Prompt-Based Modality Bridging for Unified Text-to-Face Generation and Manipulation

YIYANG MA and HAOWEI KUANG, Wangxuan Institute of Computer Technology, Peking University, Beijing, China
HUAN YANG and JIANLONG FU, Microsoft Research, Beijing, China
JIAYING LIU, Wangxuan Institute of Computer Technology, State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China

Text-driven face image generation and manipulation are significant tasks. However, such tasks are quite challenging due to the gap between text and image modalities. It is difficult to utilize current methods to deal with both of the two problems because these methods are usually designed for one certain task, limiting their application in real scenarios. To address the two problems in one framework, we propose a **Unified Prompt-based Cross-Modal Framework (UPCM-Frame)** to bridge the gap between the text modality and image modality with CLIP and StyleGAN, which are two large-scale pre-trained models. The proposed framework is combined with two main modules: a Text Embedding-to-Image Embedding projection module based on a special prompt embedding pair, and a projection module mapping Image Embeddings to semantically aligned StyleGAN Embeddings which can be used in both image generation and manipulation. The proposed framework is able to handle complicated descriptions and generate impressive results with high quality due to the utilization of large-scale pre-trained models. In order to demonstrate the effectiveness of the proposed method in the two tasks, we conduct experiments to evaluate the results of our method both quantitatively and qualitatively.

CCS Concepts: • **Computing methodologies** → **Computer vision**;

Additional Key Words and Phrases: Text-to-image translation, text-to-image semantic alignment, prompt embedding

ACM Reference format:
Yiyang Ma, Haowei Kuang, Huan Yang, Jianlong Fu, and Jiaying Liu. 2024. Prompt-Based Modality Bridging for Unified Text-to-Face Generation and Manipulation. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 12, Article 386 (November 2024), 23 pages.
https://doi.org/10.1145/3694974

Yiyang Ma and Haowei Kuang contributed equally to this research.
This work was supported in part by the National Natural Science Foundation of China under Grant 62332010, and in part by the State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China.
Authors' Contact Information: Yiyang Ma, Wangxuan Institute of Computer Technology, Peking University, Beijing, China; e-mail: myy12769@pku.edu.cn; Haowei Kuang, Wangxuan Institute of Computer Technology, Peking University, Beijing, China; e-mail: kuanghw@stu.pku.edu.cn; Huan Yang, Microsoft Research, Beijing, China; e-mail: hyang@fastmail.com; Jianlong Fu, Microsoft Research, Beijing, China; e-mail: jianf@microsoft.com; Jiaying Liu (corresponding author), Wangxuan Institute of Computer Technology, State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China; e-mail: liujiaying@PKU.EDU.CN.

ACM 1551-6865/2024/11-ART386
https://doi.org/10.1145/3694974

1 Introduction

Text-driven facial image generation and manipulation have drawn great attention, which are widely desired in many visual tasks, such as book character generation with diverse emotions and automatic portrait creation under different conditions. Such scenarios containing text descriptions of persons lead to the tasks of face image generation and manipulation. The tasks are also vital in many other practical application scenes, such as automatic portrait drawing and image retouching. In the past, only humans were able to accomplish these tasks due to the challenges below. The first one is that text descriptions can be abstract and intricate, making it challenging to generate high-quality visualizations accurately. These descriptions may encompass not only specific attributes such as hair length or skin color but also abstract characters like mood or personality. The next one is that generated or manipulated face images must have impressively fine-grained visual details with high quality, otherwise, they would not satisfy the users. In this article, we manage to handle both of the tasks of face image generation and manipulation within one unified framework and solve the two problems simultaneously.

We briefly discuss existing remarkable text-to-image generation methods [3, 12, 18, 22, 29, 31, 33, 36, 39, 41–43, 48, 52] first. The works in previous years [18, 29, 33, 39, 41, 43, 48, 52] train text-to-image models with paired data, which heavily rely on the quality of dataset during training, limiting the quality of the generated images and restricting their ability to handle open-world words. TediGAN-B [42] utilizes a pre-trained language model to extract semantics contained in texts and forces the semantics of images to be close to the text semantics by optimizing the images themselves. However, it still encounters difficulties in certain scenarios (e.g., the text contains complex semantics, as shown in Figure 4) because its optimization method cannot take full advantage of pre-trained vision-language models and is limited by the initial image. Large-scale models trained on open-domain data like DALLE-2 [31] or Imagen [36] can handle the semantics in texts but may fail in the task of generating face images because of the gap between their training data and the certain domain of face images. Besides, such big models are quite time-consuming and computing resource-consuming, making them impractical for most normal users.

Then, we conclude image manipulation methods [1, 2, 6, 28] in the aspects of their advantages and drawbacks. There have been several methods [6, 28] with combined **Contrastive Language-Image Pre-Training (CLIP)** [30] and StyleGAN [15, 16] which utilize CLIP to align the semantics in images and texts. However, there are several drawbacks in these methods. StyleCLIP-O [28] utilizes optimization-based manipulation method which are time-inefficient. StyleCLIP-LM [28] trains specific mappers which can only manipulate one attribute with one model. StyleCLIP-GD [28] sweeps complicated hyper-parameters during inference. StyleGAN-NADA [6] focuses on the transformation between different domains of images, but cannot manipulate images in single attribute. Diffusion models [11], as powerful generative models, are also utilized in the task of image manipulation. Blended-Diffusion [1] leverages CLIP [30] to constrain the semantics of manipulated images, but its requirement of mask limits its application. InstructPix2Pix [2] applies pre-trained StableDiffusion [34] as the generator, but performs unsatisfactorily in disentanglement because the manipulating guidance and the image generator are entangled. Most importantly, most of the current image manipulation methods take the image to be manipulated as a part of input. This characteristic means that these methods are dependent on existing images. Although they can be leveraged as image generation methods by manipulating an existing image guided by texts, the semantics contained in the origin image which are not controlled may be contrary to the texts, causing unsatisfactory results.

Given the current situation that existing methods are not able to handle both the tasks of image generation and manipulation, we are motivated to design an approach to unify the two tasks

into one single framework. The main challenge lies in the gap between text modality and image modality. It is critical to bridge the gap, otherwise it would be difficult to transform the semantics contained in the input texts to the generated or manipulated images. Furthermore, the quality of output images is also vital. Facing these challenges, our key idea is to utilize two large-scale pre-trained models, StyleGAN [15, 16] and CLIP [30]. StyleGAN, as one of the most notable GAN [7] backbones, ensures the image quality. Meanwhile, CLIP is a large-scale text-image aligning model, which could extract aligned semantic representations from different modalities. With the two pre-trained models, we can transform the task of bridging the abstract gap between different modalities into bridging different latent spaces of the pre-trained models, which is more concrete to be solved.

Based on this idea, we propose an efficient framework named as **Unified Prompt-Based Cross-Modal Framework (UPCM-Frame)** which is able to address both the text-to-image generation and text-guided image manipulation simultaneously. To solve the problem of bridging the gap between different latent spaces, we introduce two main modules. Overall, we manage to encode input texts into CLIP Text Embeddings and project the CLIP Text Embeddings to their corresponding CLIP Image Embeddings in the first module. Then we build another mapping module from image embeddings to StyleGAN Embeddings within the $\mathcal{W}$ space of StyleGAN, which can be used to generate semantically aligned images with high quality. To be more specific, in the first projecting module, we propose to employ a specific pair of CLIP Embeddings as prompts, which help to bridge the gap between the two latent spaces. In our work, prompt embeddings represent "a neutral text description" and "a neutral image." We add the distance between the input text embedding and text prompt embedding to image prompt embedding. The second image-to-StyleGAN mapping module, containing a trainable deep neural network named **CLIP-to-StyleGAN (C2S)**, maps the input image embeddings to the StyleGAN $\mathcal{W}$ latent space. The training data of the network are randomly sampled with help of the pre-trained StyleGAN. Thus, the training of the network does not need any external data, avoiding the limitation of training datasets we discuss before. In order to keep the semantics during the projection, we further design a semantic consistency loss, guaranteeing the projected StyleGAN embedding can be utilized to generate semantically aligned images.

