3D-Aware Text-driven Talking Avatar Generation

Xiuzhe Wu[1], Yang-Tian Sun[1], Handi Chen[1], Hang Zhou[2], Jingdong Wang[2], Zhengzhe Liu[3], and Xiaojuan Qi[1⋆]

[1]HKU [2]Baidu Inc. [3]CUHK

{xzwu,sunyt98,hdchen}@connect.hku.hk zhouhang@link.cuhk.edu.hk
wangjingdong@outlook.com zzliu@cse.cuhk.edu.hk xjqi@eee.hku.hk

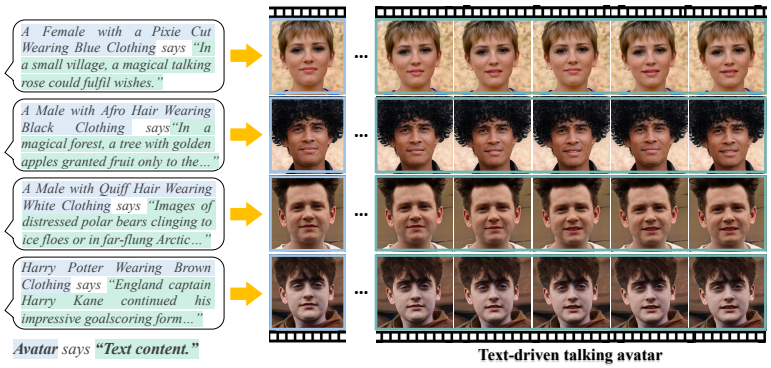

Fig. 1: Given a text description and text content, our model generates high-quality 3D facial avatars following the description and creates a high-fidelity video of the avatar speaking out the content.

Abstract. This paper introduces text-driven talking avatar generation, a task that uses text to instruct both the generation and animation of an avatar. One significant obstacle in this task is the absence of paired text and talking avatar data for model training, limiting data-driven methodologies. To this end, we present a zero-shot approach that adapts an existing 3D-aware image generation model, trained on a large-scale image dataset for high-quality avatar creation, to align with textual instructions and be animated to produce talking avatars, eliminating the need for paired text and talking avatar data. Our approach's core lies in the seamless integration of a 3D-aware image generation model (i.e., EG3D), the explicit 3DMM model, and a newly developed self-supervised inpainting technique, to create and animate the avatar and generate a temporal consistent talking video. Thorough evaluations demonstrate the effectiveness of our proposed approach in generating realistic avatars based on textual descriptions and empowering avatars to express user-specified text. Notably, our approach is highly controllable and can generate rich expressions and head poses.

Keywords: Text · Talking Avatar · Training Efficiency

⋆. corresponding author

1 Introduction

Controllable generation and animation of lifelike 3D facial avatars are crucial in various applications, such as digital human creation, video dubbing, and user-generated content (UGC) video creation. In this work, we explore a new task of text-guided talking avatar generation (see Figure 1): given a text description of the avatar and also the text to speak out, the task aims to create a talking avatar with appearance following the description and speaking the user-specified content. By this means, we can assist users in readily generating a desired avatar and letting it speak naturally, enabling diverse applications.

However, it remains challenging to create such talking avatars from pure texts based on existing methods. One major challenge is the absence of large-scale paired text and talking avatar data, which requires extensive crowd-sourcing and annotation. Fortunately, compared to paired text and video datasets, the image-level face datasets are easy to scale up (e.g., FFHQ [24]) and have facilitated the development of a wide array of research works [5,9,19,22,35,41,64,66], significantly advancing the area of photorealistic avatar creation. While the combination of text-to-2D image and 2D image animation method seems feasible, 2D image-based approaches often struggle with view inconsistency, especially for large head poses, due to limited 3D information. In contrast, EG3D [4] can synthesize high-quality avatars with rich details and 3D controllability. The above motivates us to enquire whether we can leverage such pre-trained 3D-aware image generation models for text-guided talking avatar generation.

In this work, we introduce a zero-shot approach to adapt a pre-trained 3D-aware avatar generation model (*i.e.* EG3D) for text-guided talking avatar generation, eliminating the need for paired text and video data for training. Our approach seamlessly integrates EG3D for pose-controllable high-fidelity appearance generation, CLIP to align text and rendered images for text-driven generation, and 3DMM for text-driven vivid expression generation in a talking video.

To be more specific, first, given a text description of the avatar, a high-fidelity static 3D-aware avatar is generated by optimizing a learnable code in the latent space of EG3D. This is achieved by minimizing the CLIP distance between the textual description and the rendered image in the front view. Next, we estimate a sequence of expressions that match the text content to speak out. With the expressions, we exploit the explicit parametric 3D representation, 3DMM, to animate the generated avatar with the desired facial expressions in 3D. This allows us to compute a 3D motion flow from the generated avatar to the desired avatar, which is then projected to 2D for warping the latent features to the target frame. To enhance avatar realism and mitigate warping artifacts, especially in dynamic facial areas, we introduce InpaintNet, which is trained in a self-supervised manner with self-reconstruction and consistency objectives to rectify the latent features and ensure temporal consistencies. Finally, with the refined latent features, the EG3D decoder yields high-quality video results with its high generative capabilities.

