Training-Free 3-D Face Avatars Generation by Knowledge Discovering in Foundational Models

Qilei Li, Mingze Xu, Wenzhe Zhai, David Camacho, *Senior Member, IEEE*, and Gwanggil Jeon

***Abstract*—An informative 3-D avatar, closely mirroring real-world traits, plays a pivotal role in accessing the metaverse. Traditional methods for creating 3-D avatars usually employ one-to-one training, which restricts avatar diversity. To enhance style diversity in generated 3-D avatars, we utilize synthesized images derived with prompts from ChatGPT in a conversational manner, ultimately resulting in a broader range of 3-D variations. Rather than creating models from scratch, we devise a training-free framework that utilizes established large-scale foundation models. Specifically, we employ a real-world image synthesis technique guided by text prompts that are generated by ChatGPT in a conversational manner, to describe the desired characteristics of the synthesized image. As a result, these informative latent representations can accurately reflect the distinct style of the synthesized image, and further lead to the creation of photorealistic and diverse 3D avatars. Our training-free design allows this proposed method to achieve competitive performance compared to existing generation models, while requiring minimal computational resources.**

***Index Terms*—3-D avatars generation, ChatGPT, image synthesis, metaverse, multimodal learning.**

I. Introduction

IN today's rapidly evolving digital landscape, computational social systems play an integral role in shaping virtual interactions and societal dynamics. As individuals increasingly engage in immersive environments like the metaverse, 3-D avatars serve as vital digital representations, facilitating communication, identity expression, and social interaction [1], [2]. It can provide immersive and interactive experiences for users who can explore, communicate, and create in the virtual space [3]. A 3-D avatar, which is a digital representation of a user, is the key to entering the virtual world. A 3-D avatar can convey rich information and features that are very similar to the real world, such as appearance, expression, gesture, and emotion [4]. There are two distinct gender-specific examples in Fig. 1, namely the real-world images (located on the left), and their corresponding 3-D metaverse avatars (positioned on the right). However, creating a realistic and diverse 3-D avatar is a difficult task. It requires a lot of data, computation, and expertise. In addition, current research has primarily focused on utilizing only one modality at a time. For instance, Yu et al. [5] propose a meta-learning-based adversarial training (MLAT) algorithm, which alternately optimizes adversarial samples and network parameters dynamically. Additionally, a meta-learning framework is introduced to train the deep 3-DFR model, contributing to enhancing the accuracy of the 3-DFR model. Avola et al. [6] reconstructed a 3-D model from a single 2-D image of a face. This approach employs a generative network to synthesize depth and correspondence maps from the input image, subsequently used to reconstruct a complete 3-D mesh. The initial raw model undergoes refinement and smoothing, with some fine details restored using the original texture, resulting in the final 3-D model. Taherkhani et al. [7] propose a novel framework aimed at nonlinearly mapping 3-D facial grids to two compact and disentangled identity and expression subspaces using a pair of supervised autoencoders and utilizing a conditional generative adversarial network trained on a high-resolution texture map to synthesize high-quality 3-D models with detailed textures and shapes. The majority of methods generate 3-D models from 2-D images, therefore the exploration of utilizing multiple conditions remains insufficient.

The 3-D avatars generation has been studied from various perspectives, such as metaverse, diffusion model, 3-D generative model, and large foundation models [8]. The metaverse is a concept of a shared virtual space that connects multiple platforms and applications. It can enable users to create and customize their own 3-D avatars and interact with others in the virtual world [2]. The diffusion model is a probabilistic framework that can generate high-quality images by reversing the diffusion process [8]. It can be applied to generate 3-D avatars from 2-D images or sketches. The 3-D generative model

Received 27 October 2024; revised 9 January 2025 and 18 March 2025; accepted 21 March 2025. This work was supported in part by the H2020 TMA-MSCA-DN TUAI project "Towards an Understanding of Artificial Intelligence via a transparent, open and explainable perspective" (HORIZON-MSCA-2023-DN-01-01) under Grant 101168344, in part by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program, in part by the EMIF managed by Calouste Gulbenkian Foundation, in the project MuseAI, and in part by the Comunidad Autonoma de Madrid, CIRMA-CAM Project under Grant TEC-2024/COM-404. *(Corresponding author: Gwanggil Jeon.)*

Qilei Li is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China and also with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: q.li@qmul.ac.uk).

Mingze Xu and Wenzhe Zhai are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: sdut_mingzexu@outlook.com; wenzhezhai@outlook.com).

David Camacho is with the School of Computer Systems Engineering, Universidad Politecnica de Madrid, 28040 Madrid, Spain (e-mail: david.camacho@upm.es).

Gwanggil Jeon is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China, and also with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail: ggjeon@gmail.com).

Digital Object Identifier 10.1109/TCSS.2025.3554455

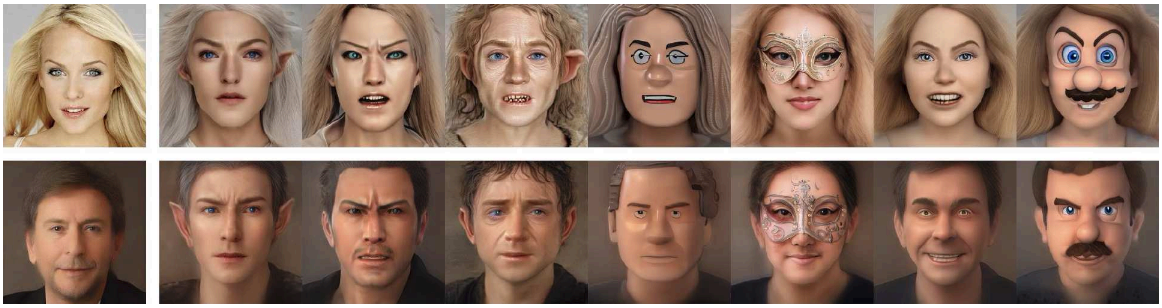


Fig. 1. Visualization of real-world person images (first column) and generated 3-D metaverse avatars in different styles (other columns).

is a neural network that can learn the distribution of 3-D shapes and generate novel ones [9]. Specifically, Kim et al. [10] proposed a domain adaptation-based 3-D generative framework with multimodal text-to-image diffusion models, which can be used to produce 3-D avatars with various expressions. The large foundation models are pretrained models that can learn from massive amounts of data and perform various tasks across domains. They can be leveraged to generate 3-D avatars from natural language descriptions. Current models for generating 3-D avatars are typically trained one-to-one, meaning they learn to map a single input image to a single output 3-D model. This limits the variety of avatars that can be produced, as they tend to inherit the style and appearance of the input image. Moreover, these models often rely on large-scale datasets of 3-D scans or images with annotations, which are costly and time-consuming to collect and process [11], [12]. In addition, Radford et al. [11] proposed a method that can learn from both images and natural language descriptions, and perform various tasks that require both visual and linguistic understanding. Rai et al. [13] proposed AlbedoGAN, which achieves high-quality morphing and accurate 3-D shape generation by integrating 2-D facial generation models with semantic facial manipulation. Furthermore, substantial advancements have been achieved in 3-D model generation through these methodologies. Nevertheless, directly incorporating real images into these networks elevates the vulnerability to privacy breaches. As a solution, we introduce the TFAG, aimed at safeguarding user privacy while upholding the fidelity of 3-D model generation.