For image generation task, we can simply leverage the projected StyleGAN embedding at last to generate the corresponding image. For image manipulation task, we further calculate the distance between the projected StyleGAN embedding and the prompt StyleGAN embedding. Then, we apply the distance to the inverse StyleGAN embedding of the input image. Thus, the projected StyleGAN embedding can be used in the both tasks, solving the two problems simultaneously. In addition, comprehensive experiments are conducted to analyze the image generation and manipulation performance of the proposed UPCM-Frame, including comparisons between different methods and ablation studies.

This article is an extension of our previous work [24]. Our extension lies in the methodology and experiments. For the method, we extend our image generation framework to a more comprehensive joint generation-manipulation framework in Section 3.4. We design a method of migrating the semantics contained in the texts to existing images by adding the distance between the semantics of input text and prompt to the input image, solving the task of manipulation. Besides, we move the StyleGAN Embedding from $\mathcal{Z}$ space to $\mathcal{W}$ space to achieve better disentanglement and image diversity in Section 3.3. Correspondingly, we re-design the regularization loss employed in the training process of the C2S projection network, ensuring the projected embeddings are within the $\mathcal{W}$ space. In terms of extensive experimental results, we provide more comparison results with other state-of-the-art methods of both face image generation and manipulation in Section 4.3 to demonstrate the efficiency of the proposed method and more complete ablation studies in Section 4.5.

In summary, this work has the following contributions:

—We propose a unified framework that can handle the text-to-face generation and manipulation tasks with high semantic consistency, fidelity, and image quality. Compared with previous works, our method can solve the two problems simultaneously.
—We design a prompt-based projection method between the two latent spaces of CLIP and analyze the characteristics of prompt embeddings, bridging the semantic gap between different modalities.
—We develop a C2S network to map Image Embeddings to their corresponding StyleGAN Embeddings. A novel semantic consistency loss is further proposed to guarantee that the projected StyleGAN Embeddings are semantically aligned with the input texts.

The remainder of the article is organized as follows. In Section 2, we briefly review current text-to-image cross-modality works. In Section 3, we present the methodology of the proposed UPCM-Frame. Later in Section 4, we provide qualitative and quantitative comparisons to prove the effectiveness of the proposed method in both of the image generation and manipulation tasks. We also provide implementation details, more experimental results, ablation studies on loss functions and the prompt design, and discussion on limitations in Section 4. Finally, we conclude our work in Section 5.

2 Related Work

2.1 Text-to-Image Alignment

To transform a text into an image, it is essential to ensure the semantic alignment between the text and the image. In the early stages of research, various methods were developed to achieve this goal which train separate text and image encoders [4, 13, 19, 21, 38, 44–46, 49–51]. However, these methods which are utilized in specific text-to-image translation tasks are limited by the vocabulary of training datasets.

Recently, the success of attention mechanism [40] in natural language processing has led to the adoption of transformers [40] as baseline models for multimodal tasks. CLIP [30], as a notable example, is built by training two transformer encoders on a large corpus of text-image pairs collected from various sources online. This model consists of two latent spaces: one for texts and another for images. Additionally, researchers have discovered multi-modal neurons in CLIP [30], which has inspired further exploration in this area.

In this work, we adopt CLIP as our text-to-image alignment checking module. By leveraging its capabilities, we aim to achieve accurate semantic alignment between texts and images to provide supervision for the task of text-to-image translation.

2.2 Text-to-Image Translation

Text-to-image translation methods can be roughly classified into two categories based on the generation models they employ. The first category does not make use of pre-trained models, such as StyleGAN [15, 16]. This kind of methods build images generators from scratch. GAN-INT-CLS [33], as the pioneer work, employs a conditional GAN [26] guided by text embeddings extracted from a pre-trained text encoder to achieve the goal of text-to-image translation. Following this approach, DM-GAN [52] introduces a memory writing gate, and DF-GAN [39] proposes a backbone that generates images with Wasserstein distance. Another notable example is DALLE [32], which has about 12 billion parameters and exhibits great quality in the task of text-to-image translation.

The second category utilizes pre-trained generation models to improve the image quality and shorten the training process. However, due to the domain limitations of these models, the images

they are able to generate may have specific constraints. For instance, TediGAN-A [41] maps input text to the latent space of StyleGAN, and TediGAN-B [42] optimizes an embedding in StyleGAN's latent space utilizing cosine similarity of text and image embeddings encoded by CLIP as a loss function. Due to the usage of CLIP, TediGAN-B [42] can handle open-world texts. However, its performance can be random and visually unappealing. Moreover, StyleCLIP [28] proposes three methods of image manipulation, and the optimization method can be leveraged as an method of image generation by providing an origin image. StyleGAN-NADA [6] transfers images to other domains by fine-tuning StyleGAN with the guidance of texts. In our work, we adopt pre-trained StyleGAN2 [16] for generating images from textual descriptions. Our method bridges the latent spaces of different pre-trained models, leveraging the abilities of existing models and reducing the time and computing resource cost.

2.3 Text-Guided Image Manipulation

Different with text-to-image translation, text-guided image manipulation aims to manipulate input images guided by the input texts describing desired attributes, but with the properties of non-descriptive image parts left unchanged. Based on the similarity between manipulation task and generation task, numerous image manipulation methods with existing generative models such as StyleGAN [15, 16] have been proposed to achieve improved image quality and shorter training processes. These techniques [9, 27, 28, 41, 42, 47] encode images into a latent space, allowing for manipulation through modifications to the latent vectors. In TediGAN-A [41], a visual-linguistic similarity module is proposed to align two modalities in the latent space of pre-trained StyleGAN [16]. More recently, StyleCLIP [28] combines the generative power of StyleGAN [16] with the image-text representation ability of CLIP [30] to explore manipulation directions, eliminates the need for text during the training process and enables zero-shot inference. Recently, diffusion model [11, 35] based image manipulation methods [1, 2, 25, 31] have achieved excellent performance. These methods adopts pre-trained diffusion models [34] as generators, achieving high image quality and diversity. However, they usually performs unsatisfactory disentanglement because the guidance and generator are entangled. Our method designs an editing embedding which is irrelevant to the input image in the latent space of StyleGAN. Thus, our framework can be used in both image generation and manipulation because of the input-independent latent embedding.

3 Prompt-Based Modality Bridging

In this section, we depict the proposed prompt-based modality bridging method. The method projects input CLIP Text Embeddings (which is abbreviated as CTE_{input} and C denotes CLIP [30]) to their corresponding CLIP Image Embeddings (which is abbreviated as CIE_{input}) by a prompt embedding pair CTE_{prompt} and CIE_{prompt}, as shown in Figure 1(a) and (b). Then, CIE_{input} and CIE_{prompt} are further projected to their StyleGAN $\mathcal{W}$ Embeddings (abbreviated as SE_{input} and SE_{prompt} respectively). For image generation, the SE_{input} is directly leveraged by pre-trained StyleGAN [16] to generate the semantically-aligned image, as shown in Figure 1(c). (1) For image manipulation, we compute the distance between SE_{input} and SE_{prompt} and add the distance to the inverse StyleGAN Embedding of the image to be manipulated (abbreviated as $SE_{inverse}$), getting the edited StyleGAN Embedding (i.e., SE_{edited}) which can be used to generated the manipulated image, as shown in Figure 1(c). (2) We introduce the method in detail in the following subsections.