Through comprehensive evaluations, we showcase that our model can create high-quality talking avatars with facial details, rich facial expressions, nat-

Methods	Text-. Generation	Audio/Exp-. Animation	Text-. Animation	Train Time	Train Data
EG3D [4]	✗	✗	✗	8.5d, 8 V100	70k images
AnyFace [44]	✓	✗	✗	-	30k text-image
RODIN [50]	✓	✗	✗	-	100k images
OTAvatar [30]	✗	✓	✗	<2d, 4 A100	17k images
SADTalker [60]	✗	✓	✗	N/A, 8 A100	100k videos
Ours	✓	✓	✓	**<5h, 2 3090**	**No**

Table 1: Comparisons with existing works in other related tasks. Notably, the training time and training data employed for these models are sourced from published works.

ural movement, and good consistency with the user-specified text input in a self-supervised manner without any additional data, which cannot be readily achieved by any existing works as far as we know.

In summary, our primary contributions are:

1) We study a novel text-guided talking avatar generation task, which uses text to instruct both the generation and animation of a talking avatar.
2) We introduce a zero-shot approach for adapting a 3D-aware facial image generation model to this task without the need for paired training data. Our method is centered around the seamless integration of three components: EG3D for high-quality avatar generation, the CLIP model for aligning text and visual outputs, and the 3DMM model for animating the avatar with a wide range of expressions, ensuring precise expression generation and control.
3) We propose InpaintNet, a self-supervised learning model that utilizes consistency and reconstruction objectives to generate coherent talking videos effectively.
4) Comprehensive experiments demonstrate the effectiveness of our model in creating high-quality talking avatars according to text, which cannot be readily achieved by any existing works to the best of our knowledge.

2 Related Work

Since there is no existing work that can be readily applied to our task as far as we know, we compare our approach with SOTA methods in other related tasks, including 3D avatar generation and talking head video generation. Table 1 summarizes the representative-related works, and the details are given below. In summary, our method can work on a wider range of tasks compared to existing works without manually collecting the training data.

Avatar Generation and Animation A series of existing approaches [5, 19, 35, 41, 64, 66] attempt to create avatars leverage GANs [8, 18] and diffusion models [9, 22]. For instance, EG3D [4] can synthesize rich details and control the head pose; yet, it lacks animation capability such as expression control. To enable the avatar animation, existing works [2,13,15,16,34,39,48,56] leverage 3D Morphable

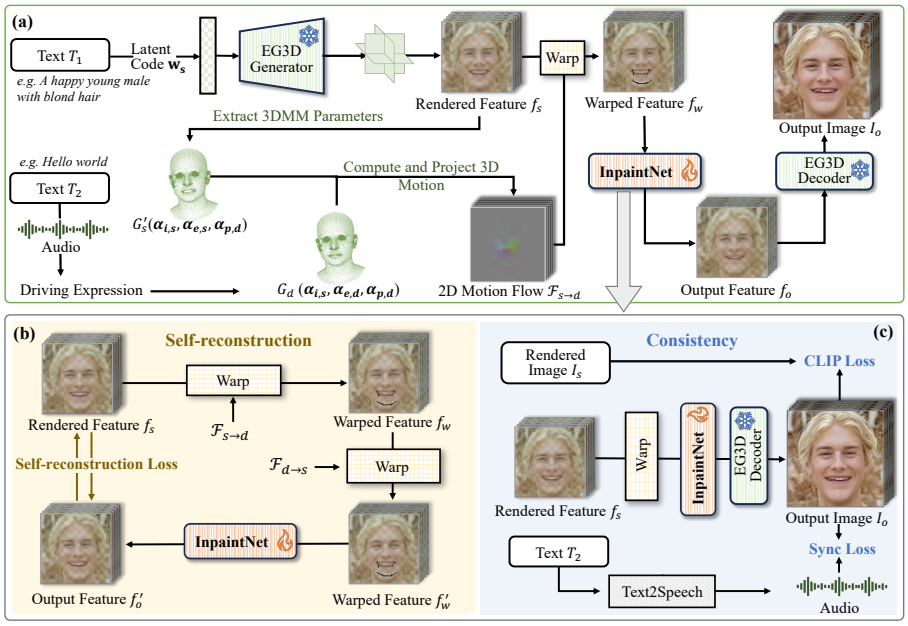

Fig. 2: (a) The overall framework of our approach. Leveraging the pre-trained EG3D model, we optimize a latent code $\mathbf{w_s}$ and create a 3D avatar satisfying the given text description T_1. A front-view image is subsequently rendered to derive the corresponding 3DMM model G_s with identity coefficient $\boldsymbol{\alpha}_{\mathbf{i,s}}$ and expression coefficient $\boldsymbol{\alpha}_{\mathbf{e,s}}$. For animation, we drive G_s with the expression parameters $\boldsymbol{\alpha}_{\mathbf{e,d}}$ extracted from the input text T_2. The 3D motion flow ($\mathbb{F}_{s\to d}$) between G_d and G_s is projected to the image plane ($\mathcal{F}_{s\to d}$) for the warping of rendered feature f_s, which is further enhanced with the InpaintNet and decoded to the output frame sequence. **(b)(c)** The training objectives of InpaintNet. To mitigate the warping artifacts, a double-warping strategy (using $\mathcal{F}_{s\to d}$ and $\mathcal{F}_{d\to s}$) is performed for the self-reconstruction loss. To enhance temporal consistency across frames and consistency between decoded and rendered images, Sync loss and CLIP loss are involved.

Models (3DMM) to offer robust control on expressions. However, these models typically cannot capture fine-grained facial details like EG3D [4]. In this work, we integrate the advantages of EG3D and 3DMM to create a high-quality facial appearance using EG3D and enable expression control with 3DMM for talking avatar generation.