To introduce diversity in the generated 3-D avatar styles, we use composite images derived from the initial landmark images as a starting point and subsequently generate different 3-D versions from these composite images. Instead of creating a model from scratch, we designed a framework that required no training, leveraging an already established large base model. Specifically, we employ a real-world image synthesis technique guided by text prompts describing the features needed to synthesize the image. To ensure that the prompts are rich and varied, we utilize ChatGPT to generate many prompts. Subsequently, we employ a collaborative diffusion model to generate images. This model comprises pretrained unimodal diffusion models that cooperate to achieve multimodal face generation and editing without the requirement for retraining. In addition, we utilize CLIP [14], a pretrained visual language model, for cross-modal regularization of potential representations to ensure they carry basic semantic information from aligned real-world image-text pairs. Furthermore, by comparing the high-dimensional semantic features of the generated images with those of synthetic images, we optimize the latent vectors through unimodality alignment. The combination of cross-modality and unimodality alignment in multimodality alignment further enhances the optimization of random noise, and thus leads to the generation of highly semantically consistent 3-D avatars. As a result, these information-rich potential representations can reflect the unique style of the synthesized images and further lead to the creation of realistic and diverse 3-D avatars. At the same time, we not only address privacy concerns but also preserve individual attributes, all without the need for substantial additional data, and thus enable the generation of a wide variety of 3-D avatars. The main contributions of our article are as follows.

1) We recognize the limitation in current 3-D face generation models, particularly their restricted capacity for creating diverse avatars. To address this challenge, we propose a novel approach that involves synthesizing a wide range of real-world images that is accomplished by incorporating multimodal supervision into the learning process. This assists the generation of discriminative yet varied latent vectors. These vectors, in turn, facilitate the creation of a diverse set of 3-D avatars in the metaverse.
2) We introduce training-free avatars generation (TFAG) method to exploit the large-scale foundational models to establish a coherent vision-language alignment for the latent vector. The TFAG framework is designed to be training-free for leveraging ChatGPT and CLIP to enable the generated images to be both diverse and representative.
3) The proposed TFAG yields impressive performance in items of generative high-quality avatars. Notably, TFAG achieves this without the need for additional training data, which significantly expands the available options it provides the audience with more options to select preferred avatars to enter the metaverse.

The rest of this article is organized as follows: Section II reviews the related work on 3-D avatar generation and real-world image synthesis. Section III describes the proposed framework in detail. Section IV presents the experimental setup and results. Section V concludes this article.

II. Related Work

A. Metaverse

The metaverse is a term that describes a virtual world that is seamlessly integrated with the real world, where people can interact with each other and with digital content across various platforms and devices. The metaverse has attracted a lot of attention and investment in recent years, especially in the field of vision, which aims to create realistic and immersive visual experiences for users. One of the main challenges in vision research for the metaverse is to develop technologies that can generate, manipulate, and render high-quality 3-D content in real-time, such as avatars, scenes, objects, and animations. Some of the current approaches include using deep learning models to synthesize 3-D content from 2-D images or videos [2], using computer graphics techniques to model and animate 3-D content [15], and using volumetric capture systems to record and stream 3-D content [16]. Another challenge is to design and implement user interfaces and interaction methods that can support natural and intuitive communication and collaboration in the metaverse. Some of the current approaches include using augmented reality (AR) and virtual reality (VR) devices to display 3-D content and provide haptic feedback [17], using eye tracking and facial expression recognition to capture and convey user emotions [18], and using speech recognition and natural language processing to enable voice-based interaction. A third challenge is to address the ethical, social, and legal issues that may arise from the widespread adoption of the metaverse, such as privacy, security, identity, ownership, regulation, and governance. Some of the current approaches include using blockchain and cryptography to ensure data integrity and user autonomy [1], using federated learning and differential privacy to protect user data and privacy [1], and using participatory design and stakeholder engagement to promote social inclusion and diversity [19]. To ensure user privacy, we employ a multimodal diffusion (MMD) module alongside textual prompts to generate images akin to the input image, which are then utilized for 3-D generation. In summary, the metaverse is a visionary concept with many potential applications and benefits for various domains and industries. However, it also poses many technical and societal challenges that require interdisciplinary research and collaboration. Vision research plays a key role in advancing the development of the metaverse, but it also needs to consider the broader implications and impacts of its innovations.

B. Diffusion Model

Diffusion models [8], [20], [21] have recently emerged as a prominent approach for image synthesis [22], [23], [24], alongside generative adversarial networks (GANs) [25]. These models have demonstrated success in diverse domains, including video generation [26], [27], [28], [29], image restoration [30], [31], semantic segmentation [32], [33], [34], and natural language processing [35]. Within the diffusion-based framework, models are trained using score-matching objectives [36], [37] across various noise levels, with sampling achieved through iterative denoising techniques. Previous research efforts have concentrated on enhancing the performance and efficiency of diffusion models via improved architectural designs [38], [39] and sampling schemes [40]. Multimodal synthesis and editing refers to the task of generating or manipulating images conditioned on multiple modalities of input, such as text, audio, sketches, and other images. This task is challenging because it requires capturing the semantic correspondence between different modalities and preserving the output's consistency and diversity. Existing diffusion models mainly focus on unimodal control, i.e, the diffusion process is driven by only one modality of condition. However, unimodal control may limit the expressiveness and flexibility of the generative process, especially when the input modality is sparse or ambiguous. For example, text-to-image synthesis may suffer from insufficient details or inaccurate alignment when the text description is vague or incomplete. Similarly, image-to-image translation may produce unrealistic or distorted results when the source image is noisy or occluded. In contrast, this study centers on leveraging existing models and presenting a concise framework for multimodal synthesis and editing without requiring extensive model retraining.

C. 3-D Generative Model

Recent advancements in 3-D generative models have led to achieving multiview consistent and explicitly pose-controlled image synthesis. While various types of 3-D generative models have been proposed, they have often suffered from issues such as low image quality, view inconsistency, and inefficiency. Notably, EG3D [41], utilizing a tri-plane hybrid representation and conditioned dual discrimination on poses, can generate images in real-time with exceptional quality, view consistency, and 3-D shapes. These 3-D generative models can be trained using single-view images and can subsequently produce infinite 3-D images in real-time. In contrast, 3-D scene representation using neural implicit fields, exemplified by neural radiance field (NeRF) [42] necessitates multiview images and substantial training time for each scene. Training contemporary 3-D generative models is notably more challenging than their 2-D counterparts. This is due to the requirement for a substantial volume of images and the knowledge of the camera parameter distribution associated with those images. To extend the applicability of state-of-the-art 3-D generative models across broader domains, we propose a text-guided domain adaptation method without additional images from the target domain. We design a novel pipeline to enable fine-tuning of EG3D [41], a state-of-the-art 3-D generator, to facilitate the synthesis of high-resolution, multiview, and consistent images in text-guided targeted domains.

