Talk3D: High-Fidelity Talking Portrait Synthesis via Personalized 3D Generative Prior

Jaehoon Ko[1], Kyusun Cho[1], Joungbin Lee[1], Heeji Yoon[1], Sangmin Lee[1], Sangjun Ahn[2], and Seungryong Kim[1]

[1] Korea University, Republic of Korea
[2] NCSOFT, Republic of Korea
https://ku-cvlab.github.io/Talk3D

ER-NeRF [32]   **Talk3D (Ours)**

Fig. 1: Visualizations of generated talking heads by state-of-the-art NeRF-based method, ER-NeRF [32], and Talk3D rendered at extreme novel camera poses. Our Talk3D shows the robustness in generating high-fidelity realistic geometry of talking heads even at unseen poses during training.

Abstract. Recent methods for audio-driven talking head synthesis often optimize neural radiance fields (NeRF) on a monocular talking portrait video, leveraging its capability to render high-fidelity and 3D-consistent novel-view frames. However, they often struggle to reconstruct complete face geometry due to the absence of comprehensive 3D information in the input monocular videos. In this paper, we introduce a novel audio-driven talking head synthesis framework, called **Talk3D**, that can faithfully reconstruct its plausible facial geometries by effectively adopting the pre-trained 3D-aware generative prior. Given the personalized 3D generative model, we present a novel audio-guided attention U-Net architecture that predicts the dynamic face variations in the NeRF space driven by audio. Furthermore, our model is further modulated by audio-unrelated conditioning tokens which effectively disentangle variations unrelated to audio features. Compared to existing methods, our method excels in

generating realistic facial geometries even under extreme head poses. We also conduct extensive experiments showing our approach surpasses state-of-the-art benchmarks in terms of both quantitative and qualitative evaluations.

1   Introduction

Audio-driven talking portrait synthesis aims to synthesize a facial video clip featuring a human portrait with lip movements synchronized to the input audio stream [12, 28, 41, 49, 51, 53, 58, 73]. This task poses various challenges including accurately capturing phonemes, generating realistic movements in facial dynamics, and achieving high-fidelity facial image synthesis. Early approaches have tackled this task using 2D generative models, focusing on image-based reconstruction of lip motion [12, 28, 41, 58], but these methods often exhibit limitation on head pose control.

To address these challenges, recent studies have integrated neural radiance fields (NeRF) [37] in talking head synthesis to leverage its capability in rendering realistic and multi-view consistent images. These approaches either directly condition NeRF on audio features [25, 32, 34, 45, 52, 64] or utilize intermediate representations like facial landmarks [65, 66]. Despite notable advancements, the task of directly constructing dynamic facial NeRF from monocular videos remains challenging. This difficulty stems from the deficiency of diverse head poses and essential 3D information on the monocular videos, resulting in a lack of visual quality when rendered from viewpoints unseen during training. Even the state-of-the-art methods struggle to generate high-fidelity images from extreme camera poses, with its corresponding depth information showing implausible holes and artifacts (see Fig. 1).

On the other hand, 3D-aware generative models [1, 8, 24, 50] recently gained popularity for the ability to generate high-fidelity 3D-aware images through the integration of 3D spatial representation. In response, several works have explored single-view 3D reconstruction on facial images by fusing these 3D generative models and GAN inversion methods [6, 22, 31, 54, 62, 67, 71]. By inverting the input image into the latent space of a pretrained 3D-aware GAN, they faithfully synthesize novel-view images from a single input image.

In this paper, we argue that this capability of 3D-aware GANs can be seamlessly extended to 3D talking head generation, offering the advantage of rendering realistic talking portraits from unseen viewpoints. To implement such a model, one of the intuitive strategies would be directly predicting the latent vector within the GAN latent space conditioned by audio. Notably, previous works have employed this strategy, either conducting audio-driven editing using GAN latent space [3] or predicting latent through 3DMM parameters [36]. However, these approaches encounter challenges due to the high dimensionality of the GAN latent space, which is a complex feature space for the model to learn the elaborate lip movement within the NeRF space. Moreover, the objective of audio-driven NeRF editing tasks is to specifically modify localized regions such

as the lip region or chin, but GAN latent space influences the entire scene, posing a disadvantage to model training.

To overcome these, we introduce a novel audio-driven talking head generation framework leveraging a 3D-aware generative prior, dubbed **Talk3D**, allowing accurate lip movement synchronization with its realistic geometry estimation. Specifically, we integrate two crucial strategies: 1) leveraging 3D-aware GAN prior for realistic geometry and 2) executing direct NeRF space editing beyond the GAN latent space. Specifically, our U-Net-based architecture takes a fixed personalized triplane to yield the triplane offset, namely deltaplane. Modulated by audio features, our model learns to predict a deltaplane that represents precise lip movement within the NeRF space, tailored to the corresponding audio. Additionally, our U-Net architecture employs an attention-based module and accommodates extra features as conditioning tokens for disentanglement. Notably, this architectural choice enables our model to disentangle the local facial variations within the portrait image such as the torso, background, and eye movements. This disentanglement also improves lip-sync accuracy by ensuring the successful mapping of audio features to lip movements.

To summarize, our proposed talking head synthesis method contains contributions as follows:

- We present to adopt 3D generative priors for synthesizing a 3D-aware talking head avatar with realistic geometry.
- We propose an audio-driven deltaplane prediction strategy to modify the NeRF space of the employed 3D generative model conditioned by audio.
- Our audio-guided attention U-Net architecture successfully disentangles the local variations within frame, such as background, torso, and eye movement.
- Our model shows state-of-the-art talking head generation results, proved on extensive experimental results.

2   Related Work

Audio-driven talking portrait synthesis.   Talking portrait synthesis is a challenging task, as the generated lips must be synced to the given audio, while also producing a consistent identity with realistic facial movements. Early deep-learning based methods [14,17,41] employed 2D generative adversarial networks (GANs) [23] to synthesize audio-synchronized lip motions while maintaining realistic facial structure, but lacked control over head poses. Although subsequent works attempted pose control using facial landmarks and 3D facial models [35,53,55,73], the reliance on intermediate representations inadvertently led to information loss during the training process [46].

AD-NeRF [25] first incorporated neural radiance fields (NeRF) [37] to address the challenges of 3D head structure in audio-driven talking portrait synthesis. SSP-NeRF [34] incorporated a semantic sampling strategy to predict the differential impact of audio on facial areas. RAD-NeRF [52] and ER-NeRF [32] achieved substantial improvements in visual quality and efficiency leveraging Instant-NGP [38]. Nevertheless, these NeRF-based methods [25, 32, 34, 45, 52, 64–66]

Table 1: Comparison of our model with existing methods.