Audio-driven Talking Head Video Generation Audio-driven talking avatar generation, aimed at talking avatar videos from audio, can be roughly divided into two primary categories: speaker-specific and speaker-independent methods. Early speaker-specific approaches typically require a long video of the target avatar for model training [25, 45], while recent ones leverage NeRF [1, 12, 32, 53] or 3D geometry and correspondence information [17, 26, 33, 46, 47, 55, 63] to build avatar-specific 3D models [20,27,28,40,54,57] to reduce the reliance on the train-

ing data. Speaker-independent methods [5, 6, 35, 64, 66] predominantly leverage GANs. Besides, some recent works develop other techniques including targeted facial region refinement [38], lip-sync technique [35], head pose integration [65], and leverage 2D facial landmarks and expression parameters [5, 10, 43, 49, 66] to improve the animation realism.

Video-driven Talking Head Video Geneation Video-driven talking avatar generation aims to transfer the facial motion from a source avatar to a target one. Various approaches have been proposed in this area, including landmark-based [23, 42, 51, 52, 62], 3D Morphable Model (3DMM) based [14, 37, 58], and latent animation approaches [31]. Recently, Ren et al. [37] further improves motion representation for enhanced facial reanimation, and Doukas et al. [14] enhances the model accuracy via 3DMM mesh fitting. Besides, Yin et al. [58] adopt StyleGAN for feature map manipulation. These approaches further improve the performance of the video-driven talking avatar generation. However, both audio- and video-based approaches require a data collection process, which can be tedious and laborious. In response to this limitation, our approach eliminates the need for data collection and allows users to leverage a piece of text to create a talking video, providing more convenient and adaptable interaction options.

3 Our Method

In this work, we study a new task for text-guided talking avatar generation. This task aims to produce a video sequence of an avatar that corresponds to a textual description T_1 while speaking out the content provided in another text, designated as T_2. To this end, we propose a zero-shot method that adapts a pre-trained 3D-aware facial avatar generation model to produce a text-driven talking avatar, eliminating the need for paired training data. As illustrated in Figure 2, our framework consists of three modules. (i) text-guided 3D avatar generation (Section 3.1). To do this, we optimize the avatar latent code w_s in the pre-trained EG3D latent space to create a 3D avatar whose 3DMM model is represented as $G_s = \{\boldsymbol{\alpha}_{\mathbf{i,s}}, \boldsymbol{\alpha}_{\mathbf{e,s}}, \boldsymbol{\alpha}_{\mathbf{p}}\}$ ($\boldsymbol{\alpha}_{\mathbf{i,s}}$: identity coefficient; $\boldsymbol{\alpha}_{\mathbf{e,s}}$: expression coefficient; and $\boldsymbol{\alpha}_{\mathbf{p}}$: head pose). Thanks to the 3D awareness of EG3D latent space, $\boldsymbol{\alpha}_{\mathbf{p}}$ can be an arbitrary head pose. To make it consistent with the text description T_1, we encourage rendered output image I_s to align with T_1 by using CLIP consistency regularization. (ii) text-guided avatar animation in 3D (Section 3.2). We leverage the explicit 3D representation of 3DMM for animation in 3D. Specifically, given G_s and a desired expression $\boldsymbol{\alpha}_{\mathbf{e,d}}$ from text T_2 and head pose $\boldsymbol{\alpha}_{\mathbf{p,d}}$, we obtain the 3DMM of the animated avatar $G_d = \{\boldsymbol{\alpha}_{\mathbf{i,s}}, \boldsymbol{\alpha}_{\mathbf{e,d}}, \boldsymbol{\alpha}_{\mathbf{p,d}}\}$ with expression aligned with the content of T_2. Note that we can easily obtain the source avatar $G'_s = \{\boldsymbol{\alpha}_{\mathbf{i,s}}, \boldsymbol{\alpha}_{\mathbf{e,s}}, \boldsymbol{\alpha}_{\mathbf{p,d}}\}$ at the target head pose $\boldsymbol{\alpha}_{\mathbf{p,d}}$. Then, given G'_s and G_d, we are able to compute the 3D motion flow $\mathbb{F}_{s\to d}$ that maps the surface points from the source G'_s to its corresponding location at target G_d. This 3D motion flow is further projected onto a 2D motion flow $\mathcal{F}_{s\to d}$ within the image plane. (iii) high fidelity talking avatar video frame generation in 2D

(Section 3.3). We maintain the identity and intricate details of the source avatar while creating a high-quality talking video. This is achieved by animating the source avatar to exhibit the desired expressions through warping its latent feature map f_s of EG3D in accordance to the motion flow $\mathcal{F}_{s\to d}$ to produce f_w and then leveraging the decoder of EG3D to synthesize a high-quality talking video. To alleviate the warping artifacts in f_w, we incorporate an InpaintNet to generate a complete feature map f_o for further decoding with EG3D decoder to produce high-quality results.

3.1 Text-Guided 3D-aware Avatar Generation

Given a text description T_1, we aim to create a 3D-aware facial avatar with 3DMM representation $G_s = \{\boldsymbol{\alpha}_{\mathbf{i,s}}, \boldsymbol{\alpha}_{\mathbf{e,s}}, \boldsymbol{\alpha}_{\mathbf{p}}\}$ following it. As shown in Figure 2 (a), we leverage a pre-trained avatar generator EG3D for high-fidelity avatar generation and derive a latent code $\mathbf{w_s}$ in its latent space Ω which can further generate an output image I_s with the EG3D decoder that is consistent with T_1.