D. Large Foundation Models

In recent years, the fields of computer vision and natural language processing have witnessed remarkable advancements thanks to groundbreaking models like vision transformer (ViT)

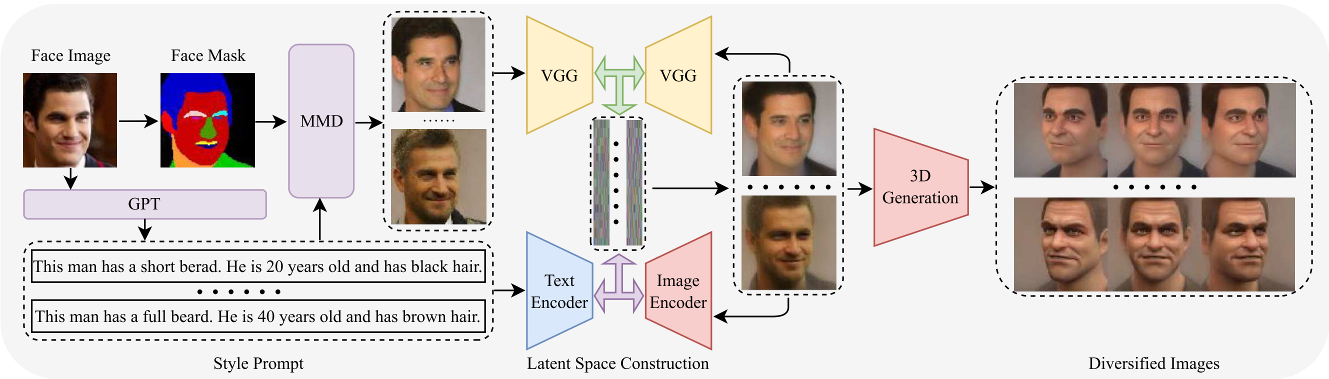

Face Image
Face Mask
MMD
GPT
This man has a short berad. He is 20 years old and has black hair.
This man has a full beard. He is 40 years old and has brown hair.
Style Prompt
VGG
VGG
Text Encoder
Image Encoder
Latent Space Construction
3D Generation
Diversified Images

Fig. 2. Framework of the proposed TFAG. It is designed to generate diverse 3-D avatars while addressing privacy concerns and preserving identity attributes in the metaverse.

[43], contrastive language-image pretraining (CLIP) [11], bootstrapping language-image pretraining (BLIP) [44], and BLIP-2. These models have fundamentally transformed our ability to understand and work with visual and textual data. In particular, ViT [43] is a model that applies the transformer architecture to image recognition tasks. ViT divides an input image into patches and treats them as tokens for the Transformer encoder. ViT achieves competitive results on several image classification benchmarks but requires a large amount of labeled data for pretraining. Contrastive language-image pretraining (CLIP) [11] has emerged as a powerful tool for bridging the gap between text and images. By training on an extensive dataset of images and text, CLIP has unlocked the potential for tasks like image-text matching. It can understand images and their associated descriptions, enabling it to perform tasks like finding matching images for given textual queries. Bootstrapping language-image pretraining (BLIP) [44] is a model that leverages existing pretrained vision models and large language models (LLMs) for vision-language pretraining. BLIP bridges the modality gap with a lightweight querying transformer, which is pretrained in two stages. The first step used masked language modeling on text-image pairs. The second step built contrastive learning on image-text matching. BLIP-2 [12] is an extension of BLIP that improves its efficiency and generality. BLIP-2 uses frozen pre-trained vision models and LLMs as feature extractors and only trains the querying transformer end-to-end. BLIP-2 introduces a data augmentation technique called caption filtering, which filters out noisy or irrelevant captions from the pretraining data. BLIP and BLIP-2 further push the boundaries of cross-modal understanding by leveraging even larger datasets. These models excel at fine-grained image classification, visual question answering, and other tasks that demand a deep understanding of the relationships between images and text. In a world increasingly driven by visual and textual data, ViT, CLIP, BLIP, and BLIP-2 are at the forefront, opening new possibilities for research, applications, and creative exploration at the intersection of vision and language. This article employs large foundational models such as diffusion and ChatGPT, aiming to associate visual and linguistic modalities for high-quality model generation while simultaneously protecting user privacy.

III. Methodology

A. Problem Definition

To produce 3-D avatars that vividly present one's identity in the metaverse, current models generally train a single mapping from a real-world face image. This mapping aims to learn

$$f_{\theta(x_i)} \rightarrow y_i \tag{1}$$

where x_i denotes the input person image and y_i represents the resulting 3-D image portraying the same person, and θ is the learnable parameter of the mapping function. However, this mapping function typically operates in a one-to-one manner, and thus means one real-world image can produce only one corresponding 3-D image. This inherent limitation arises due to the constrained diversity of generated 3-D images. One straightforward solution is to perform data augmentation on the input image, such as color jitter, random erasure, and random cropping. This transforms the mapping into

$$f_{\theta(x_i+\triangle)} \rightarrow y_i \tag{2}$$

where $\triangle$ represents the augmentation applied to the input image. Nevertheless, it is essential to acknowledge that these augmentation methods unavoidably result in information loss in the input image, and thus potentially degrade the quality of the generated counterpart. Alternatively, another plausible approach is to upload more real-world images to generate a broader range of 3-D avatars in different styles. However, contemporary privacy concerns have led to hesitations in sharing additional real-world images.

B. Framework Overview

To achieve the generation of diverse 3-D avatars representing various identities in the metaverse, while mitigating privacy concerns and preserving the appreciable attributes of each identity, we design a training-free framework termed as training-free avatars generation (TFAG). This framework capitalizes on the capabilities of large-scale pretrained foundational models to simulate real-world images related to various identity attributes, and thus generate the corresponding diversified 3-D outputs for the input face image. The proposed TFAG framework, as depicted in Fig. 2, achieves its goals through a structured sequence

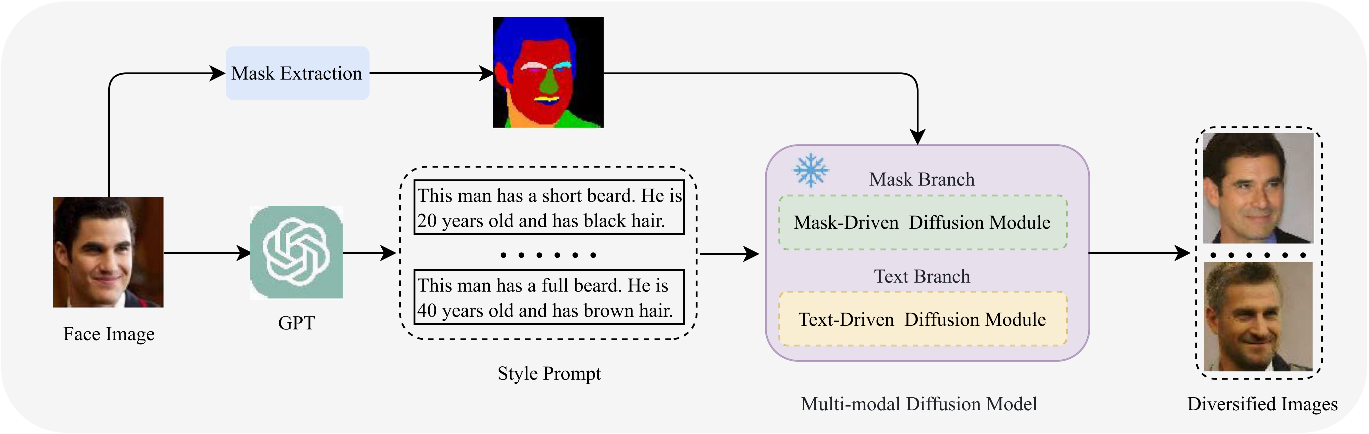

Mask Extraction
Face Image
GPT
This man has a short beard. He is 20 years old and has black hair.
• • • • • •
This man has a full beard. He is 40 years old and has brown hair.
Style Prompt
Mask Branch
Mask-Driven Diffusion Module
Text Branch
Text-Driven Diffusion Module
Multi-modal Diffusion Model
Diversified Images

Fig. 3. GPT-driven face style diversification.

of four key steps. The proposed TFAG framework involves the following steps.