	Wav2Lip [41]	AD-NeRF [25]	SSP-NeRF [34]	DFA-NeRF [64]	RAD-NeRF [52]	ER-NeRF [32]	HFA-GP [3]	OTAvatar [36]	Talk3D (Ours)
Audio-driven?	✓	✓	✓	✓	✓	✓	✓	✗	✓
3D-aware?	✗	✓	✓	✓	✓	✓	✓	✓	✓
Generalizability to unseen poses?	✗	✗	✗	✗	✗	✗	✓	✓	✓
Disentangled facial attributes?	✗	✗	✗	✓	✓	✓	✗	NA	✓
Region-aware condition mapping?	✗	✗	✓	✗	✗	✓	✗	✗	✓
Volumetric torso?	✗	✓	✓	✓	✗	✗	✓	✓	✓
Joint volume of head and torso?	✗	✗	✓	✗	✗	✗	✓	✓	✓

struggle to achieve multi-view consistency and realistic geometries, because they learn facial representations only from a single monocular video. Detailed comparison of relevant existing works with our Talk3D is summarized in Tab. 1.
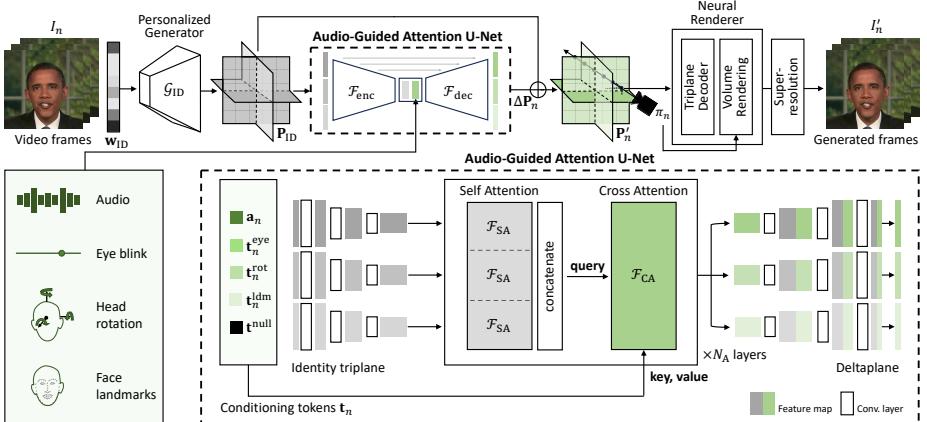
NeRF-based 3D-aware GANs. Some works have extended NeRF for generating images preserving multi-view consistency, by conditioning NeRF on sampled random vector [9] or semantic codes [39, 43]. Subsequent works enabled high-resolution image synthesis by using a super-resolution module [8, 24, 40] or its hybrid 3D representations [8, 20, 44] that integrate NeRF representations with the merits of explicit representations. Notably, EG3D [8], which we use as our base representation, has achieved state-of-the-art image generation quality while maintaining 3D consistency through the design of an efficient triplane hybrid 3D representation.

Facial reconstruction with generative priors. Subsequent works took advantage of EG3D's efficient representation and diverse latent space for usage in 3D facial reconstruction. Some works [6, 29, 31, 33, 54, 60, 67, 71] perform GAN inversion on 3D-aware GANs to reconstruct and edit a 3D facial avatar given a single image. Extensive works explored their various approaches such as direct optimization strategies [31, 67], encoder-based inversion [6, 29, 54, 71], or extending method for inverting consecutive video frames [22, 62]. Recent works [3, 36, 50, 61] enabled facial animation from a given triplane representation, either by presenting a deformation field to morph the reconstructed 3D model, or by sequentially estimating the latent codes that correspond to the given condition.

3   Methodology

3.1   Preliminary: NeRF-based 3D-aware GANs

While conventional neural radiance field (NeRF) [10, 37, 38, 68] aims to be optimized for a single static scene, NeRF-based 3D-aware GANs [8, 9, 18, 39, 43, 44, 59, 63] achieved explicitly pose-controllable image generation by conditioning their NeRF space with random-sampled latent code $\mathbf{w}$. Among these works, EG3D [8] demonstrates its superior performance using three stages. First, EG3D employs a plane generator $\mathcal{G}(\cdot; \theta_{\mathcal{G}})$ parametrized by $\theta_{\mathcal{G}}$ that efficiently synthesizes low resolution feature plane $\mathbf{P}$ such that $\mathbf{P} = \mathcal{G}(\mathbf{w}; \theta_{\mathcal{G}})$. This feature plane is reshaped to three orthogonal feature planes, $\{\mathbf{P}^{\mathrm{xy}}, \mathbf{P}^{\mathrm{yz}}, \mathbf{P}^{\mathrm{zx}}\}$. EG3D then utilizes an MLP

Fig. 2: Overview of our Talk3D framework. Our model mainly utilizes a personalized generator. Given identity triplane $\mathbf{P}_{\mathrm{ID}}$ and n-th frame's conditioning tokens t_n, audio-guided attention U-Net predicts the deltaplane $\mathbf{\Delta P}_n$ which represents the dynamic residual scene variation of the corresponding ground-truth image I_n. This is further combined with $\mathbf{P}_{\mathrm{ID}}$ through summation, forming $\mathbf{P}'_n$, and fed to the neural renderer with given camera viewpoint π_n to generate final output image I'_n.

that takes features aggregated from the orthogonal planes and maps it to volume density σ and feature $\mathbf{f}$. This feature field is rendered to a low resolution 2D feature map $\mathbf{F}$, Finally, the produced feature map $\mathbf{F}$ undergoes processing in a 2D super-resolution module comprised of several convolutional layers to generate the final image I. We denote $\mathcal{R}(\cdot; \theta_{\mathcal{R}})$ as this sequential process involving volume rendering and super-resolution module. Given $\theta_{\mathcal{R}}$ as the learnable parameters, the final synthesized image can be formulated as: $I = \mathcal{R}(\mathbf{P}, \pi; \theta_{\mathcal{R}})$.

3.2 Problem Formulation and Overview

In this section, we describe the main components of our method, **Talk3D**, which enables pose-controllable audio-driven high-fidelity talking portrait synthesis. Given N number of video frames for a specific identity, $\mathcal{V} = \{I_n\}$, our model takes n-th frame image I_n with corresponding audio feature $\mathbf{a}_n$, and camera parameter π_n. We then formulate the audio-driven rendering process as:

$$
\begin{aligned}
\mathbf{P} &= \mathcal{G}(\mathbf{w}; \theta_{\mathcal{G}}), \\
I'_n &= \mathcal{R}(\mathbf{P}, \pi_n, \mathbf{a}_n; \theta_{\mathcal{R}}).
\end{aligned}
\tag{1}
$$

To attain the rendered portrait image I'_n that best replicates the lip movement of the frame I_n, our model aims to find the optimal EG3D [8] parameters denoted as $\{\theta_{\mathcal{G}}^*, \theta_{\mathcal{R}}^*\}$, and the optimal triplane $\mathbf{P}^*$ which encapsulates the appropriate scene encodings. At inference, given new audio $\mathbf{a}_n^{\mathrm{novel}}$, we reformulate (1) as:

$$
\begin{aligned}
\mathbf{P}_{\mathrm{ID}} &= \mathcal{G}(\mathbf{w}_{\mathrm{ID}}; \theta_{\mathcal{G}}^*), \\
I_n^{\mathrm{novel}} &= \mathcal{R}(\mathbf{P}_{\mathrm{ID}}, \pi_n, \mathbf{a}_n^{\mathrm{novel}}; \theta_{\mathcal{R}}^*),
\end{aligned}
\tag{2}
$$

where $\mathbf{w}_{\mathrm{ID}}$ denotes an identity latent code that corresponds to a specific person's facial identity. Then, such a personalized generator generates $\mathbf{P}_{\mathrm{ID}}$, namely identity triplane.