Specifically, to generate a high-fidelity 3D facial avatar, we leverage the pre-trained EG3D 3D-aware avatar generator, which generates a latent triplane representation with three mutually perpendicular feature planes $\{f^0, f^1, f^2\}$. Then, we query the corresponding features of point p from each feature plane as $\{f_p^0, f_p^1, f_p^2\}$ and concatenate them to be a feature vector $f_p = \{f_p^0 \oplus f_p^1 \oplus f_p^2\}$. With the feature f_p, we adopt an MLP to predict the occupancy σ_p and color c_p of the point p. This representation can be used to effectively render a 2D image by grid-sampling to form a feature map $f \in \mathbb{R}^{H/r\times W/r\times C}$ from a camera pose and then decode it to an image $I_s \in \mathbb{R}^{H\times W\times 3}$ using the pre-trained decoder, where r indicates the downsampling rate in EG3D and C means the sum of feature dimensions of the three feature planes $\{f_p^0, f_p^1, f_p^2\}$.

Then, we leverage the feature spaces of the pre-trained vision-language model CLIP [36] to connect the avatar generated by EG3D [4] and the given text description without needing paired data. We optimize the latent code $\mathbf{w}_s$ in EG3D by encouraging the CLIP feature f_s^I of rendered images I_s from the triplane representation produced by $\mathbf{w}_s$ to be consistent with the CLIP text feature f^{T_1} of the input text T_1, such that the generated 3D avatar can follow the description of T_1. The objective function in the optimization process is then formulated as Equation (1).

$$\mathcal{L}_{\text{clip}} = \|f_s^I - f^{T_1}\|^2. \tag{1}$$

With $\mathbf{w}_s$ available, we can render image I_s at any given pose $\boldsymbol{\alpha}_{\mathbf{p}}$, allowing us to estimate the 3DMM parameters and obtain G_s. Figure 5 illustrates several visualization results of our text-guided avatar generation, demonstrating that our method can create high-fidelity 3D avatars that are consistent with the text inputs.

3.2 Text-guided Avatar Animation in 3D

With the high-quality 3D-aware avatar G_s, we further aim to drive the avatar to speak out the user-provided text content T_2 by controlling its 3DMM coefficients. As shown in Figure 2 (b), to drive the avatar in accordance with the input context text T_2, we utilize a publicly available text-to-speech tool to convert input text descriptions into audio. Afterward, a pre-trained audio-to-expression model [60] is further employed to generate the corresponding 3DMM coefficients for expression $\boldsymbol{\alpha}_{\mathbf{e,d}}$ and head pose $\boldsymbol{\alpha}_{\mathbf{p,d}}$, which are further utilized to drive the avatar to speak out T_2.

Fig. 3: Facial image generated by 3DMM coefficients. It fails in the detailed and complete generation.

An intuitive approach is to naively create a video sequence using the derived 3DMM coefficient sets. However, the generative fidelity of this simple baseline is far from satisfactory due to the limited capability of 3DMM for hair region representation, identity preservation, and facial detail generation (see Figure 3). To this end, we propose a warping-based animation approach to drive the high-quality source avatar G_s to make the desired expression under the control of 3DMM coefficients in 3D, which is further mapped into 2D and leverage the decoder of EG3D for generating high-quality results. This design helps to preserve the identity and facial details of the source avatar G_s while taking advantage of the pre-trained EG3D model, producing high-quality talking videos. Specifically, we create the 3D meshes G_s, G_d using the source and driving 3DMM expression coefficients $\boldsymbol{\alpha}_{\mathbf{e,s}}$ and $\boldsymbol{\alpha}_{\mathbf{e,d}}$, respectively. Note that G_s, G_d are created using the shared identity coefficient $\boldsymbol{\alpha}_{\mathbf{i,s}}$ that was utilized to create the source 3D avatar G_s in order to preserve the identity of the source avatar G_s. Also, they share the same pose coefficient $\boldsymbol{\alpha}_{\mathbf{p,d}}$ for better alignment. Then, we derive the 3D motion flow $\mathbb{F}_{s\to d}(p)$ of each query surface point p from G_s to G_d so that we can drive the source avatar to make the target expression through warping, as shown in Equation (2).

$$\mathbb{F}_{s\to d}(p) = \mathcal{T}(G_d, p) - p, \tag{2}$$

where $\mathcal{T}(\cdot,\cdot)$ denotes the operation that finds the 3D coordinates of p's corresponding point in G_d.

However, the meshes created by 3DMM do not include some crucial details, including colors and hair details; hence, the created talking video will lose details through the 3D warping. Thus, we further project the 3D motion flow $\mathbb{F}_{s\to d}(p)$ onto the image plane to derive the 2D motion flow $\mathcal{F}_{s\to d}$ and leverage EG3D for high-quality talking video generation through the 2D warping as detailed in the following section.

3.3 Talking Avatar Generation and Self-supervised InpaintNet

Given the 2D warp field $\mathcal{F}_{s\to d}$, we warp the latent EG3D feature map of the source avatar f_s to produce the latent EG3D feature map of the target avatar f_w. This will be further processed by the proposed InpaintNet to handle warp artifacts and produce f_o which is further fed to the EG3D decoder to produce the high-quality target frames for text-driven talking avatar generation. This design leverages the strong generative capability of the pre-trained EG3D model for high-quality avatar generation and is efficient. In the following, we will elaborate on the InpaintNet, which is trained in a self-supervised manner to hallucinate the incomplete feature map with warping artifacts and produce temporal consistent latent features f_o.

Due to the inability of the 3DMM model to capture details such as teeth, the warped feature map may exhibit empty holes in these areas. To mitigate these artifacts effectively, we introduce InpaintNet to produce a complete feature map f_o. InpaintNet is designed to denoise the warped feature map f_w and generate a high-quality feature map f_o suitable for decoding using the decoder of EG3D. Specifically, the InpaintNet takes five consecutive f_w as inputs and then outputs five f_o to enhance the temporal consistency. It employs a residual architecture to learn the residue between f_w and f_o. Then, f_o is processed effectively by the superresolution decoder in our pre-trained 3D facial avatar generation model to produce high-quality results. This enables us to leverage the refined expertise of the pre-trained decoder, which has been trained on large, high-fidelity datasets, ensuring detailed, high-quality outputs while maintaining 3D consistency.