1) Employ ChatGPT to generate a wide range of textual descriptions that describe an identity from various facets. This process leads to the creation of diverse text prompts, which diversify the input image by guiding it along varied yet meaningful directions.
2) Synthesize the input face image based on the face mask and the diverse text prompt, to create a few of the samples that represent the attribute of the identity.
3) Learn the corresponding latent vector with the text prompt supervised by the multimodality regularization.
4) Generate the 3-D avatar from the latent vectors using a pretrained 3-D generation model.

C. Text Prompt Generation With ChatGPT

When given a real-world face image, we initially extract the facial mask, which serves as the whiteboard for our subsequent operations. The whiteboard is pivotal for generating diverse synthetic face images, guided by a set of varied text descriptions, referred to as prompts which act as priors. To ensure the diversity of the generated text, we harness the capabilities of ChatGPT, a versatile language model that can generate prompts to describe the person differently. As illustrated in Fig. 3, we make use of ChatGPT 3.5 and ask it to provide us with a set of descriptions regarding the attributes of a person. We formulate the query (q) and the ChatGPT will reply to us with a list of descriptions. Specifically, we have devised clear prompts for ChatGPT, which enables it to provide diverse descriptions of facial attributes based on our instructions. We mathematically denote the interaction with ChatGPT as $h_{\text{gpt}}(\cdot)$ and this process is denoted as

$$\mathcal{P} = h_{\text{gpt}}(x_i, q) \tag{3}$$

where $\mathcal{P} = \{p_j\}_{j=0}^{N}$ is the set of replies from ChatGPT that describe the identity that will be used as prompts to generate new and q is the query to get the person attribute. We summarized the detailed prompt-generation process in Algorithm 1.

Algorithm 1: Description Generation Algorithm.

```
Require: Text Prompt Generation
Input: Men's Description:
 - Hair color: brown, black, silver-gray, white, gold, green, purple, blue, red, pink.
 - Beard type: full, sideburns, small, big, or short beard.
 - Hair type: straight or curly.
 - Eyelash length: long or short.
 Output:
 - Generated non-repetitive descriptions for men, e.g., "This man has a medium-length beard, brown hair, and straight hair."
Input: Women's Description:
 - Hair color: brown, black, silver-gray, white, gold, green, purple, blue, red, pink.
 - Hair type: straight, curly, or wavy.
 - Eyelash length: long or short.
 Output:
 - Generated non-repetitive descriptions for women, e.g., "The woman has long black hair and curly hair."
```

D. Diverse Synthetic Images Generation

To obtain various counterparts that describe a person in different styles, instead of relying on multiple images captured under various conditions, we synthesize additional images by employing unique text prompts, which serve as constraints during image generation. To accomplish this goal, we leverage the pretrained multimodal diffusion (MMD) model by freezing the learnable weight. We first extract the mask of the landmark image to indicate the appearance attribute of the identities. The face mask m_i and the corresponding text prompt will be used pair-wise and sent into the MMD to generate realistic-like person images with distinct styles. Assuming the function of MMD model is $h_{\text{mmd}}(\cdot)$, we formulate this process as follows:

$$\begin{aligned} m_i &= h_{\text{mask}}(x_i), \\ x_i^j &= h_{\text{mmd}}(p_i^j, x_i), \quad \text{for} \quad p_i^j \in \mathcal{P}_i \end{aligned} \tag{4}$$

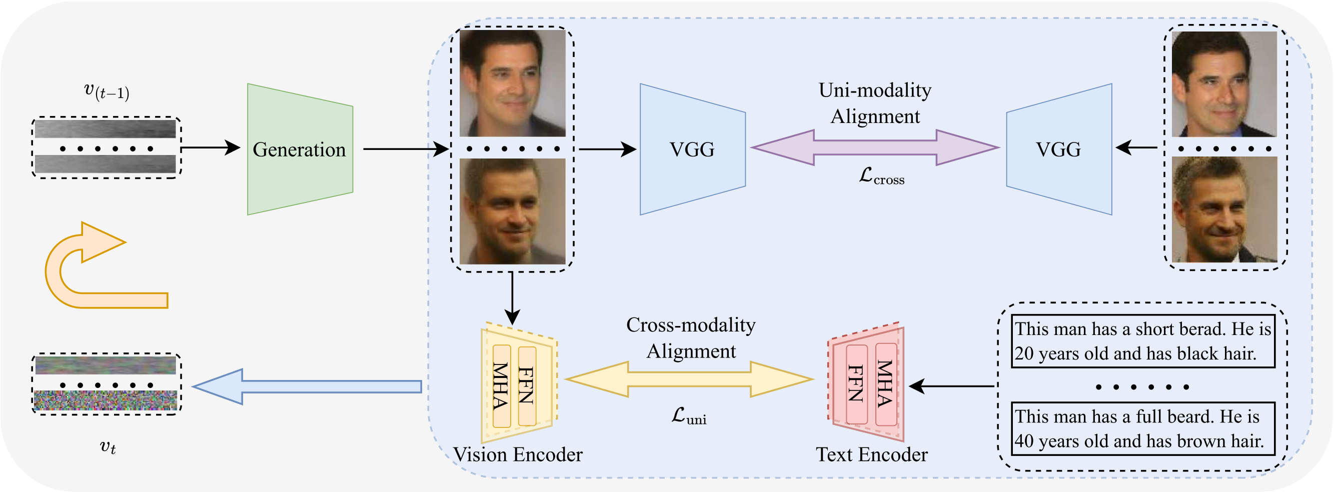

$v_{(t-1)}$
Generation
VGG
Uni-modality Alignment
$\mathcal{L}_{\text{cross}}$
VGG
MHA
FFN
Cross-modality Alignment
$\mathcal{L}_{\text{uni}}$
FFN
MHA
Vision Encoder
Text Encoder
This man has a short berad. He is 20 years old and has black hair.
This man has a full beard. He is 40 years old and has brown hair.
v_t

Fig. 4. Latent vector generation through multimodal alignment.

where $h_{\text{mask}}(\cdot)$ is the function for mask extraction. Therefore, we can have diverse synthetic image $\mathcal{X}_i = \{x_i^j\}_{j=1}^N$. Specifically, to enhance the diversity of image generation, we employ a dynamic diffuser consisting of a textual modality diffusion model and a mask modality diffusion model. These two diffusion models collaborate throughout each denoising iteration to facilitate multimodal guided synthesis. The dynamic diffuser assesses individual collaborators' input magnitude by forecasting spatial–temporal influence functions. The resulting impact of each pretrained diffusion model denoted as I_m, which is dynamically ascertained through the adaptive modulation of a dynamic diffuser D_m. The regularized variable I_{mmd} is formulated as

$$\begin{aligned} I_m &= D_m(X_t, t, c_m) \\ \hat{I}_m &= \frac{\exp(I_m)}{\sum_{j=1}^{M} \exp(I_j)} \end{aligned} \tag{5}$$

where m denotes the number of diffusion models. c_m is the condition of the diffusion model. X_t is the noisy image at time t. We utilize the acquired influence functions to manage the contribution of each pre-trained diffusion model at every denoising stage

$$\begin{aligned} \epsilon_{\text{pred}} &= \hat{I}_{m,\text{text}} \odot \boldsymbol{h}_{\text{mmd,text}}(x_t, t, c_m) \\ &\quad + \hat{I}_{m,\text{mask}} \odot \boldsymbol{h}_{\text{mmd,mask}}(x_t, t, c_m) \end{aligned} \tag{6}$$

where $\boldsymbol{h}_{\text{mmd,text}}$ is the diffusion model of text and the $\odot$ is a pixel-wise multiplication function. $\boldsymbol{h}_{\text{mmd,mask}}$ is the diffusion model of mask. Therefore, the diffusion model flexibly integrates models with varying weights, architectures, and modalities.