To formulate this, we first train a personalized generator that gives $\mathbf{w}_{\mathrm{ID}}$ and $\{\theta_{\mathcal{G}}^*, \theta_{\mathcal{R}}^*\}$. In the renderer $\mathcal{R}(\cdot)$, to condition $\mathbf{a}_n^{\mathrm{novel}}$, we propose a deltaplane generator that generates a new plane $\Delta\mathbf{P}_n^{\mathrm{novel}}$ from $\mathbf{a}_n^{\mathrm{novel}}$ to manipulate the identity plane $\mathbf{P}_{\mathrm{ID}}$ such that $\mathbf{P}' = \mathbf{P}_{\mathrm{ID}} + \Delta\mathbf{P}_n^{\mathrm{novel}}$. Then our final renderer is defined as follows:

$$I_n^{\mathrm{novel}} = \mathcal{R}(\mathbf{P}', \pi_n; \theta_{\mathcal{R}}^*) = \mathcal{R}(\mathbf{P}_{\mathrm{ID}} + \Delta\mathbf{P}_n^{\mathrm{novel}}, \pi_n; \theta_{\mathcal{R}}^*). \tag{3}$$

In the following, we first explain our generator fine-tuning strategy (Sec. 3.3) and our audio-conditioned deltaplane prediction method (Sec. 3.4). Finally, we describe the loss functions employed in our proposed framework (Sec. 3.5). An overview of our proposed framework is depicted in Fig. 2.

3.3 Personalized Generator for Identity Triplane Generation

3D-aware GANs are usually trained on an extensive dataset of facial images such as FFHQ [30], allowing for the generation of a wide range of personal identities. This nature of 3D-aware GANs may not be an optimal choice for our specific problem setting, which involves capturing the speech of a single person recorded by a monocular video. In this work, we adopt VIVE3D [22], a fine-tuning strategy for 3D-aware GAN to produce single-identity images. This personalizing step aims to enhance the model's editability and visual fidelity.

The personalization strategy involves a variant of pivotal tuning [42], which inverts a few selected frames to find the optimal latent vector in $\mathbf{w}$-space, and then jointly fine-tunes the generator $\mathcal{G}$ and $\mathcal{R}$. They first select M number of frames I_m and simultaneously optimize the latent vectors $\mathbf{w}_{\mathrm{ID}} + \mathbf{o}_m$ of each frame, where $\mathbf{o}_m$ is additional offset vectors that aim to capture the local variants such as facial expression or lip movement. Consequently, they jointly fine-tune the weight of the generator on I_m, while keeping the optimal latent vectors $\mathbf{w}_{\mathrm{ID}} + \mathbf{o}_m$ fixed. Finally, for the N number of the total target frames I_n, they conduct frame-by-frame video inversion on the fine-tuned generator $\mathcal{G}_{\mathrm{ID}}$ to predict a stack of offsets $\mathbf{o}_n$ and camera parameters π_n.

3.4 Audio-Guided Attention U-Net for Deltaplane Generation

Throughout the aforementioned inversion process, we achieve the personalized generator $\mathcal{G}_{\mathrm{ID}}$, the global identity $\mathbf{w}_{\mathrm{ID}}$, and the camera parameters for each frame π_n. We further derive the identity triplane $\mathbf{P}_{\mathrm{ID}} = \mathcal{G}_{\mathrm{ID}}(\mathbf{w}_{\mathrm{ID}}; \theta_{\mathcal{G}}^*)$ to integrate into our training framework. Our ultimate goal is to modulate the generator with an audio feature vector, thus conditioning the NeRF space and enabling image manipulation. While the most straightforward approach [3] involves predicting the latent vector within the generator's learned manifold, we experimentally found that it may not necessarily be the most optimal choice. Alternatively, we

introduce a training method that focuses on the direct prediction of a triplane grid rather than the **w**-space latent vector. In the following, we will explain how to manipulate the triplane with given condition $\mathbf{a}_n$.

Audio-guided attention U-Net architecture. As depicted in Fig. 2, U-Net based architecture $\mathcal{F}$ is employed, where identity triplane serves as an input, yielding offset triplane grid $\Delta\mathbf{P}_n$ such that: $\Delta\mathbf{P}_n = \mathcal{F}(\mathbf{P}_{\text{ID}}, \mathbf{a}_n; \theta)$, where $\mathbf{a}$ denotes given audio feature. This offset grid $\Delta\mathbf{P}_n$, which we call deltaplane is further combined with $\mathbf{P}_{\text{ID}}$. This training strategy offers several distinct advantages compared to GAN latent prediction. First of all, this GAN latent prediction approach struggles to represent the disentangled lip movement due to the high-dimensionality of GAN latent space. This obstacle leads to undesired prediction such as flickering in background and torso area. Furthermore, the triplane grid directly represents the 3D grid structure of the NeRF space which guides the model to understand and manipulate the spatial relationships within the scene. Lastly, the triplane grid is basically a 2D feature map returned from convolutional networks, which offers the convenience of leveraging existing 2D-based network architectures.

Attention design. In an ideal setting, the deltaplane should seamlessly amalgamate temporal motion signals with the identity triplane, ensuring that the signals are appropriately synchronized with the relevant facial segments. This becomes imperative for audio, as their impact on the entirety of the facial movements is not uniform. We incorporate cross-attention at the deepest hidden layer of U-Net architecture to effectively capture localized facial dynamics during the generation of the deltaplane. Specifically, the U-Net encoder $\mathcal{F}_{\text{enc}}$ encodes $\mathbf{P}$ into a low-resolution feature map as $\mathbf{E} = \mathcal{F}_{\text{enc}}(\mathbf{P})$. Consequently, this feature map is passed through N_A number of attention layers, each comprised of self-attention (SA) and cross-attention (CA) layer, which we denote as: $\mathcal{F}_{\text{SA}}$ and $\mathcal{F}_{\text{CA}}$. Specifically, SA and CA can be defined as:

$$\begin{aligned} \mathbf{e} &= \mathcal{F}_{\text{SA}}(\text{flatten}(\mathbf{E} + \mathbf{E}^{\text{pos}})). \\ \mathbf{E}_n^{\text{out}} &= \mathcal{F}_{\text{CA}}(\mathbf{e}, \mathbf{a}_n), \end{aligned} \tag{4}$$

where $\mathbf{E}^{\text{pos}}$ denotes 3D positional encoding. Finally, the U-Net decoder generates the deltaplane by $\Delta\mathbf{P}_n = \mathcal{F}_{\text{dec}}(\mathbf{E}_n^{\text{out}})$.