3.4 Self-supervised Training Objectives

During training, we design a self-supervised training objective for training InpaintNet, including self-reconstruction and consistency loss to align synthesized images and text descriptions of the avatar T_1.

Self-reconstruction Loss We synthesize warping artifacts to create paired data simulating f_w and f_o using a double-warping strategy. As shown in Figure 2, this approach involves initially warping from the original pose space to a random pose space and then back again, resulting in a final warped feature map with artifacts. This feature is then fed into InpaintNet, allowing us to utilize the rendered feature f_s under the original pose to further produce the supervision signal. Specifically, we randomly sample a latent code $\mathbf{w_s}$ in the triplane feature space and then render a feature map $f_s \in \mathbb{R}^{H/r\times W/r\times C}$ with a randomly sampled pose. Further, we randomly sample a 3DMM expression coefficient $\boldsymbol{\alpha}_{\mathbf{e,d}}$ to compute the $\mathcal{F}_{s\to d}$ to warp f_s to be f_w to let the avatar make the associated expression. To create paired training data, we further warp f_w back to the original expression of f_s using $\mathcal{F}_{d\to s}$, where the double warped feature map is denoted as f'_w. By doing so, we can utilize f'_w to simulate the warped feature map in our framework and create the paired warped-clean dataset $\{f'_w, f_s\}$, which is utilized to train our InpaintNet in a self-supervised manner. Our InpaintNet takes f'_w as input

and outputs f'_o. For the training loss, the self-reconstruction loss comprises two components. The first one is a image reconstruction loss $\mathcal{L}_r$ to minimize the Euclidean distance between the decoded image I'_o of f'_o and the rendered ground truth image I_s of f_s: $\mathcal{L}_r = ||I'_o - I_s||_2$. The second component is the perceptual loss function, which works on the whole image $\mathcal{L}_p$, to enhance the overall image quality, employing the LPIPS metric [59]: $\mathcal{L}_p = \text{LPIPS}(I'_o, I_s)$.

Consistency Loss During training, we utilize text-based instructions for convenience and introduce two losses to enhance consistency. Initially, we input meaningful text, converting it into text-driven audio features a and further into expressions. Utilizing our method, frames are animated by these expressions, producing five decoded images I_o. Subsequently, since ground truth is unavailable after just one warping, we introduce two types of loss. Firstly, the sync loss $\mathcal{L}_s$ quantifies the discrepancy between text-driven audio a and decoded images I_o using the negative cosine similarity through a binary cross-entropy loss: $\mathcal{L}_s = \text{BCELoss}(-\text{cosine\_similarity}(a, I_o), 1)$. This enhances the temporal consistency of generated frames. Secondly, the CLIP loss $\mathcal{L}_c$, inspired by CLIP's capabilities, computes the L2 distance between the features of the decoded image I_o and the rendered image I_s, ensuring consistency in CLIP's feature space.

Overall Loss Functions The overall training objective function combines these loss components, weighted by trade-off hyperparameters λ_r, λ_p, λ_c, and λ_s as

$$\mathcal{L}_{\text{total}} = \lambda_r \mathcal{L}_r + \lambda_p \mathcal{L}_p + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s. \tag{3}$$

4 Experiments

4.1 Implementation Details

For our model's training, we set the total batch size to 16 and utilize the Adam optimizer to fine-tune the network. This optimizer is initialized with a learning rate of $1e^{-4}$ to train InpaintNet. We implement our framework with PyTorch with two NVIDIA 3090 GPUs. The network training takes less than five hours, demonstrating the efficiency of our approach. In our implementation, λ_r and λ_c are empirically set to 1 while λ_p is set to 0.01. Additionally, λ_s is further adjusted to 0.001. Besides, for the text-to-speech conversion, we utilize a specific software solution. For the audio-to-expression aspect, we employ a pre-trained model as described in [60]. Besides, to further eliminate artifacts resulting from warping operations, the rendered high-resolution images are also warped, resulting in the warped high-resolution image I_w based on the driving motion signal. I_w is then multiplied by a corresponding high-resolution mask and added to the image after InpaintNet correction I_i. The high-resolution mask is derived using mouth landmarks from a landmark detector [3]. Within the mouth region, I_i effectively eliminates artifacts, while outside the region, I_w provides reliable and high-quality results.

Fig. 4: Qualitative comparisons with baselines and existing works. For each approach, We present several frames with different expressions (Please zoom in for a better view).

Methods	Task	No Cost Data	FID ↓	CSIM ↑	AED ↓	Sync ↑
SadTalker [60]	Audio-Driven	✗	196.733	0.831	0.510	**5.769**
DreamTalk [29]	Audio-Driven	✗	200.856	0.027	0.500	5.676
OTAvatar [30]	Exp-Driven	✗	193.264	0.571	0.506	2.880
Ours	Text-Driven	✓	**187.006**	**0.843**	**0.492**	3.907

Table 2: Quantitative results compared with the several works in other related tasks.

4.2 Dataset and Evaluation Metrics

To the best of our knowledge, no existing dataset is tailored for evaluating our proposed task of text-controllable talking avatar video generation. Consequently, we introduce a novel dataset designed specifically for our task. This dataset includes detailed text descriptions of avatars, as well as the textual content to let the avatar speak out, providing a comprehensive benchmark for both training and evaluation. It further facilitates future development in this field.