E. Latent Vector Optimization With Cross-Modality Regularization

To generate a 3-D counterpart from its 2-D input, we optimize a hidden latent space that serves as a map of the 2-D image. We commence by sampling random noise as z and then iteratively optimize it to refine its representational capability. A latent vector sampled from the latent space refers to a compact, lower-dimensional representation of data that captures essential features or patterns. In diffusion models, which are generative models used for density estimation and image generation, the latent vector represents a point in the latent space from which samples are generated. The process of latent vector generation through multimodal alignment is illustrated in Fig. 4. For simplicity in notation, we denote v_i^j as the latent vector corresponding to the image x_i^j and omit the index i in v_i^j for the remainder of the section. We initialize v^j as a random noise such that $v_{(0)}^j = z$. In optimizing the latent vector, we propose cross-modality regularization to ensure that the text prompt contributes to enhancing the semantic information within the latent vector. We employ the pretrained vision and text encoders from the CLIP model to align embeddings from different modalities. Specifically, we transform the latent vector at the $t-1$ iteration, denoted as $v_{(t-1)}^j$, into a person image using a generative model $h_{\text{g}}(\cdot)$. The cross-modality regularization is formulated as follows:

$$\begin{aligned} &I_{\text{(t-1)}}^j = h_{\text{vis}}(h_{\text{g}}(v_{\text{(t-1)}}^j)), \quad T_{\text{(t-1)}}^j = h_{\text{tex}}(p^j)), \\ &\mathcal{L}_{\text{cross}} = \frac{1}{N}\sum_{j=0}^{N} \|I_{\text{(t-1)}}^j - T_{\text{(t-1)}}^j\| \end{aligned} \tag{7}$$

where $I_{(t-1)}^j$ represents the vision embedding extracted by the vision encoder $h_{\text{vis}}(\cdot)$. p^j denotes the corresponding text prompt. $T_{(t-1)}^j$ denotes the text embedding extracted by the text encoder $h_{\text{tex}}(\cdot)$. Additionally, the optimization of the latent vector is also supervised by unimodality alignment, as expressed in the following equation:

$$\mathcal{L}_{\text{uni}} = \frac{1}{N}\sum_{j=0}^{N} \|h_{\text{vgg}}(h_{\text{g}}(v_{\text{(t-1)}}^j)) - h_{\text{vgg}}(x^j)\| \tag{8}$$

where $h_{\text{vgg}}(\cdot)$ denotes the mapping of input images to high-dimensional semantic features. The process aims to align the features in the latent vector with those in the original input image. By comparing the semantic features of the generated image with those of the real image, we can quantify their similarity. Consequently, the combination of unimodality alignment and cross-modality alignment in multimodality alignment helps optimize random noise, and thus enables the subsequent generation of highly semantically consistent 3-D avatars. It is worthwhile to highlight that both the unimodality and cross-modality learning objectives are utilized to optimize the latent vector v, while all the modules remain frozen in accordance with the training-free design.

F. 3-D Face Avatar Generation

Once we obtain the latent vectors v_i^j mapped from the original diversified real-world image set $\mathcal{X}_i$, we can proceed to generate the 3-D avatars leveraging a 3-D generation framework, i.e. DATID [10], in which an EG3D [41] model is adopted as the backbone. Assuming the 3-D generation model is represented by the function $h_{3\text{d}}(\cdot)$, we can generate the 3-D avatar as follows:

$$\begin{aligned} y_i^j &= h_{3\text{d}}(v_i^j) \\ \mathcal{Y}_i &= \{y_i^j\}_{j=1}^N \end{aligned} \tag{9}$$

where y_i^j represents the 3-D avatar corresponding to the latent vector v_i^j. $\mathcal{Y}_i$ is the collection composed of these 3-D avatars, which represents the diversity of the final generated 3-D images. These 3-D avatars serve as visual representations of individuals in the metaverse, which offer diversity and realism and enable users to express their unique identities and characteristics in virtual environments. Through the process of 3-D avatar generation, we can create a diverse and realistic collection of avatars that faithfully represent various individual identities in the metaverse, all without the need for substantial additional data or concerns about privacy issues.

IV. Experiment

A. Implementation Details

We utilized the collaborative diffusion model [45] as the multimodal generation model to synthesize real-world images. The mask was extracted by referring to [46]. The dimension of the random noise I_0 in (7) was set to 512, which was then updated for 300 iterations using the Adam optimizer. The learning rate was set to 0.1. Following DATID [10], an ImageNet pre-trained VGG16 [47] was used to extract vision feature for unimodality alignment. We used the pretrained CLIP model with a `ViT Base` and the patch size is set to 32 for the cross-modality alignment.

B. Benchmark

In the experiments, we employed Celeba-HQ [48] dataset to evaluate the performance of the proposed methods. This dataset contains 30 000 high-resolution images of celebrity faces and is widely used in computer vision tasks, such as face recognition and image generation. It encompasses a diverse range of attributes, including variations in age, gender, ethnicity, hairstyle, and facial expressions. Such diversity ensures that our method is validated under varied conditions and supports the generation of avatars representative of different demographic groups.

C. Evaluation Metrics

We employed two image quality assessment metrics, termed as perceptual image quality evaluator (PIQE) [49] and naturalness image quality evaluator (NIQE) [50], [51], to conduct a comprehensive quality evaluation of the generated 3-D images. These metrics assist us in quantifying the perceptual quality and naturalness of the generated images and thus provide better insights into their visual realism and acceptability. For both metrics, the lower scores are indeed better. The PIQE metric focuses on the perceptual quality of images, which aims to simulate the perception of the human visual system. It relies on various image features such as contrast, sharpness, and color vividness to generate a score representing image quality. Higher PIQE scores are typically associated with better perceptual quality. It is formulated as follows:

$$\text{PIQE} = \frac{\left(\sum_{k=1}^{N_{SA}} D_{sk}\right) + C_1}{(N_{SA} + C_1)} \tag{10}$$

where N_{SA} represents the number of sub-regions the image divides. D_{sk} denotes the feature value of the kth subregion of the image. C_1 is a constant term introduced to prevent division by zero in the denominator.

