32. Liu, Z., Luo, P., Wang, X. & Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)* (2015).
33. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit.* 586–595 (2018).
34. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Process.* **13**, 600–612 (2004).
35. Li, J., Li, D., Xiong, C. & Hoi, S. C. H. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning* (2022).
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695 (2022).

37. Kim, G., Kwon, T. & Ye, J. C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2426–2435 (2022).
38. Ulhaq, A. & Akhtar, N. Efficient diffusion models for vision: A survey. *arXiv preprint* arXiv:2210.09292 (2022).
39. Zheng, D. Diffusion models on the edge: Challenges, optimizations, and applications. *arXiv preprint* arXiv:2504.15298 (2025).

Author contributions
Conceptualization by Y.P. and J.P; methodology by Y.P.; experiments by Y.P., Y.W. and C.W.; resources by X.W.; writing-review and editing by Y.P., Y.W., C.W. and J.P.. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported in part by the Natural Science Foundation of Chongqing under Grant CSTB2022N-SCQ-LZX0040, CSTB2023NSCQ-LZX0012, CSTB2023NSCQ-LZX0160.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to J.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.