Text Descriptions The evaluation dataset has 256 video pairs, consisting of 16 textual avatar descriptions with 16 text contents. Avatar descriptions follow a specific pattern: A $\{\text{adj}_{\text{mood}}\}$ $\{\text{adj}_{\text{age}}\}$ $\{\text{n}_{\text{gender}}\}$ with $\{\text{n}_{\text{hair}}\}$. adj_{mood} can be "happy" or "sad", adj_{age} can be "young" or "old", n_{gender} can be "female" or "male", and n_{hair} can be "short black" or "long blond". This structured representation aids in generating diverse avatars. Text content includes 2 news, 2 fairy tales, 2 scientific research, 2 speeches, 4 wiki entries, and 4 translations (English, Chinese, French, and Japanese).

Text Contents To ensure our application's relevance to a variety of real-world scenarios, we collect several text content types commonly appear in practical open-source applications for further evaluation, including news, fairy tales, scien-

tific research, speeches, and wiki entries in multiple languages including English, Chinese, French, and Japanese.

Evaluation Our task includes generating 256 unique video pairs, each created by matching one of the 16 textual avatar descriptions with 16 text content. To evaluate the visual quality and lip-synchronization of the generated outputs, we utilize a range of established metrics. First, we measure the realism of our results using the Frechet Inception Distance (FID) [21]. Specifically, we adopt images from the High-definition Talking Face Dataset (HDTF) [61] as the real image reference. Identity feature preservation is evaluated with Cosine Similarity of Identity Embeddings (CSIM), based on the ArcFace prediction [11]. To evaluate our model's ability to capture facial expressions, we use Average Expression Distance (AED), measuring the L1 distance between the predicted expression and the ground truth expression from the 3DMM extracted by [13]. Lip synchronization quality, a critical aspect for talking head applications, is assessed using the SyncNet score [7]. These metrics collectively provide a comprehensive assessment of the quality, authenticity, and synchronization accuracy of our generated videos.

4.3 Evaluation Result

Quantitative Comparison In the absence of existing solutions for our novel text-guided talking avatar generation task, we benchmarked our approach against state-of-the-art (SOTA) methods in related tasks such as audio-driven or video-driven tasks. Initially, we generated text-guided 3D avatars from text descriptions, resulting in 2D images. Subsequently, we converted the text content to audio and evaluated SOTA audio-driven methods [29, 60]. Additionally, we converted the text to expressions and utilized a video-driven method [30]. Our quantitative comparison results, presented in Table 2, indicate that both Dreamtalk [29] and OTAvatar [30] struggle to preserve high-fidelity details and perform inadequately in image quality metrics. In contrast, our method outperforms image quality and identity preservation metrics, boasting the lowest FID and the highest CSIM value. Moreover, it achieves comparable results in lip-synchronization metrics including AED and Sync, highlighting its effectiveness in generating realistic images. Importantly, unlike methods in other tasks, our approach introduces the novel text-guided talking avatar generation task, expanding the application of avatar generation by enabling text as the motion signal. Furthermore, the compared methods are trained on public video datasets with significant training complexity, while our method demonstrates zero-shot ability without the need for paired training data, as depicted in Table 1.

Qualitative Comparison Qualitative comparisons between our approach and baselines are depicted in Figure 4. Based on experimental results, Dreamtalk [29] and OTAvatar [30] struggle to preserve image details. SadTalker [60] faces challenges in maintaining view consistency and often generates unnatural images in large head poses due to the lack of explicit 3D constraints during training. In

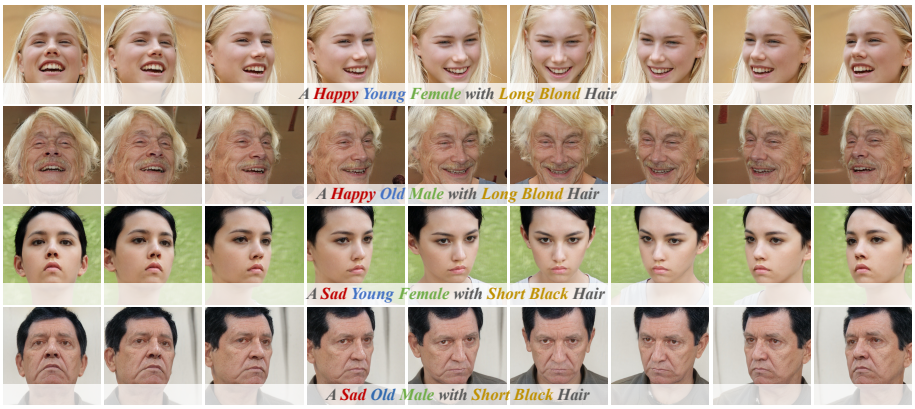

Fig. 5: Qualitative results. Our approach produces high-fidelity, view-consistent, and pose-controllable avatars that align seamlessly with the input texts.

Fig. 6: Qualitative results. Our approach produces high-fidelity talking avatar videos with vivid expressions.

contrast, our method supports a wider range of motion signals, such as text, and achieves superior performance in image quality and view consistency while preserving high-quality details. Additionally, we demonstrate superior results with diverse text descriptions (Figure 5), showcasing our method's ability to produce varied and realistic outputs. Moreover, our approach excels in maintaining view consistency, even in large head poses (Figure 5), and can generate diverse expressions (Figure 6). These figures highlight how our algorithm enables the generation of text-controllable 3D avatars and facilitates their animation with various poses and expressions following the given content.