In contrast, the NIQE metric emphasizes the naturalness of images. It evaluates the naturalness level of images by analyzing statistical features like in-band variability and naturalness ratio. Lower NIQE scores generally indicate that the images are closer to being perceived as naturally real. It is formulated as follows:

$$\text{NIQE} = \sqrt{\left((\nu_1 - \nu_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\nu_1 - \nu_2)\right)} \tag{11}$$

where ν_1 and ν_2 represent two statistical feature vectors used to describe the naturalness of an image. These vectors capture statistical information about the image. Similarly, Σ_1 and Σ_2 denote two covariance matrices, each describing statistical features between the two sets of patches. They reflect the relationships between different pixel intensities or features within each patch. These matrices are employed to measure the naturalness of the image.

D. Qualitative Visualization

We utilize 26 text prompts generated by ChatGPT to steer the creation of 3-D avatars. From this set, we randomly select three samples presented in the upper part of Figs. 6 to 5. In these visualizations, the left column displays real-world images, while the metaverse renditions in the other three columns exhibit three distinct styles: Pixar style, Lego style, and Super Mario style. In the line charts presented in Figs. 6 to 5, the horizontal axis represents 26 different character images derived from adjustments in features like hair color and eye position, while the vertical axis showcases the scores of these 27 images under two metrics. The zero point on the axis corresponds to the score of

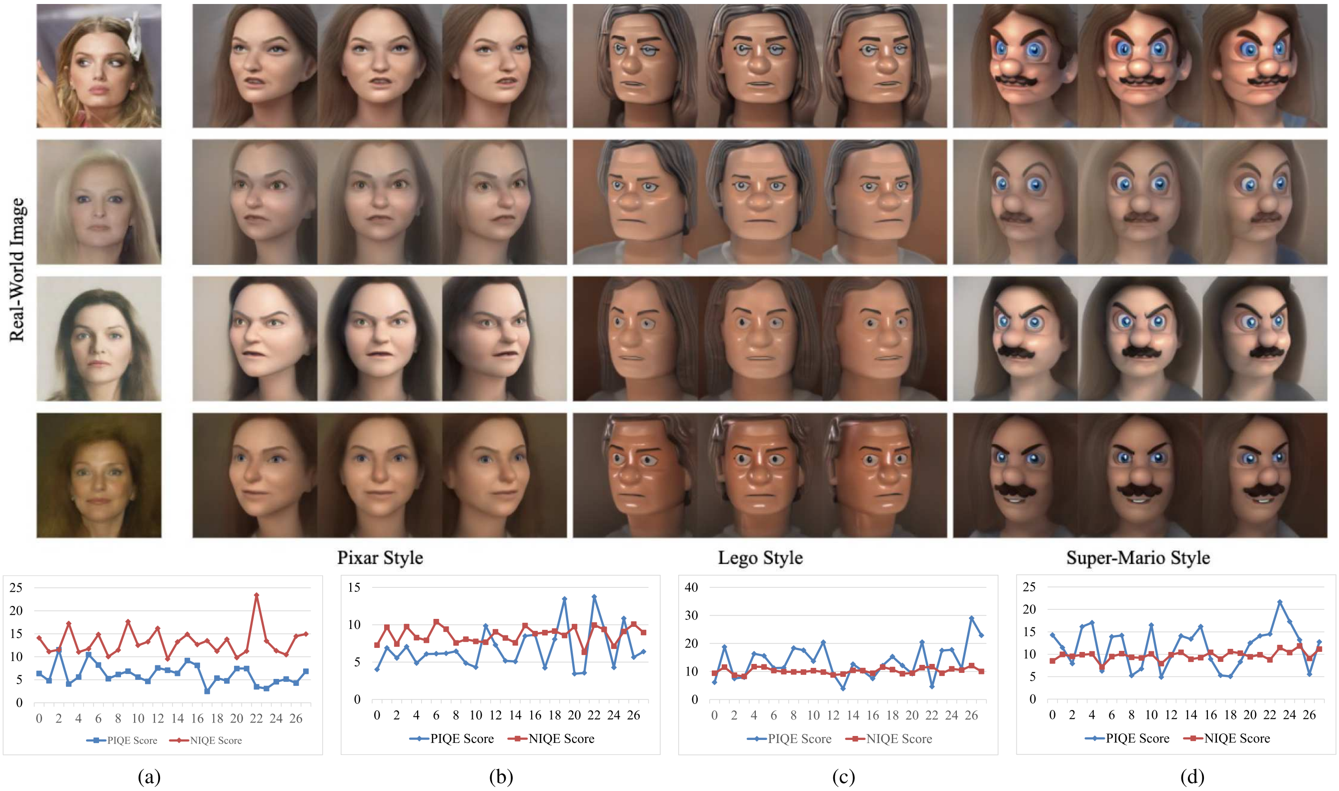


Fig. 5. Qualitative (upper part) and quantitative (lower part) comparisons for image 22. (a) Real-world. (b) Pixar style. (c) Lego style. (d) Super-Mario style.