3.1 Design of Prompt Embedding

ptThe first step of modality bridging is projecting CTE_{input} to CIE_{input} by a pair of prompt embeddings. As we point out in Section 1, CTE_{prompt} and CIE_{prompt} represent "a neutral text description" and "a neutral image" respectively, ensuring that the distances between the prompt embeddings

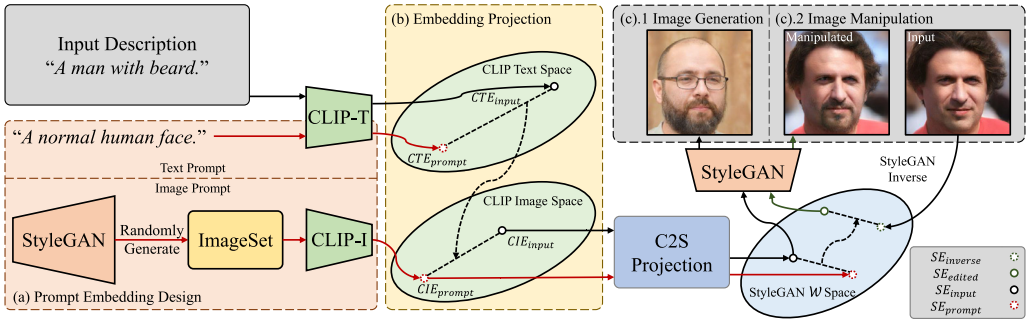

Fig. 1. The entire framework of UPCM-Frame. CLIP-T and CLIP-I denote CLIP text encoder and CLIP image encoder. In (a), we show the method of designing prompts. The CIE_{prompt} is extracted from a randomly sampled set of images from StyleGAN and the CTE_{prompt} is obtained from a certain sentence by CLIP text encoder. In (b), we illustrate the projection process with help of the prompt embedding pair obtained in (a). At last, (c) shows results of image generation and manipulation and the method of applying the Input SE and Prompt SE in these tasks. The detailed architecture of C2S projection network is further shown in Figure 2. SE, StyleGAN $\mathcal{W}$ Embeddings.

and input embeddings will not being long. As a result, the design of prompt embedding pair is vital to the projection. The method of leveraging the prompt embedding pair is introduced in Section 3.2 in detail.

In order to ensure that the prompt embeddings can represent the centers of the corresponding latent space, the average cosine similarity between the prompt embedding and all the other embeddings in the latent space should be the largest. We employ a large subset of embeddings to represent the whole latent space, containing n embeddings. We depict the method of obtaining such a subset of latent embeddings in the end of the subsection. The lengths of all the embeddings are normalized because the semantics only depend on their orientations. Denoting $\mathbf{y}$ as the prompt embedding and $\mathbf{x}_i$ as the ith embedding in the subset, the target described before can be formulated as a programming problem:

$$\begin{aligned} &\max_{\mathbf{y}} z = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{y}\cdot\mathbf{x_i}}{|\mathbf{y}|\cdot|\mathbf{x_i}|}, \\ &\text{s.t. } |\mathbf{y}| = 1, \end{aligned} \tag{1}$$

where z denotes the average costing similarity. It is difficult to solve such a non-linear programming problem directly. Thus, we manage to leverage the physical meanings of such problem to simplify it.

As we mention before, the lengths of all the embeddings are all normalized. Thus, Equation (1) can be transformed into

$$\begin{aligned} \max_{\boldsymbol{y}} z &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}\cdot\boldsymbol{x_i} \\ &= \boldsymbol{y}\cdot\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x_i}. \end{aligned} \tag{2}$$

The physical meaning of Equation (2) is a hyperplane in the dimension of the latent space and z represents the constant parameter. Larger absolute value of z means bigger distance between the hyperplane and the origin point. Besides, the feasible region of the programming problem is $|\mathbf{y}| = 1$, which is a hypersphere. As the target of the problem is maximizing z, we can move the hyperplane as possible from the origin point, until the hypersphere and the hyperplane are tangent. At this

time, $\mathbf{y}$ will be the unit normal vector of the hyperplane. The unit normal vector of the hyperplane is given as

$$\boldsymbol{y}' = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \boldsymbol{y} = \frac{\boldsymbol{y}'}{|\boldsymbol{y}'|}. \tag{3}$$

It is obvious that the vector $\mathbf{y}'$ is the arithmetic mean value of all the n embeddings in the subset.

Then, we introduce the specific method of getting image prompt embedding CIE_{prompt} and text prompt embedding CTE_{prompt}. For CIE_{prompt}, we randomly generate images by pre-trained StyleGAN [16] and extract their CIEs by CLIP [30], getting a set of CIEs which can be utilized to obtain the CIE_{prompt} via Equation (3). For CTE_{prompt}, it is non-trivial to get a descriptive texts with diverse semantics containing all the attributes. Besides, texts themselves are carriers of semantics. Thus, we appoint a specific sentence and extract its CTE as CTE_{prompt}, following the method introduced in NLP domain [20]. For instance, we set the CTE of the sentence "A normal human face." as the CTE_{prompt}. Such setting is further discussed in Section 4.5. It is noted that if there exists a high-quality text set, the proposed method can be utilized to extract a better CTE_{prompt} than the manually specified one.

3.2 CLIP Embedding Projection with Prompt Pair

In this subsection, we introduce the method of projecting CTE_{input} of the input text description to its corresponding CIE_{input} which can be leveraged in both image generation and manipulation by the prompt embedding pair CTE_{prompt} and CIE_{prompt}. As we discuss before, the prompt embedding pair represents the centers of each latent space, thus we can move the distance between CTE_{input} and CTE_{prompt} to CIE_{prompt}, getting CIE_{input}. In conclusion, the projection can be formulated as

$$CIE_{input} = CIE_{prompt} + (CTE_{input} - CTE_{prompt}). \tag{4}$$

Such a projection is simple yet effective. The simple summation represents the idea of "adding a specific prompt to implement the query" which is introduced in [20] which is the main perspective of prompt. In addition, we propose that the subtraction can be regarded as "removing a certain part of the origin embedding to exclude the exact formulation." Then, CIE_{input} can be further projected to SE_{input}.

We give a brief discussion of why such a simple linear operation works. In short, it is guaranteed by the characteristic of CLIP [30] itself. We explain it by giving an example. Assuming that there exists a text-image pair which has aligned semantics (e.g., "A girl with short hair." and a corresponding image), we manually edit one single attribute of the image without affecting any other attributes and change the text correspondingly (e.g., lengthen the hair of the girl and replace the origin sentence with "A girl with long hair."). In this way, we obtain a new pair text and image which is still semantically aligned. Thus, the *CTE-CIE* pair also has large cosine similarity due to the ability of CLIP (otherwise, we can repeat such manipulating operations and finally getting a semantically aligned text-image pair with *CTE-CIE* pair which has low cosine similarity, and this case is in conflict with the characteristic of CLIP). This example proves that if we manipulate both the text and image with the same semantic, their CTE and CIE will change roughly collinearly. Similar linear operations are also used in previous works [6, 28].

Another perspective to explain the linear operation is thinking about the "semantic distance" in two spaces. Assuming that there are two semantically aligned pair of texts and images, we name their CIEs and CTEs as CIE_1, CTE_1 and CIE_2, CTE_2, respectively. Due to the characteristic of CLIP [30], CIE_1 and CTE_1 are roughly co-linear, also for CIE_2 and CTE_2. Thus, the distances between CIE_1 and CIE_2 should be close to the distance between CTE_1 and CTE_2, otherwise it will

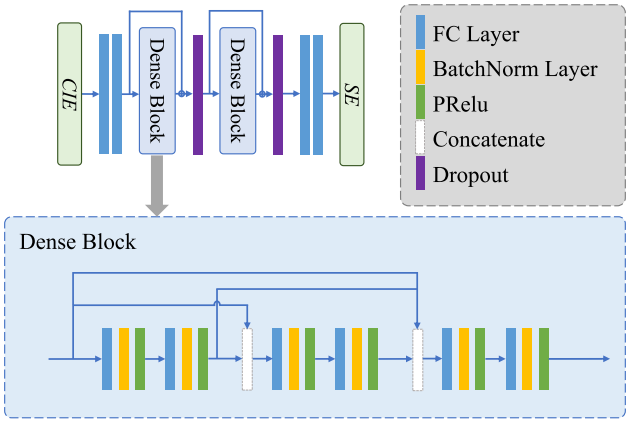

Fig. 2. The detailed architecture of C2S projection network. FC, fully connected layer.

be contrary to the co-linearity. Hence, we have

$$CIE_1 - CIE_2 = CTE_1 - CTE_2. \tag{5}$$

In our scenario, the prompt embedding pair CIE_{prompt} and CTE_{prompt} represent the centers of each latent space. Thus, they are semantically aligned embeddings. Thus, we can simply replace one of two pairs in Equation (5) and transform it into

$$CIE_1 = CIE_{prompt} + CTE_1 - CTE_{prompt}, \tag{6}$$

where CIE_1 and CTE_1 are also a pair of aligned embeddings. In this way, we can get the semantically aligned CIE of a certain input CTE_{input}, which is actually Equation (4).