Split-convolution. The original EG3D [8] employs a single convolution network to generate the triplane, where each plane, $\mathbf{P}^{\text{xy}}$, $\mathbf{P}^{\text{yz}}$, and $\mathbf{P}^{\text{zx}}$, is channel-wise concatenated. However, we observed a performance decline when utilizing the $\mathcal{F}_{\text{enc}}$ structure as a singular model. This degradation stems from the orthogonality of each plane within the NeRF space, and the channel-wise concatenation hinders the 3D-awareness of the triplane. To address this issue, our architecture processes each plane independently to maintain the individual plane's characteristics. Nevertheless, since each plane's features equally contribute to the query sampled points by concatenation, the aforementioned split convolution structure may impede the learning of the correlation between each plane. Therefore, we incorporate the roll-out method [13, 56] to appropriately blend features from each plane.

Augmenting condition. Following previous NeRF-based works [25, 32, 34, 52], our dataset settings closely align with theirs, except for variations in image crop regions. This disparity is caused by the image cropping process in the utilization of the EG3D [8]. Consequently, a specific challenge arises, wherein alterations to the crop area may give the appearance of unnecessary movement between the background and the torso's position, which interferes with the learning of audio features. To mitigate this challenge, we encode additional signals with causal relationships to the torso and background movements. Features capturing independent actions, such as background motion (inferred from facial landmarks coordinate in the original video), and torso dynamics (correlated with head rotation), are tokenized as $\mathbf{t}^{\text{rot}}$ and $\mathbf{t}^{\text{ldm}}$ and then incorporated through cross-attention layers. The intuition here lies in the effectiveness of our model's cross-attention layer, allowing diverse tokens to be efficiently learned for local editing within the triplane. As utilized in prior work [32], we employ the AU45 [21] features to describe eye movements, which also be tokenized into $\mathbf{t}^{\text{eye}}$. Additionally, a single null-token $\mathbf{t}^{\text{null}}$ incorporated uniformly across all frames, serving the purpose of global features across video frames. Again, (4) can be reformulated as:

$$\mathbf{E}_n^{\text{out}} = \mathcal{F}_{\text{CA}}(\mathbf{e}, \mathbf{t}_n), \quad \mathbf{t}_n = \{\mathbf{a}_n, \mathbf{t}_n^{\text{eye}}, \mathbf{t}_n^{\text{rot}}, \mathbf{t}_n^{\text{ldm}}, \mathbf{t}^{\text{null}}\}, \tag{5}$$

where $\mathbf{t}_n$ denotes the concatenation of all tokens.

3.5 Loss Functions

To train the model, we mainly adopt $L1$ loss $\mathcal{L}_{\text{L1}}$ and LPIPS loss $\mathcal{L}_{\text{lpips}}$ [72] to reconstruct given input frame I. Let $\mathcal{L}_{\text{rec}}$ denotes the combination of the above reconstruction loss as: $\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{L1}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}}$. Similarly, we obtain a semantic segmentation of facial region with BiSeNet [69, 70], to enhance the reconstruction loss on the local image area. We give additional reconstruction loss on lip segment $S_{\text{lip}}(I)$. Moreover, we adopt ID similarity loss $\mathcal{L}_{\text{id}}$ and syncnet loss $\mathcal{L}_{\text{sync}}$ [15, 41] to further optimize the generation results. In summary, our total loss function $\mathcal{L}$ can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}}(I, I') + \lambda_{\text{lip}}\mathcal{L}_{\text{lip}}(S_{\text{lip}}(I), S_{\text{lip}}(I')) + \lambda_{\text{id}}\mathcal{L}_{\text{id}}(I, I') + \lambda_{\text{sync}}\mathcal{L}_{\text{sync}}(I, I'). \tag{6}$$

We additionally take a few epochs to update the super-resolution module. This additional fine-tuning step aims to boost rendering quality to generate sharper image results. During the fine-tuning stage, only the reconstruction loss $\mathcal{L}_{\text{rec}}$ is applied, resulting in the following loss function: $\mathcal{L}_{\text{tune}} = \mathcal{L}_{\text{rec}}(I, I')$.

4 Experiments

4.1 Experimental Settings

Dataset. To perform audio-driven talking head synthesis, we require a few minutes of speaking portrait video paired with an audio track. Specifically, in

Table 2: Quantitative comparison under the *novel view synthesis* setting. We measure the image fidelity and lip synchronization accuracy of the generated talking portraits using different NeRF-based methods rendered at different head poses. The head poses are selected with $15°$ yaw intervals and $10°$ pitch intervals. The top, second-best, and third-best results are shown in red, orange, and yellow, respectively.

Head angle (yaw, pitch)	$(-30°, -20°)$			$(-15°, -10°)$			$(+15°, +10°)$			$(+30°, +20°)$		
	Sync↑	FID↓	IDSIM↑	Sync↑	FID↓	IDSIM↑	Sync↑	FID↓	IDSIM↑	Sync↑	FID↓	IDSIM↑
AD-NeRF [25]	2.236	212.845	0.068	3.474	175.978	0.280	3.821	152.018	0.481	2.523	193.343	0.034
RAD-NeRF [52]	4.938	167.834	0.186	5.543	123.924	0.378	6.831	94.674	0.607	5.447	185.718	0.283
ER-NeRF [32]	4.774	198.291	0.226	7.335	87.594	0.575	6.652	80.562	0.503	2.702	141.625	0.022
Talk3D (Ours)	7.201	81.113	0.611	7.932	37.774	0.766	8.144	39.971	0.797	7.766	68.680	0.643

Table 3: The quantitative results of the *self-driven* setting. Red, orange, and yellow highlights indicate the 1st, 2nd, and 3rd-best performing technique for each metric. Note that PSNR, SSIM, and LPIPS are not valid for Wav2Lip [41], as it can see the ground-truth during the self-driven evaluation.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	LMD ↓	AUE ↓	Sync ↑	IDSIM ↑
Ground Truth	N/A	N/A	0	0	0	0	9.077	1
Wav2Lip [41]	28.678	0.862	0.053	33.074	4.658	3.040	10.096	0.893
PC-AVS [75]	20.729	0.638	0.112	42.646	3.419	2.497	8.945	0.520
AD-NeRF [25]	27.611	0.877	0.049	20.243	5.692	2.331	5.692	0.904
RAD-NeRF [52]	28.797	0.886	0.038	14.218	3.467	2.163	6.316	0.921
ER-NeRF [32]	29.284	0.891	0.032	11.860	3.417	2.025	6.724	0.940
Talk3D (Ours)	30.185	0.895	0.027	8.626	2.932	1.920	7.383	0.944

order to compare with the state-of-the-art method, we directly employ datasets from [25, 35], comprising person-centric videos averaging 6,000 frames at 25 fps. Following the training methodology of previous NeRF-based works [25, 32, 52], we split the video into training and testing sets. Furthermore, we utilize a pre-trained Wav2Vec model [2] to extract audio features from each speech audio.