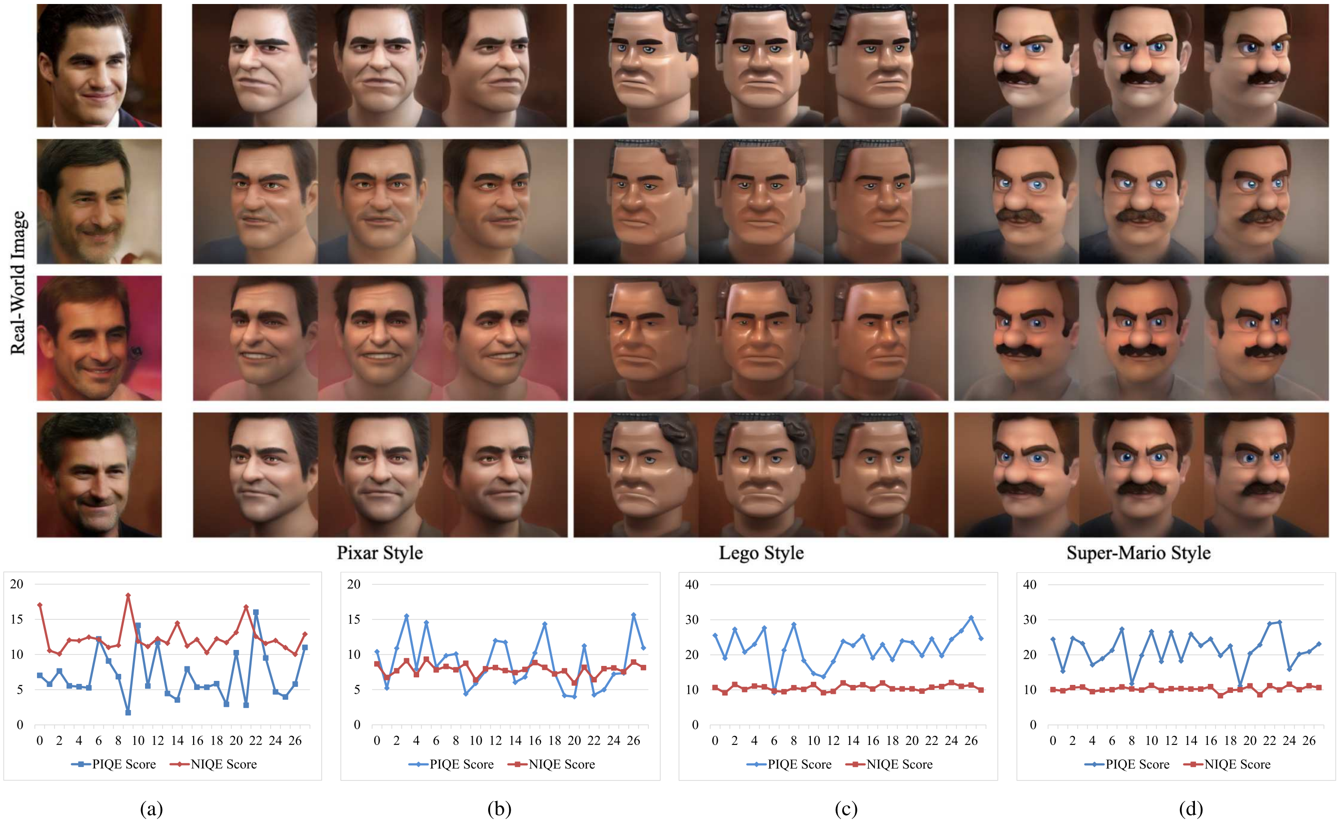
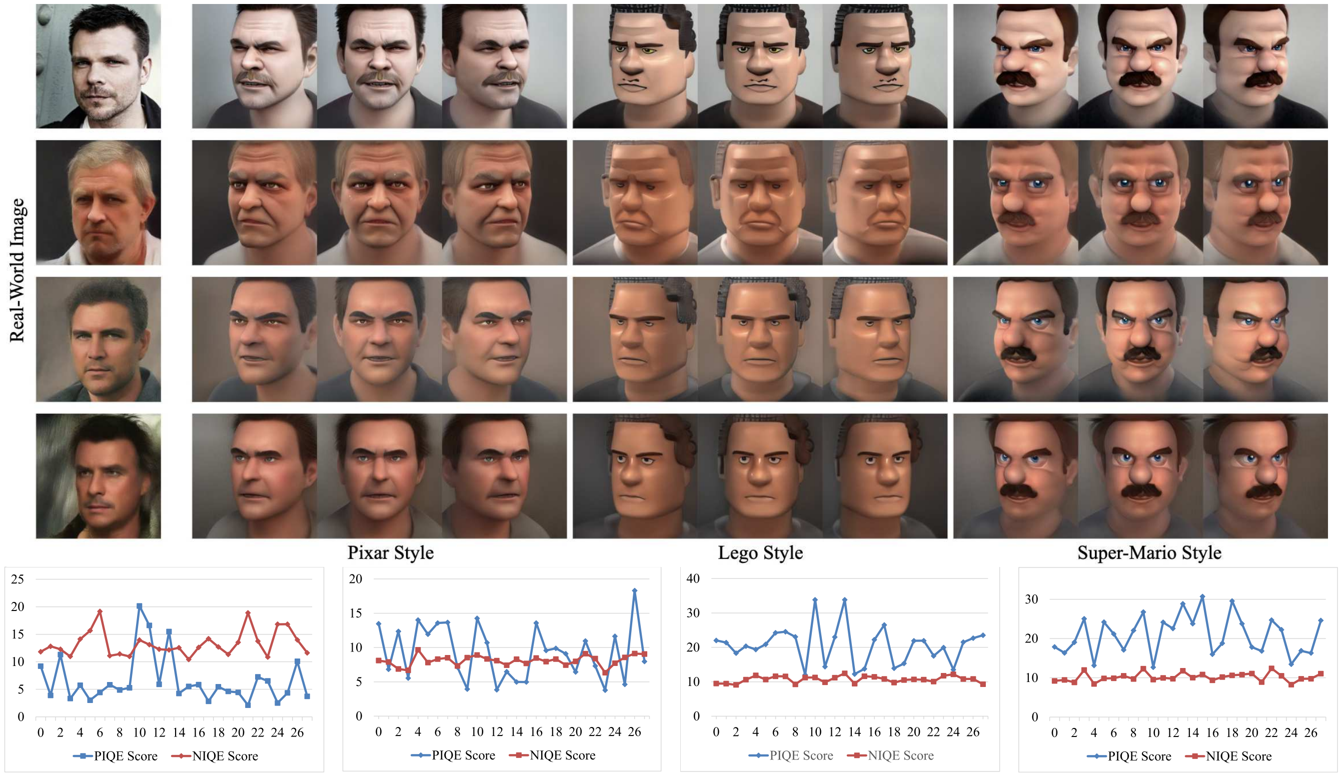


Fig. 6. Qualitative (upper part) and quantitative (lower part) comparisons for image 1123. (a) Real-world. (b) Pixar style. (c) Lego style. (d) Super-Mario style.



(a) (b) (c) (d)

Fig. 7. Qualitative (upper part) and quantitative (lower part) comparisons for image 1149. (a) Real-world. (b) Pixar style. (c) Lego style. (d) Super-Mario style.

the 3-D model generated from the authentic image under the two metrics, whereas each subsequent data point represents the score of the 3-D model obtained from 26 different images. It is clear from the chart that the scores of all data points are very close to those of the 3-D model from the real image, which indicates that the 3-D model generated by 26 different images guarantees a relatively smooth and informative visual output. When compared to the 3-D avatars generated from the original images without the use of any text prompts, our model showcases a wider range of styles, irrespective of age or hairstyle. Remarkably, it retains a wealth of detail from real-world images while accurately translating these distinguishing attributes into the metaverse. This accomplishment can be attributed to the additional guidance provided by the text prompts, which steer the generation of 3-D avatars. This guidance is achieved through cross-modality alignment, which aligns the image and text embeddings in a shared feature space. Consequently, it facilitates the creation of a meaningful inverse latent feature. As a result, we have effectively demonstrated the prowess of the proposed method in generating distinctive 3-D avatars.

E. Quantitative Performance Evaluation

The quantitative performance of these generated 3-D avatars in various styles is presented in the lower section of each figure. When examining the generated images in the real world, it is evident that these images closely resemble the original facial images, as evidenced in the index of 0. We observed that the averaged metrics remained consistently comparable for the 3-D avatars generated from diverse images, with some even exhibiting superior quality, denoted by lower NIQE and PIQE scores. In the image of Lego type with ID 22, the NIQE score might not reach a satisfactory level due to the highly block-like structure inherent to Lego constructions. This discrepancy arises as NIQE predominantly appraises the naturalness of images, posing a contradiction with the Lego aesthetic. Conversely, PIQE emphasizes holistic perceptual cues rather than localized natural coherence. Consequently, while NIQE performance may suffer in the Lego style, PIQE could potentially offer a more representative evaluation. In general, TFAG demonstrated superior results with lower PIQE and NIQE scores.

V. Conclusion

In this work, we propose a framework termed as training-free avatars generation (TFAG) to produce a wide range of 3-D avatars in response to textual prompts. Specifically, we employed ChatGPT to produce diverse and informative prompts, which were subsequently used to generate synthesis images. We utilized CLIP to extract essential semantic feature from aligned real-world image-text pairs. By doing so, these informative latent representations can effectively capture the unique style of the generated image and facilitate the generation of photo-realistic 3-D avatars. The proposed TFAG framework was designed in a training-free diagram, which can achieve competitive performance compared to existing generation models while maintaining minimal computational resource requirements. Extensive experiments and ablation studies demonstrated the superiority of the proposed method in 3-D avatar generation.