User Study Perceptual quality is assessed via user studies on 16 generated video clips. Fifteen participants rate lip sync, image fidelity, and realness using Mean Opinion Scores (MOS). Results in Table 3 show our algorithm's superior visual quality and satisfactory lip synchronization.

4.4 Contributions of InpaintNet

We assessed the necessity of our InpaintNet by evaluating its performance alongside two baseline models. The first baseline, Naive 3DMM in Image Space (N3-I),

Methods	Lip Synchronization ↑	Image Fidelity ↑	Image Realness ↑
DreamTalk	3.51	2.88	2.71
SadTalker	**3.58**	3.48	3.23
OTAvatar	2.05	1.77	1.75
Ours	3.44	**3.78**	**3.38**

Table 3: User study.

Method	FID ↓	CSIM ↑	AED ↓	Sync ↑
N3-I (**w/o** InpaintNet)	193.697	0.776	0.554	3.418
N3-F (**w/o** InpaintNet)	195.165	0.808	0.510	2.687
PI-I (**w/** InpaintNet)	190.006	0.842	0.492	2.929
Ours (**w/** InpaintNet)	187.006	0.843	0.492	3.907

Table 4: Ablation study about InpaintNet. We demonstrate the effectiveness of our InpaintNet's design.

utilized a 3DMM-based warping approach on EG3D output images without InpaintNet. We employed our dynamic animation method to warp the final image output from the EG3D decoder. The second baseline, Naive 3DMM in Feature Space (N3-F), applied a 3DMM-based warping method in the EG3D feature space without InpaintNet. Our proposed method, Proposed InpaintNet in Image Space (PI-I), trained an image-level InpaintNet to denoise EG3D output images instead of the feature space using a 3DMM-based animation technique.

The comparison results, shown in Table 4, indicate the effectiveness of our InpaintNet. Specifically, comparing our full method with N3-F and PI-I with N3-I highlights improvements, particularly in image quality metrics such as FID. Moreover, the results between the full method and PI-I affirm the advantages of our approach to warp in the feature space rather than the image space.

4.5 Contributions of Loss Function

The impact of each loss function to train InpaintNet is shown in Table 5. The study assesses three configurations: $\mathcal{L}_r$, $\mathcal{L}_p$ plus $\mathcal{L}_c$, and the full combination including $\mathcal{L}_s$. Results show that the full combination configuration leads to superior results in Sync, reflecting the importance of consistency loss design. However, the image quality metric like FID primarily assesses global image quality, whereas InpaintNet specifically addresses the removal of artifacts from warping, which only affects small parts of the image. Similarly, AED evaluates key point accuracy using 3DMM, and CSIM measures identity preservation. Therefore, the improvements brought by InpaintNet may not be fully captured by these metrics. To better demonstrate the effectiveness of InpaintNet, we introduce an additional metric, FID_l. To compute FID_l, we first crop both the ground truth and final images around regions where the warping operation might introduce artifacts (e.g., around the lip and chin regions) and then calculate the FID score

Loss	FID ↓	FID_l ↓	CSIM ↑	AED ↓	Sync ↑
$\mathcal{L}_r$	187.709	154.683	0.842	0.492	2.840
$\mathcal{L}_r + \mathcal{L}_p$	187.687	153.816	0.842	0.493	2.847
$\mathcal{L}_r + \mathcal{L}_p + \mathcal{L}_c$	187.635	151.193	0.843	0.492	2.863
$\mathcal{L}_r + \mathcal{L}_p + \mathcal{L}_c + \mathcal{L}_s$ (Ours)	187.006	149.784	0.843	0.492	3.907

Table 5: Ablation study about loss function.

for these cropped areas. This approach allows us to focus on the areas where InpaintNet has the most impact, as shown in Table 5. Overall, these findings underscore the effectiveness of our integrated loss functions in producing high-quality, realistic, and synchronized talking avatar videos.

5 Conclusion

In this paper, we introduce the novel task of text-controllable talking avatar generation. We introduce a zero-shot approach that adapts a 3D-aware avatar generation model to this task without requiring data. Our approach seamlessly integrates EG3D for high-fidelity avatar creation, CLIP for aligning text and visual outputs, and 3DMM for dynamic 3D animation capabilities. Further, we propose motion-flow-based warping in the latent feature space of EG3D along with a new InpaintNet to rectify warping artifacts. The outputs are then further refined by the EG3D decoder to produce high-quality and temporally consistent avatar videos. Extensive experiments show that our approach can generate high-fidelity talking avatars with rich facial details, natural expressions, and accurate alignment with user-specified text, which cannot be readily achieved by any existing works as far as we know.

Limitations. Despite the high generative quality achieved, our method's performance is constrained by the accuracy of the 3DMM model. Notably, the model we utilize [34] lacks control over eye regions, resulting in avatars without realistic eye movements. It's worth mentioning that our approach remains independent of the specific 3DMM model employed, suggesting that this issue could potentially be addressed by adopting a more advanced 3DMM model in the future.

Societal Impact. Our study aims to generate positive impacts for users to easily create talking avatars. However, we acknowledge the risks associated with misuse, given our model's ability to generate realistic face images. We strongly urge caution in its use and emphasize the importance of marking synthesized content as fake, with users assuming full responsibility. Our goal is to ensure that research in this area is directed solely towards positive applications.

Future Direction. Our model can create photorealistic talking avatar videos, opening up a new avenue in avatar creation and animation. Future work could extend our approach to create full-body photorealistic 3D talking avatars using texts.