Comparison baselines. We compare our method with 2D talking head research, such as Wav2Lip [41] and PC-AVS [75]. In addition, we also compare our method with several NeRF-based models: AD-NeRF [25], RAD-NeRF [52], and ER-NeRF [32]. Furthermore, we evaluate our method directly on the ground-truth to provide a clearer comparison.

In the supplementary material (Sec. 2), we extend our comparisons to include GeneFace [66] and HFA-GP [3]. is a powerful NeRF-based talking portrait synthesis model but is designed for different settings compared to our paper. HFA-GP is also a noteworthy related work, but the comparison is conducted separately due to the instability of its source code.

4.2 Quantitative Evaluation

Comparison settings and metrics. In quantitative evaluation, we compare the synthesized quality of the head and synchronized lips by constructing three distinct settings: 1) the *novel-view synthesis* experiment, 2) the *self-driven*

Table 4: Quantitative comparison under the *cross-driven* setting. We extract two audio clips from the demo of SynObama [51] to drive each method and compare the audio-lips synchronization and lips movement consistency.

Methods	Testset A			Testset B		
	Sync↑	LMD↓	AUE↓	Sync↑	LMD↓	AUE↓
Ground Truth	7.850	0	0	6.976	0	0
Wav2Lip [41]	8.272	7.039	4.154	7.907	5.561	3.967
PC-AVS [75]	8.408	7.754	6.278	7.533	6.560	4.518
AD-NeRF [25]	5.670	7.378	4.736	5.076	5.542	3.711
RAD-NeRF [52]	6.532	5.848	4.717	5.472	5.599	3.666
ER-NeRF [32]	6.507	6.181	4.489	5.160	5.374	3.519
Talk3D (Ours)	6.827	5.352	4.693	5.780	4.814	3.132

experiment, and 3) the *cross-driven* experiment. The first *novel-view synthesis* setting involves assessing the robustness of viewpoint editing by rendering from diverse novel viewpoints. Specifically, we alter a certain amount of angle degree (yaw: $-30°\sim30°$, pitch: $-20°\sim20°$) from the canonical viewpoint. The *self-driven* shares the same training settings as *novel-view synthesis* differing only in the rendering camera viewpoints, which are extracted from the ground-truth video. Finally, for the *cross-driven* setting, the model is driven by entirely unrelated audio tracks to measure lip synchronization performance using two extracted audio clips from demos of SynObama [51].

In the *self-driven* setting, we utilize evaluation metrics commonly employed in talking portrait synthesis research, including peak signal-to-noise ratio (**PSNR**), structural similarity index measure (**SSIM**), learned perceptual image patch similarity (**LPIPS**) [72] to evaluate image reconstruction quality. Given the absence of ground-truth images for the same identity in the other two settings, we adhere the methodology of previous works [25, 32, 52] and introduce identity-agnostic metrics such as the Fréchet inception distance (**FID**) [26], landmark distance (**LMD**) [11], SyncNet confidence score (**Sync**) [15, 16] for lip synchronization, and action units error (**AUE**) [4, 5] to evaluate face motion accuracy. We also use an identity similarity metric (**ID-SIM**) calculated using an off-the-shelf identity detection network [27] to measure the similarity of facial identities. For a fair comparison, each generated result is cropped and rescaled to the facial area into the same cropping region. Furthermore, since the NeRF-based methods utilize a pre-defined background extracted from the video, we exclusively measure reconstruction metrics on the facial region of the generated result.

Novel pose synthesis evaluation. The result of the *novel-view synthesis* setting is shown in Tab. 2. We compare our method against prior works capable of explicit control of the camera viewpoint. While the majority of methods exhibit comparable performance in frontal view rendering, a notable decline in scores is observed for other NeRF-based techniques when the viewpoint is rotated to different angles. In contrast, our method demonstrates consistently high scores across all metrics, highlighting its efficacy in maintaining performance across diverse viewing angles compared to its counterparts.

Fig. 3: Visualization of synthesized portraits from head poses unseen during training. We show a randomly selected frame from synthesized talking portraits using different rendered at different yaw and pitch (**y**, **p**) angles. Our method demonstrates its robustness on rendering facial images at large head angles which are rarely shown in the training video.

Self-driven evaluation. The *self-driven* evaluation results are presented in Tab. 3. Our method achieves the best quality in most image quality metrics, while also showing the best lip synchronization among NeRF-based methods. Despite the one-shot 2D-based methods, Wav2Lip and PC-AVS present superior Sync scores, their subpar scores in image fidelity highlight their inadequacy in the accurate reconstruction of the specific portrait. Our method demonstrates superior performance in lip synchronization metrics, while also showing comparable image fidelity to other NeRF-based methods.

Cross-driven evaluation. The evaluation results for the *cross-driven* setting, as depicted in Tab. 4, demonstrate the successful performance on general audio input to synthesize the corresponding lip movement. Our model consistently presents the highest scores on most comparisons among NeRF-based methods. The additional use of the sync loss function enforces the model to produce accurate lip shapes even from audio unseen during training. This distinguishes our approach from previous works, which cannot employ sync loss.

4.3   Qualitative Evaluation

In this section, we present generated results from each of the evaluation settings. We first visualize the generated facial images rendered from various viewpoints. As can be seen in Fig. 3, previous NeRF-based methods suffer from performance degradation when generated from camera angles far from the canonical viewpoint, with its results frequently showing inconsistent facial color and artifacts. Especially for the torso region, RAD-NeRF and ER-NeRF experience a substantial decline in quality caused by their torso modeling namely *pseudo-3D deformable module*, showing irregular torso boundary and geometry deterioration.

Fig. 4: The comparison of the keyframes and details of generated portraits. We show visualizations of our method and previous methods under the self-driven setting (left side) and the cross-driven setting (right side). Best viewed in zoom.

For AD-NeRF, where the head and torso volumes are learned independently, the synthesized heads appear disembodied when viewed from side angles. We also show the sampled results from *self-driven* and *cross-driven* experiments in Fig. 4. We choose four key frames from each of the two experiments to compare the lip-sync accuracy and reconstruction quality. As mentioned in 4.2, although 2D-based methods such as Wav2lip and PC-AVS demonstrate high lip-sync accuracy, the generated results cannot fully reconstruct the given scene. Although the NeRF-based methods manage to create a full portrait, the separated rendering pipeline shows unnatural head movements at the neck part, and also shows blurry textures, especially at the hair part. In contrast, our Talk3D demonstrates robust and accurate results, attributed to its unified generation process. Notably, Talk3D avoids the issue of open lips in the absence of speech audio, demonstrating accurate results by appropriately closing the lips, an aspect where other NeRF-based methods falter.

4.4   User Study

We present a user study to assess the visual quality of the generated heads. We invited 31 participants to compare 9 randomly selected video clips from the quan-

Table 5: User study results. The rating is of scale 1-5, the higher the better. The top, second-best, and third-best results are shown in red, orange, and yellow, respectively.