Data Availability Statement

The implementation is available at https://github.com/wenzhezhai/TFAG.

References

[1] B. Kye, N. Han, E. Kim, Y. Park, and S. Jo, “Educational applications of metaverse: Possibilities and limitations,” *J. Educ. Eval. Health Prof.*, vol. 18, 2021, Art. no. 32.

[2] H. Lin, S. Wan, W. Gan, J. Chen, and H.-C. Chao, “Metaverse in education: Vision, opportunities, and challenges,” in *Proc. IEEE Int. Conf. Big Data (Big Data)*. Piscataway, NJ, USA: IEEE Press, 2022, pp. 2857–2866.

[3] S. N. Gunkel, H. M. Stokking, M. J. Prins, N. van der Stap, F. B. T. Haar, and O. A. Niamut, “Virtual reality conferencing: Multi-user immersive VR experiences on the web,” in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 498–501.

[4] G. Freeman and D. Maloney, “Body, avatar, and me: The presentation and perception of self in social virtual reality,” *Proc. ACM Human-Comput. Interaction*, vol. 4, no. CSCW3, pp. 1–27, 2021.

[5] C. Yu, Z. Zhang, H. Li, J. Sun, and Z. Xu, “Meta-learning-based adversarial training for deep 3d face recognition on point clouds,” *Pattern Recognit.*, vol. 134, 2023, Art. no. 109065.

[6] D. Avola et al., “Facevision-GAN: A 3d model face reconstruction method from a single image using gans,” in *In Proc. 13th Int. Conf. Pattern Recognit. Appl. Methods*, 2024, pp. 628–632.

[7] F. Taherkhani et al., “Controllable 3d generative adversarial face model via disentangling shape and appearance,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vision*, 2023, pp. 826–836.

[8] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” 2020, *arXiv:2006.11239*.

[9] Y. Wu, Y. Deng, J. Yang, F. Wei, Q. Chen, and X. Tong, “Anifacegan: Animatable 3d-aware face image generation for video avatars,” *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 36188–36201, 2022.

[10] G. Kim and S. Y. Chun, “Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2023, pp. 14203–14213.

[11] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8748–8763.

[12] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” 2023, *arXiv:2301.12597*.

[13] A. Rai et al., “Towards realistic generative 3d face models,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vision*, 2024, pp. 3738–3748.

[14] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8748–8763.

[15] H.-S. Choi and S.-H. Kim, “A content service deployment plan for metaverse museum exhibitions—Centering on the combination of beacons and HMDS,” *Int. J. Inf. Manage.*, vol. 37, no. 1, pp. 1519–1527, 2017.

[16] J. Huggett, “Virtually real or really virtual: Towards a heritage metaverse,” *Stud. Digit. Heritage*, vol. 4, no. 1, pp. 1–15, 2020.

[17] M. Hu, X. Luo, J. Chen, Y. C. Lee, Y. Zhou, and D. Wu, “Virtual reality: A survey of enabling technologies and its applications in IOT,” *J. Netw. Comput. Appl.*, vol. 178, 2021, Art. no. 102970.

[18] M. R. Miller, F. Herrera, H. Jun, J. A. Landay, and J. N. Bailenson, “Personal identifiability of user tracking data during observation of 360-degree VR video,” *Sci. Rep.*, vol. 10, no. 1, 2020, Art. no. 17404.

[19] W. Gan, J. Chun-Wei, H.-C. Chao, S.-L. Wang, and S. Y. Philip, “Privacy preserving utility mining: A survey,” in *Proc. IEEE Int. Conf. Big Data (Big Data)*. Piscataway, NJ, USA: IEEE Press, 2018, pp. 2617–2626.

[20] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Proc. ICML*, 2015.

[21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *Proc. ICLR*, 2021.

[22] P. Dhariwal and A. Nichol, “Diffusion models beat GANs on image synthesis,” 2021, *arXiv:2105.05233*.

[23] P. Esser, R. Rombach, A. Blattmann, and B. Ommer, “ImageBART: Bidirectional context with multinomial diffusion for autoregressive image synthesis,” in *Proc. NeurIPS*, 2021.

[24] C. Meng et al., “SDEdit: Guided image synthesis and editing with stochastic differential equations,” in *Proc. ICLR*, 2022.

[25] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” *IEEE signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018.

[26] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood, “Flexible diffusion modeling of long videos,” 2022, *arXiv:2205.11495*.

[27] R. Villegas et al., “Phenaki: Variable length video generation from open domain textual description,” 2022, *arXiv:2210.02399*.

[28] U. Singer et al., “Make-a-video: Text-to-video generation without text-video data,” 2022, *arXiv:2209.14792*.

[29] J. Ho et al., “Imagen video: High definition video generation with diffusion models,” 2022, *arXiv:2210.02303*.

[30] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, “Image super-resolution via iterative refinement,” 2022, *arXiv:2104.07636*.

[31] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, “Cascaded diffusion models for high fidelity image generation,” 2022, *arXiv:*2106.15282.

[32] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, “Label-efficient semantic segmentation with diffusion models,” in *Proc. ICLR*, 2022.

[33] A. Graikos, N. Malkin, N. Jojic, and D. Samaras, “Diffusion models as plug-and-play priors,” in *NeurIPS*, 2022.

[34] T. Amit, E. Nachmani, T. Shaharbany, and L. Wolf, “SegDiff: Image segmentation with diffusion probabilistic models,” *arXiv preprint arXiv:2112.00390*, 2021.

[35] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg, “Structured denoising diffusion models in discrete state-spaces,” in *NeurIPS*, 2021.

[36] A. Hyvärinen and P. Dayan, “Estimation of non-normalized statistical models by score matching,” *JMLR*, 2005.

[37] P. Vincent, “A connection between score matching and denoising autoencoders,” *Neural Computation*, 2011.

[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022.

[39] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, “Vector quantized diffusion model for text-to-image synthesis,” in *CVPR*, 2022.

[40] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *ICLR*, 2021.

[41] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis et al., “Efficient geometry-aware 3d generative adversarial networks,” in *Proc. IEEE/CVF Conf. Comput. vision pattern Recognit.*, 2022, pp. 16123–16133.

[42] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” in *Eur. Conf. Comput. Vision*, 2020.

[43] K. Han et al., “A survey on vision transformer,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, 2022.

[44] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *Int. Conf. Mach. Learn.* PMLR, 2022, pp. 12888–12900.

[45] Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu, “Collaborative diffusion for multi-modal face generation and editing,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2023, pp. 6080–6090.

[46] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “Maskgan: Towards diverse and interactive facial image manipulation,” in *IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020.

[47] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.

[48] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANS for improved quality, stability, and variation,” in *Proc. ICLR*, 2018.

[49] F. Venkatanath, M. Praneeth, M. C. Bh., S. S. Channappayya, and S. S. Medasani, “Blind image quality evaluation using perception based features,” *Proc. 21st Nat. Conf. Commun. (NCC)*, pp. 1–6, 2015.

[50] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ‘completely blind’ image quality analyzer,” *IEEE Signal Process. Lett.*, vol. 20, pp. 209–212, 2013.

[51] K. Singla, R. Pandey, and U. Ghanekar, “A review on single image super resolution techniques using generative adversarial network,” *Optik*, vol. 266, no. 5, 2022, Art. no. 169607.