In practice, we multiply the difference between CTE_{input} and CTE_{prompt} with a constant factor, controlling the distinctiveness of projection. The final equation of projection is

$$CIE_{input} = CIE_{prompt} + \alpha \cdot (CTE_{input} - CTE_{prompt}), \tag{7}$$

where α is the factor which can be changed. Empirically, we set it to 1.75, which is an appropriate value for most cases.

3.3 CLIP to StyleGAN Embedding Projection

In order to leverage the CIE_{input} in image generation and manipulation with the help of pre-trained StyleGAN, we project CIE_{input} to its corresponding SE_{input} which can be used to generate an image whose CIE is CIE_{input}. We build a neural network containing **fully connected (FC)** layers with dense connections, named as C2S projection network. Its architecture is given in Figure 2.

We first build a dataset containing a large amount of *CIE-SE* pairs to train the C2S network. We randomly sample *SE*s in the StyleGAN $\mathcal{W}$ space by sampling from $\mathcal{Z}$ space (standard normal distribution) and mapping the $\mathcal{Z}$ space embeddings to *SE*s by the StyleGAN mapping network. Then we generate images by feeding the *SE*s to the pre-trained StyleGAN and extract their *CIE*s of the resulting images by CLIP. By this means, we can get infinite training pairs for C2S network.

In terms of the loss function, it should meet two requirements as follows. First, the network should ensure the projected *SE*s can be leveraged to generate images which have close *CIE*s as inputs. Second, the projected *SE*s should be in the $\mathcal{W}$ space of StyleGAN.

The projected *SE*s should have the same semantics as input *CIE*s. To this end, we utilize the projected *SE*s to generate images, extract *CIE*s of them and take minimizing the cosine similarity between the *CIE*s of generated images and the input *CIE*s as a loss function. Such loss function is

called reconstructed semantics consistency loss, $\mathcal{L}_{sem\_cons}$. Denoting G as the pre-trained StyleGAN and CLIP_I as the image encoder of CLIP, the loss is given by

$$\mathcal{L}_{sem\_cons} = CosDis(CIE_{input}, \text{CLIP}_I(G(SE_{pred}))). \tag{8}$$

To guarantee that the projected *SE*s are in the $\mathcal{W}$ space, we simply leverage an **mean square error (MSE)** loss between the images generated from true *SE*s and projected *SE*s. The loss ensures the quality of generated images, which means the projected *SE*s are in the latent space of StyleGAN. The loss is called regularization loss. Denoting the *SE* in the training *CIE-SE* pairs as SE_{true}, the loss is

$$\mathcal{L}_{reg} = ||G(SE_{pred}) - G(SE_{true})||_2. \tag{9}$$

We also leverage a vanilla L1 loss between the projected *SE*s and true *SE*s as a basic constraint, which is given by

$$\mathcal{L}_{L1} = ||SE_{pred} - SE_{true}||_1. \tag{10}$$

In summation, the entire loss employed during training is

$$\mathcal{L} = \lambda_{sem\_cons} \cdot \mathcal{L}_{sem\_cons} + \lambda_{L1} \cdot \mathcal{L}_{L1} + \lambda_{reg} \cdot \mathcal{L}_{reg}, \tag{11}$$

where λ_{sem_cons}, λ_{L1}, and λ_{reg} are the weights of each of the loss terms. The specific values of them are given in Section 4.2.

3.4 Image Generation and Manipulation from StyleGAN Embedding

After getting the StyleGAN embedding SE_{input}, we can use it in both image generation and manipulation. For image generation, we can simply feed SE_{input} to the pre-trained StyleGAN. For image manipulation, we first inverse the origin image I_{origin} to its corresponding SE_{origin} by optimizing a randomly initialized *SE* minimizing a combination of MSE and LPIPS between the origin image and the generated image

$$\begin{aligned} SE_{inverse} = \arg\min_{SE} (&||I_{origin} - G(SE)||_2 \\ &+ \lambda_{percep} \cdot \text{LPIPS}(I_{origin}, G(SE))). \end{aligned} \tag{12}$$

Then we add the difference between SE_{input} and SE_{prompt} to $SE_{inverse}$, getting the manipulated SE_{edited}. The calculation is given by

$$SE_{edited} = SE_{inverse} + (SE_{input} - SE_{prompt}). \tag{13}$$

Such a simple manipulation is supported by the disentanglement of StyleGAN $\mathcal{W}$ space, which is validated in the origin paper of StyleGAN [15, 16]. Then, the SE_{edited} can be used to generate manipulated image by a pre-trained StyleGAN. Thus, benefiting from the design of prompt embeddings, the proposed method can be a unified framework for both image generation and manipulation. Although, we can utilize image manipulation methods as image generation methods by giving origin images and manipulating the images by the input texts, the proposed method outperforms significantly than such pseudo image generation methods, as we will show later in Section 4.3.

4 Experimental Results

In this section, we conduct experiments to prove the performance of the proposed method. We provide the compared baseline methods of both image generation and image manipulation in Section 4.1, then give implementation details to ensure the reproducibility in Section 4.2. In Section 4.3, we show the results of user studies and compare the results of our method and other baseline methods both quantitatively and qualitatively. We give further experimental results in Section 4.4.

In Section 4.5, we conduct ablation studies on the design of prompt embeddings and loss functions. At last, we give failure cases and discuss the reasons to the limitations of the proposed method in Section 4.6.

4.1 Baselines

To demonstrate the superiority of our method, we compare our results with several state-of-the-art baseline methods. The baselines include two aspects: image generation baselines and image manipulation baselines. The image generation baselines include DF-GAN (CVPR'22) [39], TediGAN-B (arXiv'21) [42], StyleCLIP-O (ICCV'21, "O" denotes "Optimization") [28], and AI Illustrator (ACM MM'22) [24]. The image manipulation baselines include TediGAN-B (arXiv'21) [42], InstructPix2Pix (CVPR'23) [2], and StyleCLIP-GD (ICCV'21, "GD" denotes "Global Direction") [28]. It should be noticed that we leverage TediGAN-B [42], which is an image manipulation method, as a pseudo image generation method by giving it initial image embeddings randomly. Such method of extending TediGAN-B is proposed in its paper own. The results of baseline methods except DF-GAN [39] are generated by their official codes, hyper-parameters and models. For DF-GAN, we re-train it on Multi-Modal CelebA-HQ dataset [41].

4.2 Implementation Details

In this section, we provide implementations details to ensure that all of our results are reproducible. The implementation details include two parts: details of the projection from CTE_{input} to CIE_{input} and details of the projection from CIE_{input} to SE_{input}. They are given in sequence.

We give the details of the projection from CTE_{input} to CIE_{input} first. As we state in Section 3.2, such the projection is a linear operation, thus the details are only about getting prompt embeddings. For the process of getting CIE_{prompt}, we randomly sample 150,000 images to compute the average *CIE*. We employ the sentence "A normal human face." to extract the CTE_{prompt}. Then, we state the details of the projection from CIE_{input} to SE_{input}. The *CIE-SE* pairs of the 150,000 images sampled for getting CIE_{prompt} is also leveraged in the training of the C2S network. The abstract architecture of C2S network is given in Figure 2. The full architecture is shown in Table 1 and the specific architecture of dense blocks is given in Table 2. The entire network is combined with five dense blocks as the body part, two FC layers as the head part, and the tail part respectively. PReLU [8] is used as activation function. Dropout layers [37] with the ratio of 0.1 and BatchNorm layers [14] are applied to help convergence. During training, the batch size is set to 16 and the model is trained for 380,000 iterations. All the training are done on two NVIDIA 2080Ti GPUs. We utilize Adam [17] as the optimizer. The learning rate is set to 1×10^{-4} initially. We use cosine annealing to the learning rate and it will drop to 1×10^{-7} at last. The loss weights, λ_{L1}, λ_{sem_cons}, and λ_{reg} are set to 0.3, 1.0 and 0.3. During the training process of C2S network, we normalize the length of input *CIE* to $\sqrt{512}$ instead of 1. The reason is the outputs of the C2S network are StyleGAN $\mathcal{W}$ embeddings [15], whose lengths are obviously not 1. As the StyleGAN $\mathcal{W}$ embeddings are projected from StyleGAN $\mathcal{Z}$ embeddings, we estimate the lengths of StyleGAN $\mathcal{W}$ embeddings using the lengths of StyleGAN $\mathcal{Z}$ embeddings. The StyleGAN $\mathcal{Z}$ space contains 512 dimensions and the distribution of value of each dimension is i.i.d. standard Gaussian distribution (i.e., $\mathcal{N}(0, 1)$). Thus, the average length of StyleGAN $\mathcal{Z}$ embeddings is $\sqrt{512}$. Hence, we normalize the lengths of input *CIE* to $\sqrt{512}$, reducing the gap between the inputs and outputs of the network. This trick would help the model converge faster and better.