Settings	Methods	Wav2Lip [41]	PC-AVS [75]	AD-NeRF [25]	RAD-NeRF [52]	ER-NeRF [32]	**Talk3D(Ours)**
novel-view synthesis	Lip-sync Accuracy	–	–	2.056	2.411	2.983	3.103
	Image Quality	–	–	0.924	1.417	2.532	3.123
	Video Realness	–	–	1.205	0.834	2.163	2.242
self-driven	Lip-sync Accuracy	3.455	2.511	2.455	2.636	2.909	3.394
	Image Quality	2.623	0.607	3.723	3.650	3.789	3.970
	Video Realness	2.868	0.757	2.936	2.991	3.223	3.467
cross-driven	Lip-sync Accuracy	2.933	1.767	2.867	2.467	2.667	3.301
	Image Quality	2.967	0.767	3.733	3.441	3.763	3.798
	Video Realness	2.801	0.878	3.233	2.731	3.183	3.267

Table 6: Ablation study on the sync loss function.

Method	PSNR ↑	LPIPS ↓	LMD ↓	AUE ↓	Sync ↑
Ground Truth	-	-	0	0	8.605
w/o sync	26.180	0.068	3.149	1.715	6.137
All (Ours)	26.799	0.054	3.227	1.540	6.529

titative evaluation of the main study. Utilizing the mean opinion scores (MOS) rating protocol, participants first provided ratings for the generated videos of the *novel-view synthesis setting*, *self-driven setting* and *cross-driven setting*, each based on three criteria: (1) lip-sync accuracy; (2) image quality; and (3) video realness. The average scores for each method are presented in Tab. 5, revealing that our Talk3D outperforms most of the criteria. These results demonstrate the outstanding visual quality of our method, in light of both facial reconstruction and novel view synthesis.

4.5 Ablation Study

In this section, we present the ablation study to validate the efficacy of our primary contributions. All ablation studies are conducted under a slightly different setting than the *self-driven* scenario, with the key distinction being the measurement of metrics on the entire image pixels.

Use of the sync loss. Due to the computationally expensive nature of NeRF limits full-image rendering during training time, prior NeRF-based works [25, 32, 34, 52] solely employ pixel-based MSE loss and patch-wise LPIPS loss. On the other hand, leveraging the efficient representation of EG3D, our model is capable of utilizing full image-based loss functions such as the sync loss function. In Tab. 6, we assess the significance of the sync loss by comparing results without its utilization. While forgoing the sync loss function marginally enhances reconstruction accuracy, it is essential for generating well-synchronized lips.

Feature token selection. We also investigate the significance of using augmented conditions, such as eye blink, head rotation, and facial landmarks. In Tab. 7, we measure the impact of each feature on image fidelity by turning them on and off in turn. The lower PSNR scores are attributed to the low lip-sync accuracy due to the feature entanglement or the absence of eye blink features

Table 7: Ablation study on use of each feature token.

Method	PSNR ↑	LPIPS ↓	LMD ↓	AUE ↓	Sync ↑
Ground Truth	-	-	3.322	1.815	8.605
w/o null-vec	25.745	0.064	2.781	1.650	6.267
w/o eye feature	25.862	0.062	3.335	1.598	6.414
w/o landmark tokens	26.195	0.059	3.392	1.719	5.498
w/o angle tokens	26.152	0.060	3.313	1.920	6.508
All (Ours)	26.799	0.054	3.227	1.540	6.529

Table 8: Ablation study on specific design selections for deltaplane prediction.

Method	PSNR ↑	LPIPS ↓	LMD ↓	AUE ↓	Sync ↑
Ground Truth	-	-	0	0	8.605
w/o deltaplane	19.180	0.187	4.675	2.939	1.192
w/o attention	24.925	0.071	3.485	2.115	4.591
w/o split	24.403	0.096	3.793	2.730	1.024
w/o rollout	25.621	0.064	3.233	1.962	6.438
All (Ours)	26.799	0.054	3.227	1.540	6.529

preventing proper eye closure. We also show detailed visualizations in Sec. 4 of the supplementary material.

Deltaplane predictor design. We further ablate on the network designs of deltaplane predictor. Tab. 8 shows four different design choices, which are predicting $\mathbf{w}$ latent vector instead of deltaplane (w/o deltaplane), replacing attention module to conditional affine layer (w/o attention), merging split-convolution layer into a single convolution (w/o split), and removing roll-out method utilized in Sec. 3.4. Our model exhibits superior image quality and lip-sync accuracy compared to other design choices, highlighting the effectiveness of our model architecture.

5   Conclusion

In this paper, we introduced **Talk3D**, a novel framework that incorporates 3D-aware GAN prior and a region-aware motion for high-fidelity 3D talking head synthesis. Our framework incorporates a personalized generator fine-tuned using the VIVE3D framework, allowing for the synthesis of 3D-aware talking head avatars with its realistic geometry and explicit rendering viewpoint control. Furthermore, our proposed audio-guided attention U-Net architecture enhances the disentanglement of local variations within image frames, such as background, torso, and eye movements. Through extensive experiments, we demonstrate that our proposed model not only produces accurate lip movements corresponding to the input audio but also enables rendering from novel viewpoints, addressing limitations observed in previous state-of-the-art approaches. We anticipate our work will significantly impact digital media experiences, and virtual interactions, and find applications in film-making, virtual avatars, and video conferencing.

Appendix

In the following, we provide implementation details and offer further analysis of our experiment along with extensive qualitative results. Specifically, in Sec. A we describe the implementation details including data preprocessing pipeline, feature extraction, and network architecture. In Sec. B, we present additional experiments and comparisons that demonstrate the robustness and effectiveness of our method. In Sec. C, we also provide further analyses to support the efficacy of our contributions, presenting qualitative results of ablation studies, and visualizations of attention map and triplane. In Sec. D, we also introduce semantic manipulation over the generated portraits. Finally, in Sec. E and Sec. F, we briefly explain our supplementary video and further discuss the limitations and ethical considerations of our research.

A   Additional Implementation Details

A.1   Pre-processing

We follow the same image cropping as VIVE3D [22]. They detect the 6 facial landmarks from every video frame, using an off-the-shelf detector [7] and perform Gaussian smoothing on the landmarks along the temporal axis to stabilize the transition of the cropping area. They additionally detect landmarks on a single reference image generated from the personalized generator. This reference image serves as an anchor for every frame to calculate the affine transformation matrices. Using these affine matrices, we calculate the cropping boundaries for each of the raw images. More specifically, they utilize a slightly wider cropping boundary compared to EG3D [8] which employs Deep3DFace [19] for image cropping.

A.2   Augmenting feature extraction

For the audio feature extraction, we follow [52] which employ the pre-trained Wav2Vec [2] model and further encode with several layers of 1D convolutions. On the other hand, the augmenting features such as the eye (scalar factor), head rotation angles (3-dimensional vector), and landmarks (6-dimensional vector) are comparatively low-dimensional feature vectors. Therefore, we upsample these augmenting features using the positional encodings and further encode with several layers of MLP. Each of the output features is a 64-dimensional feature token and is fed to our cross-attention network $\mathcal{F}_{\mathrm{CA}}$.