Acknowledgments This work has been supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621), General Research Fund Scheme (Grant No. 17202422), Theme-based Research (Grant No. T45-701/22-R) and RGC Matching Fund Scheme (RMGS). Part of the described research work is conducted in the JC STEM Lab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust. We would like to thank Jiahui Liu for his insightful discussions.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5855–5864 (2021)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision (2017)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
5. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7832–7841 (2019)
6. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? arXiv preprint arXiv:1705.02966 (2017)
7. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 251–263. Springer (2017)
8. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE signal processing magazine **35**(1), 53–65 (2018)
9. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
10. Das, D., Biswas, S., Sinha, S., Bhowmick, B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In: European Conference on Computer Vision. pp. 408–424. Springer (2020)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 4690–4699 (2019)
12. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)

13. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
14. Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: One-shot neural head synthesis and editing. In: Proceedings of the IEEE/CVF International conference on Computer Vision. pp. 14398–14407 (2021)
15. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (TOG) **40**(4), 1–13 (2021)
16. Garrido, P., Zollhöfer, M., Casas, D., Valgaerts, L., Varanasi, K., Pérez, P., Theobalt, C.: Reconstruction of personalized 3d face rigs from monocular video. ACM Transactions on Graphics (TOG) **35**(3), 1–15 (2016)
17. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
19. Guan, J., Zhang, Z., Zhou, H., Hu, T., Wang, K., He, D., Feng, H., Liu, J., Ding, E., Liu, Z., et al.: Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1505–1515 (2023)
20. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5784–5794 (2021)
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
23. Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3397–3406 (2022)
24. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
25. Kumar, R., Sotelo, J., Kumar, K., de Brébisson, A., Bengio, Y.: Obamanet: Photo-realistic lip-sync from text. arXiv preprint arXiv:1801.01442 (2017)
26. Liu, J., Chang, C., Liu, J., Wu, X., Ma, L., Qi, X.: Mars3d: A plug-and-play motion-aware model for semantic segmentation on multi-scan 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9372–9381 (2023)
27. Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. arXiv preprint arXiv:2201.07786 (2022)
28. Lu, Y., Chai, J., Cao, X.: Live speech portraits: real-time photorealistic talking-head animation. ACM Transactions on Graphics (TOG) **40**(6), 1–17 (2021)
29. Ma, Y., Zhang, S., Wang, J., Wang, X., Zhang, Y., Deng, Z.: Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. arXiv preprint arXiv:2312.09767 (2023)

30. Ma, Z., Zhu, X., Qi, G.J., Lei, Z., Zhang, L.: Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16901–16910 (2023)
31. Mallya, A., Wang, T.C., Liu, M.Y.: Implicit Warping for Animation with Image Sets. In: NeurIPS (2022)
32. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
33. Ming, Y., Meng, X., Fan, C., Yu, H.: Deep learning for monocular depth estimation: A review. Neurocomputing **438**, 14–33 (2021)
34. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. pp. 296–301 (2009)
35. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 484–492 (2020)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
37. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13759–13768 (2021)
38. Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1173–1182 (2021)
39. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7763–7772 (2019)
40. Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., Lu, J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In: European conference on computer vision (2022)
41. Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., Lu, J.: Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1982–1991 (2023)
42. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019)
43. Song, L., Wu, W., Qian, C., He, R., Loy, C.C.: Everybody's talkin': Let me talk as you want. IEEE Transactions on Information Forensics and Security (2022)
44. Sun, J., Deng, Q., Li, Q., Sun, M., Ren, M., Sun, Z.: Anyface: Free-style text-to-face synthesis and manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18687–18696 (2022)
45. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)

46. Tan, H.R., Wang, C., Wu, S.T., Wang, T.Q., Zhang, X.Y., Liu, C.L.: Proxy graph matching with proximal matching networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 9808–9815 (2021)
47. Tan, H., Wang, C., Wu, S., Zhang, X.Y., Yin, F., Liu, C.L.: Ensemble quadratic assignment network for graph matching. International Journal of Computer Vision pp. 1–23 (2024)
48. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Fml: Face model learning from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10812–10822 (2019)
49. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: European conference on computer vision. pp. 716–731. Springer (2020)
50. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4563–4573 (2023)
51. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. arXiv preprint arXiv:1910.12713 (2019)
52. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 10039–10049 (2021)
53. Wu, X., Dai, P., Deng, W., Chen, H., Wu, Y., Cao, Y.P., Shan, Y., Qi, X.: Cl-nerf: continual learning of neural radiance fields for evolving scene representation. Advances in Neural Information Processing Systems **36** (2024)
54. Wu, X., Hu, P., Wu, Y., Lyu, X., Cao, Y.P., Shan, Y., Yang, W., Sun, Z., Qi, X.: Speech2lip: High-fidelity speech to lip generation by learning from a short video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22168–22177 (2023)
55. Wu, X., Lyu, X., Huang, Q., Liu, Y., Wu, Y., Shan, Y., Qi, X.: Do3d: Self-supervised learning of decomposed object-aware 3d motion and depth from monocular videos. arXiv preprint arXiv:2403.05895 (2024)
56. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. arXiv preprint arXiv:2301.02379 (2023)
57. Yao, S., Zhong, R., Yan, Y., Zhai, G., Yang, X.: Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. arXiv preprint arXiv:2201.00791 (2022)
58. Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In: European conference on computer vision. pp. 85–101. Springer (2022)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 586–595 (2018)
60. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. arXiv preprint arXiv:2211.12194 (2022)
61. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021)

62. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3657–3666 (2022)
63. Zhao, S., Qi, X.: Prototypical votenet for few-shot 3d point cloud object detection. In: Advances in Neural Information Processing Systems (2022)
64. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9299–9306 (2019)
65. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4176–4186 (2021)
66. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions on Graphics (TOG) **39**(6), 1–15 (2020)