For image manipulation, we inverse the origin image to its corresponding $SE_{inverse}$ by optimizing a randomly initialized *SE*. We optimize the *SE* for 500 iterations with the learning rate of 1×10^{-2} and the weight λ_{percep} is set to 0.01.

Table 1. The Detailed Architecture of C2S Network

	Id	Block Name	In Size	Out Size
Head part	1-0	FC+PReLU	512	512
	1-1	FC+PReLU	512	512
Body part	2-0	Dense block	512	512
	2-1	$1-1 + $2-0	-	-
	2-2	Dropout	-	-
	2-3	Dense block	512	512
	2-4	$2-2 + $2-3	-	-
	2-5	Dropout	-	-
	2-6	Dense block	512	512
	2-7	$2-5 + $2-6	-	-
	2-8	Dropout	-	-
	2-9	Dense block	512	512
	2-10	$2-8 + $2-9	-	-
	2-11	Dropout	-	-
	2-12	Dense block	512	512
	2-13	$2-11 + $2-12	-	-
	2-14	Dropout	-	-
Tail part	3-0	FC+PReLU	512	512
	3-1	FC	512	512

The detailed architecture of dense blocks is given in Table 2. All the dropout layers have a dropout ratio of 0.1.
"$" indicates the layer output with the corresponding Id, "FC" indicates "fully connected layer."

Table 2. The Detailed Architecture of Each Dense Block

Id	Layer Name	In Size	Out Size
0	FC+BN+PReLU	512	512
1	FC+BN+PReLU	512	512
2	Concat(input, $1)	-	1,024
3	FC+BN+PReLU	1,024	512
4	FC+BN+PReLU	512	512
5	Concat($2, $4)	-	1,536
6	FC+BN+PReLU	1,536	512
7	FC+BN+PReLU	512	512
8	Concat($5, $7)	-	2,048
9	FC+BN+PReLU	2,048	512
10	FC+BN+PReLU	512	512
11	Concat($8, $10)	-	2,560
12	FC+BN+PReLU	2,560	512
13	FC+BN+PReLU	512	512

"$" indicates the layer output with the corresponding Id, "FC" indicates "fully connected layer," and "BN" indicates "BatchNorm layer."

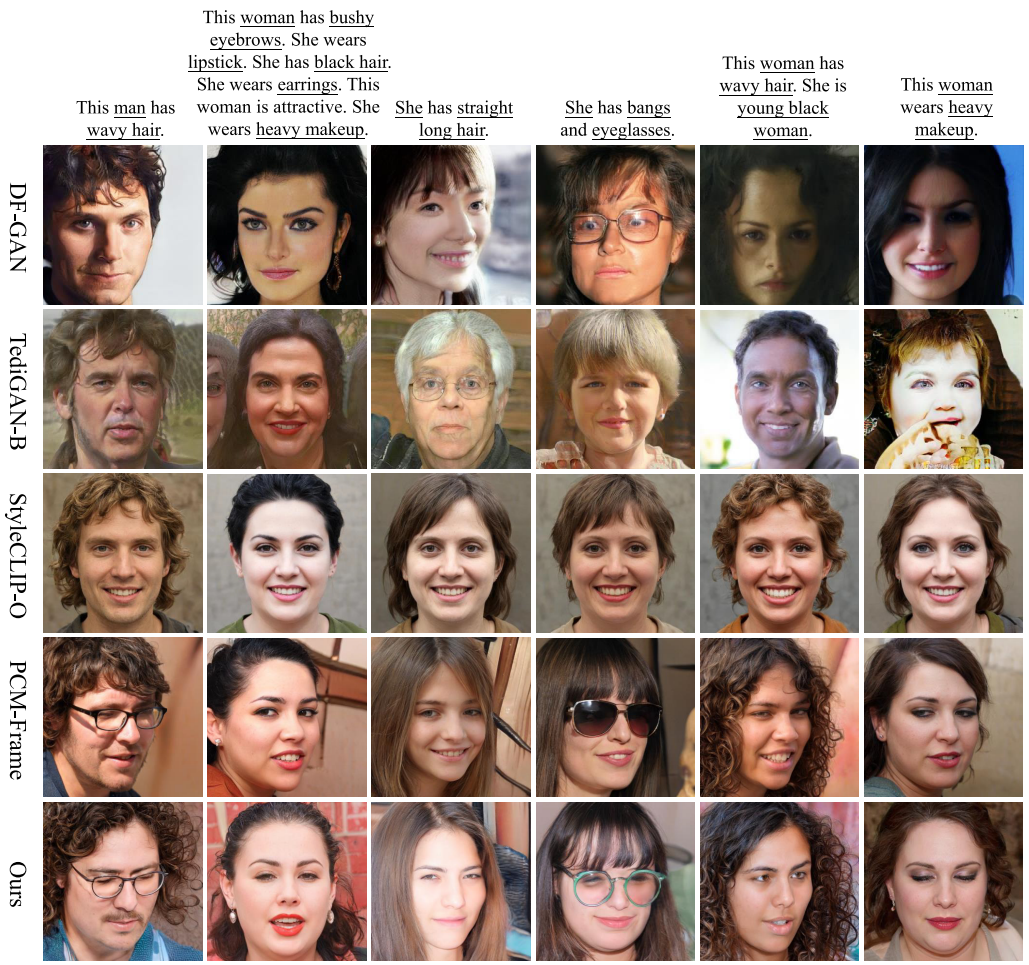

Fig. 3. Face image generation results on descriptions with limited words within a certain dictionary by several methods. The major attributes are underlined. *[Zoom in for best view]*

4.3 Comparisons to State-of-the-Art Methods

The proposed method is a unified framework for both image generation and image manipulation. Thus, we provide the results separately. We show comparisons of the results of different image generation methods first. Because DF-GAN can only generate images from sentences containing only the words in the dictionary of the training dataset, we first show image generation results with such limited sentences. Such results are shown in Figure 3. It should be noticed that the results of applying TediGAN-B as an image generation method are not satisfying, because the initial image embeddings may lead to contrary semantics of the input texts. The proposed method can translate all the semantics in the input texts to the generated images while other methods can not guarantee. Then, we show results of translating open-world texts with more complicated words and semantics in Figure 4. The results further prove the effectiveness of image generation and complicated semantic translation of the proposed method (e.g., handling "wrinkles" and "Cho," which is a Chinese name, leading to an Asian girl). In some cases, the generated facial images have closed eyes which do not in contradiction to the input texts. We speculate that the reason for this