A.3   Network architecture

Attention network. Our deltaplane predictor $\mathcal{F}$ first encodes the 256-resolution triplane $\mathbf{P}$ into 32-resolution feature map $\mathbf{E}$, while its hidden dimension is upscaled from 32 to 256.

With given flattened image feature vector $\mathbf{e}$ and conditioning tokens $\mathbf{t}_n$, our cross-attention layer predicts the low-resolution feature map $\mathbf{E}_n^{\text{out}}$ as:

$$\mathbf{E}_n^{\text{out}} = \mathcal{F}_{\text{CA}}(\mathbf{e}, \mathbf{t}_n). \tag{7}$$

Given learnable parameters of cross-attention layer $\mathbf{w}_{\text{q}}$, $\mathbf{w}_{\text{k}}$, $\mathbf{w}_{\text{v}}$, the above process can be divided into the sub-processes as:

$$Q = \mathbf{e}\mathbf{w}_{\text{q}}, \quad K_n = \mathbf{t}_n\mathbf{w}_{\text{k}}, \quad V_n = \mathbf{t}_n\mathbf{w}_{\text{v}}, \tag{8}$$

$$A_n = \text{softmax}(QK_n^{\mathsf{T}}), \quad \mathbf{E}_n^{\text{out}} = A_n V_n, \tag{9}$$

where Q, K_n, V_n denote query, key and value representation, and A_n represents attention scores. Each of the parameters represents MLP with 1 layer and 64 hidden dimensions.

Super-resolution module. We replaced the original super-resolution module in EG3D [8] with GFPGAN [57], which enhances rendering quality by reducing noise or artifacts in the background. Following the training strategy documented in the main paper, we fine-tune the pre-trained GFPGAN for a few epochs. For a fair comparison, all quantitative evaluations were measured on the results obtained by using the original EG3D's super-resolution module.

B   Additional Results and Comparisons

B.1   Novel-view synthesis and depth information

We demonstrate the robustness of Talk3D by generating images from the extreme viewpoints, shown in Fig. 5. We compare our method with the previous NeRF-based methods [25,32,52] visualizing both the generated images and their corresponding depth maps. Note that, the other NeRF-based methods do not synthesize the background and therefore lack depth information in that particular area. Furthermore, they frequently show the head and torso separation due to their separative volumetric representation for torso rendering. Especially, RAD-NeRF [52] and ER-NeRF [32] employ a 2D deformation neural field for torso rendering, thus they are not capable of generating a realistic torso geometry. In contrast, our model successfully constructs the entire image as a single NeRF representation, providing depth information for all parts of the synthesized portrait.

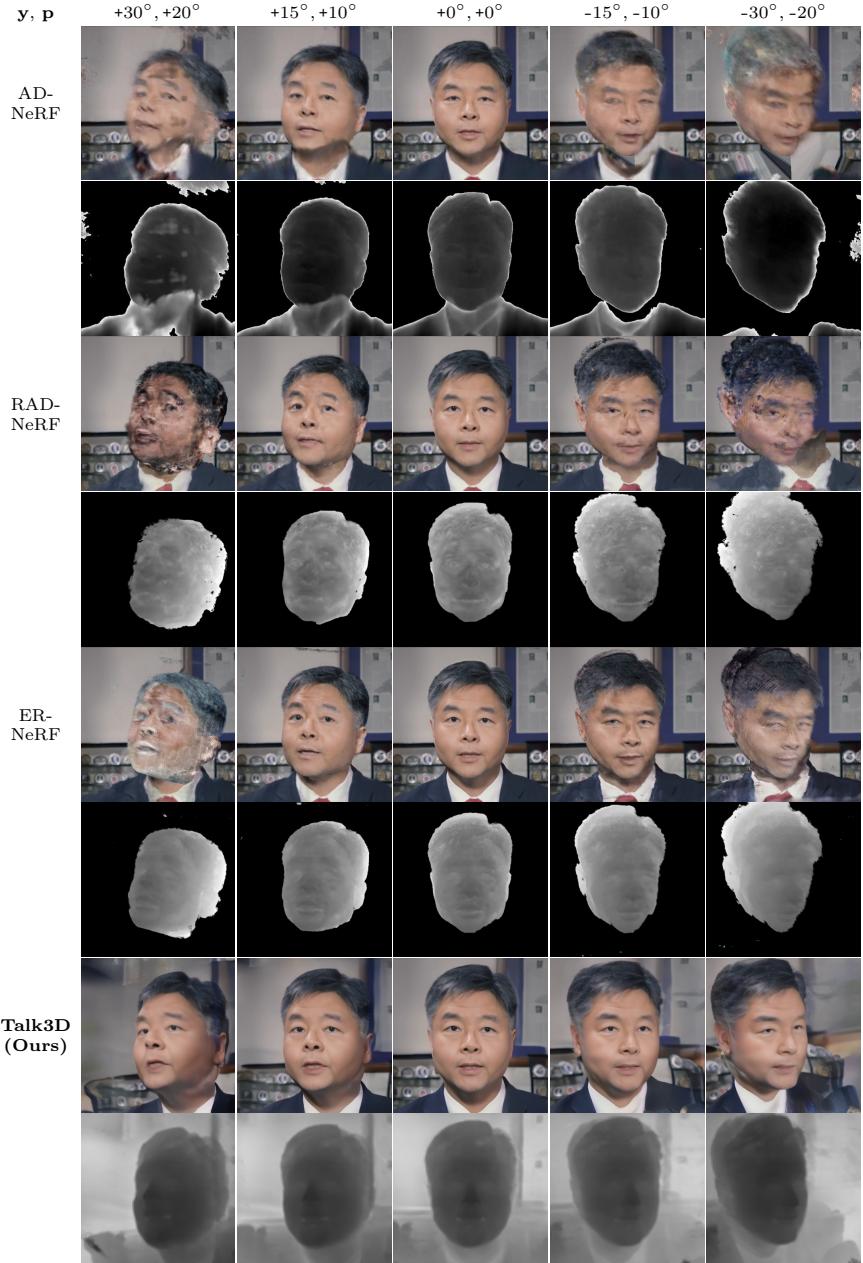## B.2   Additional dataset and qualitative results

To further demonstrate the generalizability of our method, besides the dataset [25] in the main paper, we also conduct experiments on datasets from HDTF [74], which are the YouTube video clips containing in-the-wild talking portraits of 720p or 1080p resolution. We utilize the same facial cropping method of VIVE3D [22] and split each video for both training and validation. We show additional generated results on HDTF dataset under the *self-driven* setting and the *cross-driven* setting in Fig. 6 and Fig. 7, respectively.

B.3   Comparison with HFA-GP

HFA-GP [3] is a similar work that utilizes a personalized 3D-aware generative prior for talking head synthesis. Considering the high relevance of HFA-GP, it is crucial to compare HFA-GP and our model thoroughly. However, the implementation code of HFA-GP was incomplete, lacking essential components crucial for reproducing the experimental setup outlined in the original paper. We attempted to fix errors in the code, yet the outcomes differed significantly from those reported in the original paper. These discrepancies are not entirely fair to them, thus the results in Tab. 9 and Fig. 8 are limited to provide an accurate comparison for the model performance.