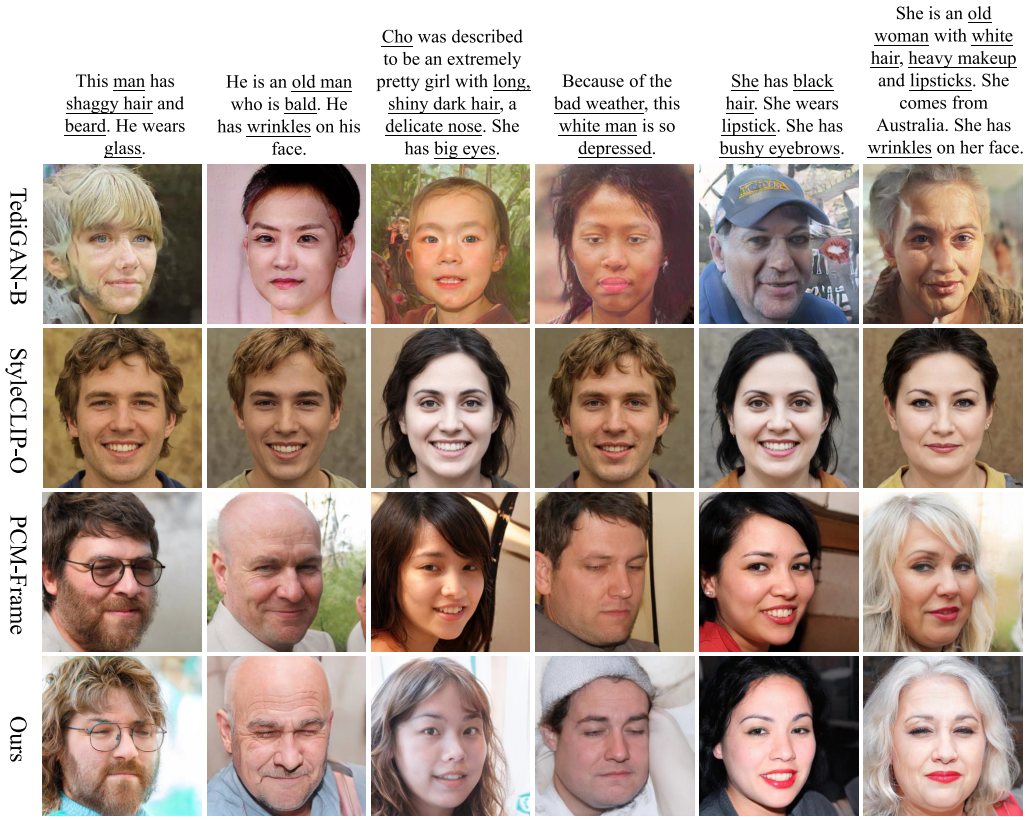

Fig. 4. Face image generation results on descriptions with open-world words by several methods. The major attributes are underlined. *[Zoom in for best view]*

phenomenon is the bias of CLIP on facial images because the input size of CLIP is 224×224, which is much smaller than the StyleGAN image size. The resize process would make the small parts on faces like eyes smaller, which means the closed eyes and slightly open eyes become semantically similar. such the issue can be solved by specifying "big eyes" or "pretty eyes" in the text inputs, which have been shown in Figures 7 and 8.

The method can generate multiple results from one single input text, as shown in Figure 5. Such diversity is supported by the style-mixing characteristic of StyleGAN. We pass a random *SE* to the first few layers of StyleGAN, getting diverse results. To make the results in other sections reproducible, they are generated without applying random style-mixing.

Then, we show image manipulation results from the proposed method and other baseline methods in Figure 6. The original images are collected from the dataset of FFHQ, CelebA [23], or generated by pre-trained StyleGAN. When manipulating, the used phrases will be placed in a whole sentence (e.g., "curly hair" will be extended as "A human face with curly hair." and "plump" will be extended as "A plump face"). It can be seen that the proposed method achieves better disentanglement and semantic alignment to the input texts than state-of-the-art baseline methods, and gets visually more pleasing results (e.g., semantic alignment of "long hair." and visual quality of "curly hair.").

We also give qualitative metrics to demonstrate the efficiency of the proposed method. We leverage FID [10] (when calculating FID, the reference set is a subset containing 10,000 images of

Fig. 5. Different results from one description by applying random style-mixing. The major attributes are underlined.

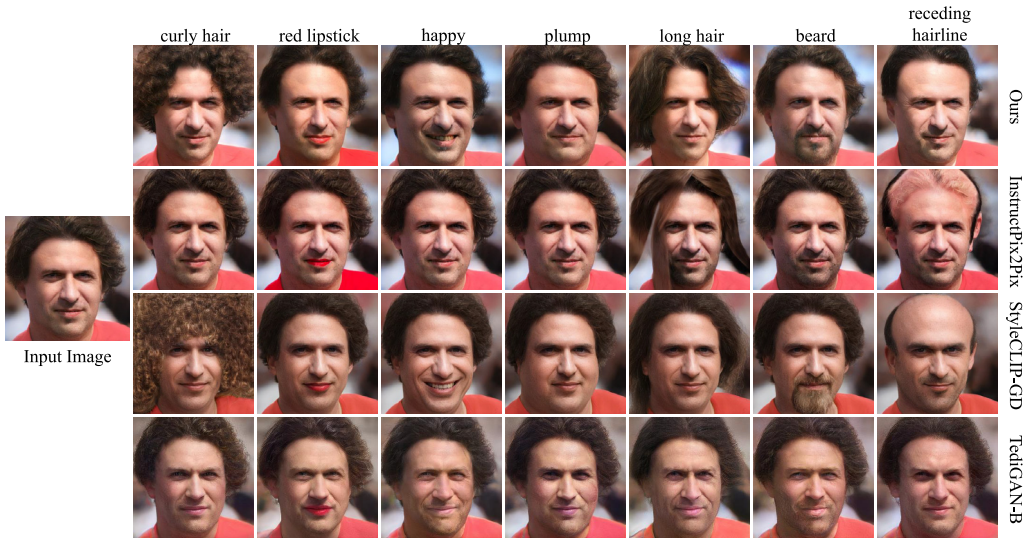

Fig. 6. Manipulation results for human faces. The input image is the inversion to a real image. *[Zoom in for best view]*

FFHQ [15] dataset. The image generation test set and manipulation test set are the sets used in the user study) to evaluate the quality of images and conduct a user study to obtain the subjective metrics. We further utilize CLIP Distance to evaluate the semantic alignment in the generation task (which is not used in the manipulation task because it is not suitable to check whether the whole manipulated images are aligned to the manipulation texts). For image generation, the users are commanded to judge the best one result in the aspect of photo-realistic (abbreviated as Real. Prefer.) and in the aspect of semantic alignment (abbreviated as Acc. Prefer.). The user study contains 80 text-image pairs. For image manipulation, the users are commanded to judge the best two results in the aspect of semantic alignment (abbreviated as Acc. Prefer.) and disentanglement (abbreviated as Dis. Prefer.). The user study contains 34 manipulation cases. In total, the user study contains 114 queries and 12 users, collecting 1,368 votes. The metrics of image generation and manipulation are reported in Tables 3 and 4, respectively. The results show that, for the subjective metric, the proposed method achieves the best FID in both image generation and manipulation; for the objective metrics, the proposed method outperforms all the other methods in the image

Table 3. The Comparisons of Metrics between Our Method and Other Methods in the Task of Face Image Generation

	Subjective		Objective	
	Acc. Prefer. (%) ↑	Real. Prefer.(%) ↑	FID ↓	CLIP Distance ↓
Ours	**49.49**	**43.32**	**114.19**	0.7295
PCM-Frame [24]	40.52	39.40	118.25	0.7417
StyleCLIP-O [28]	4.83	16.16	170.66	0.7603
TediGAN-B [42]	5.16	1.12	122.75	**0.6470**[a]

[a]TediGAN-B takes CLIP Distance as part of optimization target, leading to a much better result on this metric.
The best numbers are **bold**.

Table 4. The Comparisons of Metrics between Our Method and Other Methods in the Task of Face Image Manipulation

	Subjective		Objective
	Acc. Prefer. (%)↑	Dis. Prefer.(%) ↑	FID ↓
Ours	**40.77**	**42.66**	**167.61**
InstructPix2Pix [2]	14.10	8.46	185.26
StyleCLIP-GD [28]	39.26	41.45	170.59
TediGAN-B [42]	5.87	7.43	180.14

The best numbers are **bold**.

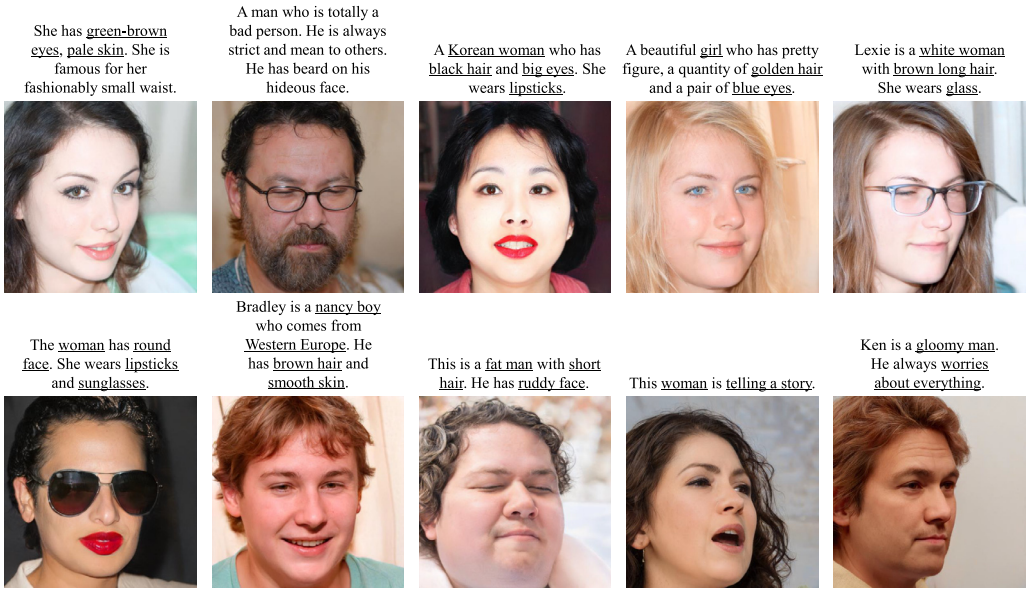

Fig. 7. Further translation results of face images of open-world text descriptions. The major attributes are underlined. *[Zoom in for best view]*

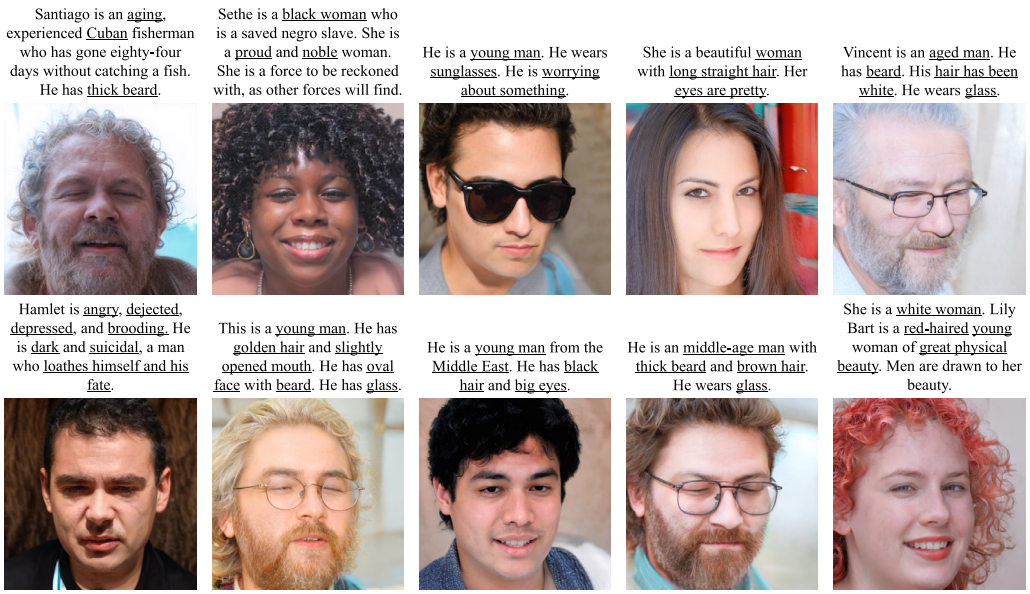

Fig. 8. Further translation results of face images of open-world text descriptions. The major attributes are underlined. *[Zoom in for best view]*

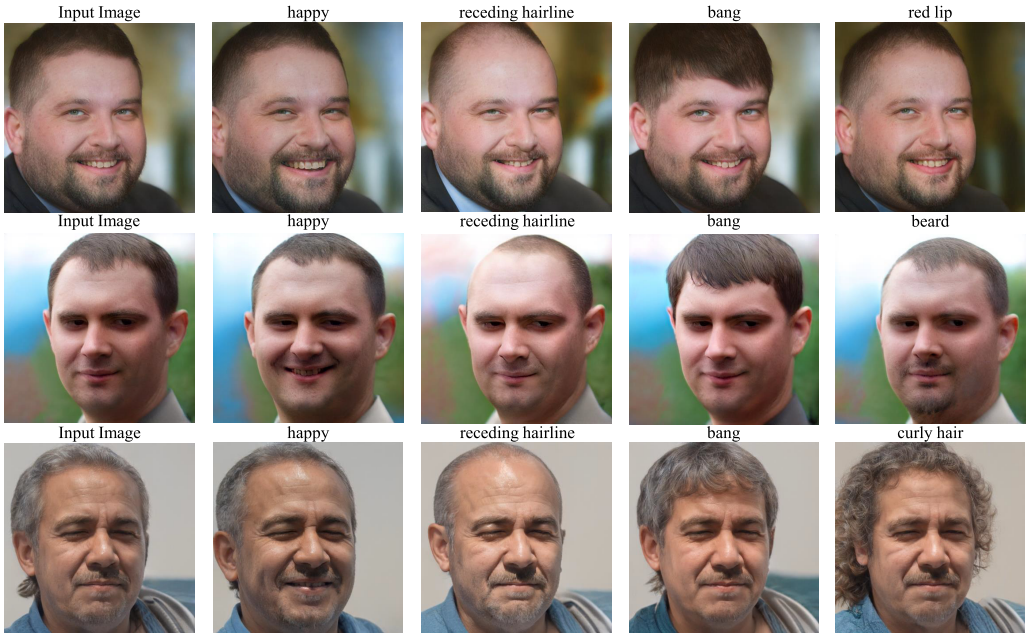

Fig. 9. Further manipulation results of face images of complicated text descriptions. *[Zoom in for best view]*

Fig. 10. Further manipulation results of face images of complicated text descriptions. *[Zoom in for best view]*

generation task and performs comparable with state-of-the-art image manipulation methods in the image manipulation task. It should be noticed that all the results of both image generation and image manipulation of our method are gotten from one unified model. Such results demonstrate the superiority of the proposed method.

4.4 Further Results of Image Generation and Manipulation

In this section, we show more diverse results in both image generation and image manipulation. Image generation results are shown in Figures 7 and 8 while image manipulation results are shown in Figures 9–11. The results show that the proposed method can handle complicated semantics in the texts.

Fig. 11. Further manipulation results of face images of complicated text descriptions. *[Zoom in for best view]*

4.5 Ablation Study

We conduct two ablation studies on the proposed modules. We first study the method of designing the prompt embedding. As we discussed in Section 3.1, the CIE_{prompt} is calculated by averaging a set of *CIE*s and the CTE_{prompt} is extracted from a certain sentence by CLIP. For CIE_{prompt}, we give an intuitive method of getting CIE_{prompt}: Generate an image from an all-zero vector in StyleGAN $\mathcal{Z}$ space and employ its *CIE* as the CIE_{prompt}. Because StyleGAN $\mathcal{Z}$ space follows standard Gaussian distribution, the all-zero vector is the center of the space. Thus, such method makes sense seemingly. We leverage these prompts to generate the 80 texts used in Section 4.3. We give visually comparison results in Figure 12 and give qualitative comparisons in Table 5. It is obvious that such a CIE_{prompt} leads to worse image quality, indicating the "physical center" of StyleGAN latent space is not the "semantic center." For CTE_{prompt}, it seems better to calculate the average embedding from a set of descriptive texts. Thus, we utilize the text set of Multi-Modal CelebA-HQ dataset to get a CTE_{prompt}. The results are also given in Figure 12 and Table 5. It can be seen that such a CTE_{prompt} has distinct bias to a younger age, leading to worse semantically alignment. We consider that the reason is the bias of the dataset itself. The captions in it of children will contain the words like "young," but the captions of adults will not be combined with explicit expressions like "grown." This phenomenon proves the claim we give in Section 3.1: the quality of text set will affect the quality of CTE_{prompt}. However, it is non-trivial to collect such a face-description dataset.

Then, we give ablation studies on the loss design in Figure 13 and Table 6. It can be seen that the regularization loss $\mathcal{L}_{reg}$ assists the network to project *CIE*s to *SE*s within the latent space which

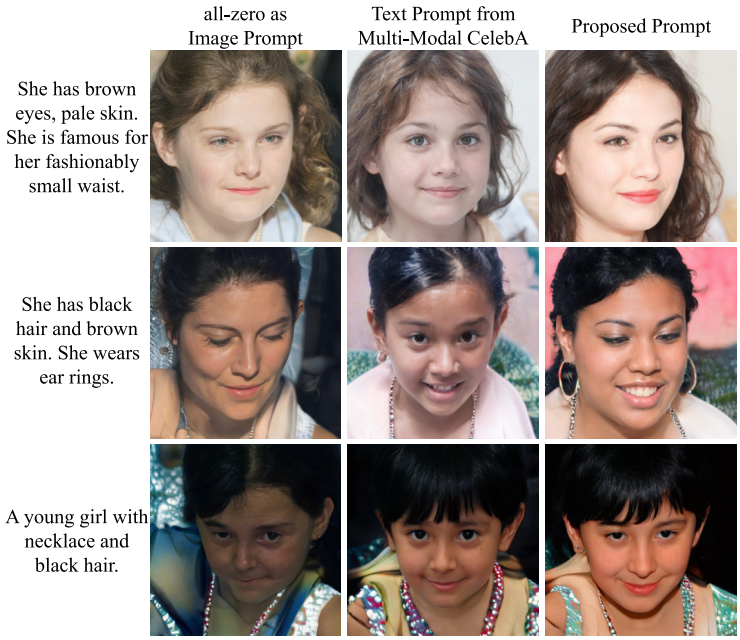

Fig. 12. The ablation study on the prompt design. The results show that the CIE_{prompt} extracted from the image generated from all-zero $\mathcal{Z}$ embedding leads to incorrect semantics and worse image quality; the CTE_{prompt} obtained from Multi-Modal CelebA dataset leads to age bias to the results, although it can be leveraged to generate semantically correct results.

Table 5. Qualitative Results of the Ablation Study on Prompt Design

	Zero CIE_{prompt}	CTE_{prompt}	Proposed
FID ↓	130.13	117.33	**114.19**
CLIP Distance ↓	0.7451	0.7382	**0.7295**

The best numbers are **bold**.

Fig. 13. The ablation study on the loss functions we propose. The results show that the $\mathcal{L}_{reg}$ assists to generate reasonable images and the $\mathcal{L}_{sem\_cons}$ keeps the semantics in the texts.

Table 6. Qualitative Results of the Ablation Study on Loss Design

	w/ L1 Loss	w/ L1 and Reg Loss	Proposed
FID ↓	211.13	**113.21**	114.19
CLIP Distance ↓	0.7811	0.7645	**0.7295**

The best numbers are **bold**.

Fig. 14. The failure cases. The successful attributes are in blue and failed attributes are in red. *[Best view in color]*

can be used to generate images with high fidelity and the reconstructed semantics consistency loss $\mathcal{L}_{sem\_cons}$ keeps the semantics contained in the input texts.

4.6 Limitations and Discussions

The proposed method is limited in two aspects. The first case is, if the expected image is out of the distribution of pre-trained StyleGAN, the image cannot be generated. The second case is, if there are several persons which are described in one input sentence, the framework will be confused and synthesize an face which has attributes from different persons. Such limited cases are shown in Figure 14. The reasons to these two limitations are relatively clear. The reason to the first failure case is that some images with rare semantics are difficult to be handled due to the limitation of StyleGAN itself. To achieve better performance, we can consider leveraging better open-world image generators. The reason to the second failure case is that CLIP is also has limited. There have been several works [5, 31, 36] proving the CLIP Text Encoder performs not satisfying enough on multi-object decomposition. This disadvantage of CLIP leads to confusing representations, making the model generate hybrid images.

5 Conclusions

In this work, we propose a unified framework for both face image generation and manipulation. The proposed framework, taking advantage of large-scale models including CLIP and StyleGAN, has the ability to handle various text inputs with complicated semantics and generate new images with impressively high semantic alignment, quality, and fidelity. The proposed method does not need any external training data. To demonstrate the efficiency, we conduct sufficient experiments and get superior objective and subjective metrics comparing with other state-of-the-art methods.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition.*

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition.*

[3] Zijun Deng, Xiangteng He, and Yuxin Peng. 2023. LFR-GAN: Local feature refinement based generative adversarial network for text-to-image generation. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 6 (2023), 1–18.

[4] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. 2017. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[5] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *Proceedings of the International Conference on Learning Representations*.

[6] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics* 41, 4 (2022), 1–13.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[9] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* 28, 11 (2019), 5464–5478.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

[12] Trang-Thi Ho, John Jethro Virtusio, Yung-Yao Chen, Chih-Ming Hsu, and Kai-Lung Hua. 2020. Sketch-guided deep portrait generation. *ACM Transactions on Multimedia Computing, Communications and Applications* 16, 3 (2020), 1–18.

[13] Tao Hu, Chengjiang Long, and Chunxia Xiao. 2021. A novel visual representation on text using diverse conditional GAN for visual recognition. *IEEE Transactions on Image Processing* 30 (2021), 3499–3512.

[14] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*.

[15] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

[18] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2019. Controllable text-to-image generation. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

[19] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. 2019. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[20] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55, 9 (2023), 1–35.

[21] Ruoyu Liu, Yao Zhao, Shikui Wei, Liang Zheng, and Yi Yang. 2019. Modality-invariant image-text embedding for image-sentence matching. *ACM Transactions on Multimedia Computing, Communications and Applications* 15, 1 (2019), 1–19.

[22] Shiguang Liu and Huixin Wang. 2023. Talking face generation via facial anatomy. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 3 (2023), 1–19.

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebfaces attributes (CelebA) dataset. *Retrieved August* 2018 (2018), 11.

[24] Yiyang Ma, Huan Yang, Bei Liu, Jianlong Fu, and Jiaying Liu. 2022. AI Illustrator: Translating raw descriptions into images by prompt-based cross-modal generation. In *Proceedings of the ACM International Conference on Multimedia*.

[25] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. 2023. Unified multi-modal latent diffusion for joint subject and text conditional image generation. arXiv:2303.09319. Retrieved from https://arxiv.org/abs/2303.09319

[26] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. arXiv:1411.1784. Retrieved from https://arxiv.org/abs/1411.1784

[27] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. 2021. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 7474–7489.

[28] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-driven manipulation of styleGAN imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[29] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. MirrorGAN: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125. Retrieved from https://arxiv.org/abs/2204.06125

[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*.

[33] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning*.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[35] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. 2023. MM-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

[37] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[38] Hongchen Tan, Xiuping Liu, Meng Liu, Baocai Yin, and Xin Li. 2020. KT-GAN: Knowledge-transfer generative adversarial network for text-to-image synthesis. *IEEE Transactions on Image Processing* 30 (2020), 1275–1290.

[39] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. 2022. DF-GAN: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

[41] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[42] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Towards open-world text-guided face image generation and manipulation. arXiv:2104.08910. Retrieved from https://arxiv.org/abs/2104.08910

[43] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[44] Liu Yang, Liping Jing, and Michael K. Ng. 2015. Robust and non-negative collective matrix factorization for text-to-image transfer learning. *IEEE Transactions on Image Processing* 12 (2015), 4701–4714.

[45] Yanhua Yang, Lei Wang, De Xie, Cheng Deng, and Dacheng Tao. 2021. Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis. *IEEE Transactions on Image Processing* 30 (2021), 2798–2809.

[46] Tao Yao, Yiru Li, Ying Li, Yingying Zhu, Gang Wang, and Jun Yue. 2023. Cross-modal semantically augmented network for image-text matching. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 4 (2023), 1–18.

[47] Zili Yi, Zhiqin Chen, Hao Cai, Wendong Mao, Minglun Gong, and Hao Zhang. 2020. BSD-GAN: Branched generative adversarial network for scale-disentangled representation learning and image synthesis. *IEEE Transactions on Image Processing* 29 (2020), 9073–9083.

[48] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. 2019. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

[49] Feifei Zhang, Mingliang Xu, and Changsheng Xu. 2022. Tell, imagine, and search: End-to-end learning for composing text and image to image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 2 (2022), 1–23.

[50] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. 2018. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2018), 1947–1962.

[51] Zizhao Zhang, Yuanpu Xie, and Lin Yang. 2018. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition.*

[52] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition.*

Received 27 December 2023; revised 24 May 2024; accepted 31 July 2024