B.4   Comparison with GeneFace

GeneFace [66] is another NeRF-based talking portrait synthesis method, but instead of directly conditioning the NeRF model on audio features, it employs audio-to-motion mapping trained on a corpus of diverse talking heads. We compare our method with GeneFace on Tab. 10 and Fig. 9. Note that the GeneFace is designed for different settings, the comparisons are not entirely fair and are just for reference.

Fig. 5: Visualization of synthesized portraits and depth map rendered from novel viewpoints. We show a randomly selected frame from synthesized talking portraits (odd rows) and corresponding depth information (even rows) using different rendering viewpoints of yaw and pitch angles (**y**, **p**) with $15°$, $10°$ intervals.
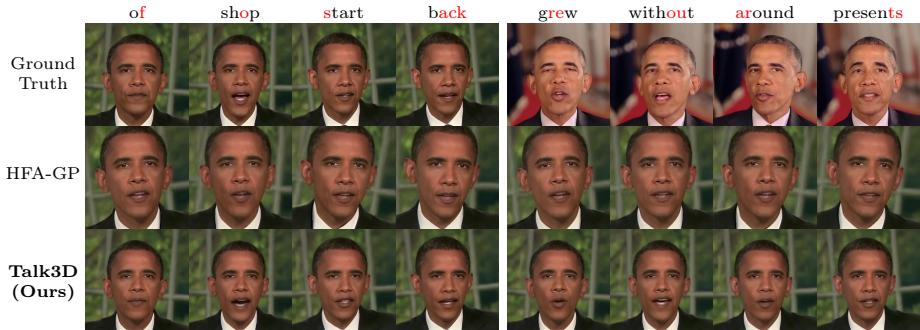
Fig. 6: The *self-driven* comparison of the key frames and details of generated portraits. We show visualizations of our method and related methods under the self-driven setting. Best viewed in zoom.

Fig. 7: The *cross-driven* comparison of the key frames and details of generated portraits. We show visualizations of our method and related methods under the cross-driven setting. Best viewed in zoom.

Table 9: Quantitative comparison against HFA-GP [3] at the *self-driven* setting. The HFA-GP results are not an accurate representation of their work and are just for reference.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	LMD ↓	AUE ↓	Sync ↑	IDSIM ↑
Ground Truth	N/A	N/A	0	0	0	0	9.077	1
HFA-GP [3]	19.612	0.712	0.164	25.397	5.161	3.146	0.474	0.913
Talk3D (Ours)	30.185	0.895	0.027	8.626	2.932	1.920	7.383	0.944

Fig. 8: The comparison of the key frames and details of generated portraits. We show visualizations of our method and HFA-GP under the self-driven setting (left side) and the cross-driven setting (right side).

Table 10: Quantitative comparison against GeneFace [66] at the *self-driven* setting. Across all evaluation metrics, Talk3D exhibits consistently higher performance than GeneFace.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	LMD ↓	AUE ↓	Sync ↑	IDSIM ↑
Ground Truth	N/A	N/A	0	0	0	0	9.077	1
GeneFace [66]	26.305	0.832	0.069	15.30	4.948	2.758	6.200	0.909
Talk3D (Ours)	30.185	0.895	0.027	8.626	2.932	1.920	7.383	0.944

Fig. 9: The comparison of the key frames and details of generated portraits. We show visualizations of our method and GeneFace under the self-driven setting (left side) and the cross-driven setting (right side).
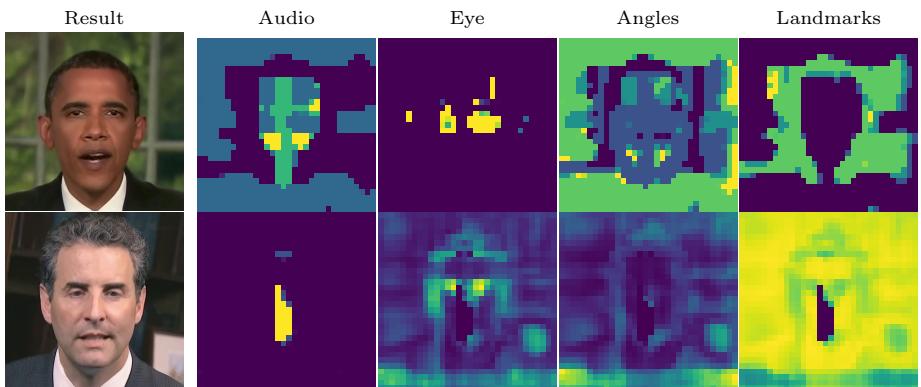
C   Further Analysis
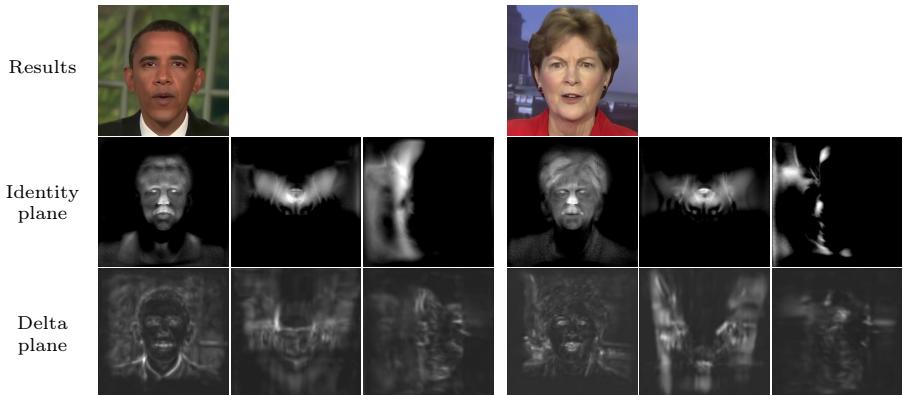
C.1   Analysis of attention

In Fig. 10, we visualize the attention map to demonstrate the efficacy of our attention-based network. The first column is the generated image result, while the rest of the columns show the attention map of the low-resolution xy-plane (the plane that is orthogonal to canonical direction) captured by each of the specific conditioning tokens. The result shows that each of the conditioning tokens successfully disentangles the local movements of the low-resolution feature map. Despite the close relationship between angles and landmarks, they capture different attention maps, since the head rotation angles are closely related to torso movement, while facial landmarks are suitable for capturing the background motion.

C.2   Analysis of triplane

Fig. 11 visualizes the generated image alongside its two corresponding triplanes: the identity plane and the deltaplane. Each column shows the three orthogonal planes. Especially xy-plane (orthogonal to canonical direction) in the 1st and 4th columns highlights the facial structure representation within the identity plane. Also, the deltaplane visualization confirms our method's ability to precisely manipulate specific regions such as the lips, eyes, torso, and background.

Fig. 10: Visualizations of attention maps. Our region attention module successfully captures the relation between diverse conditioning tokens and spatial regions.
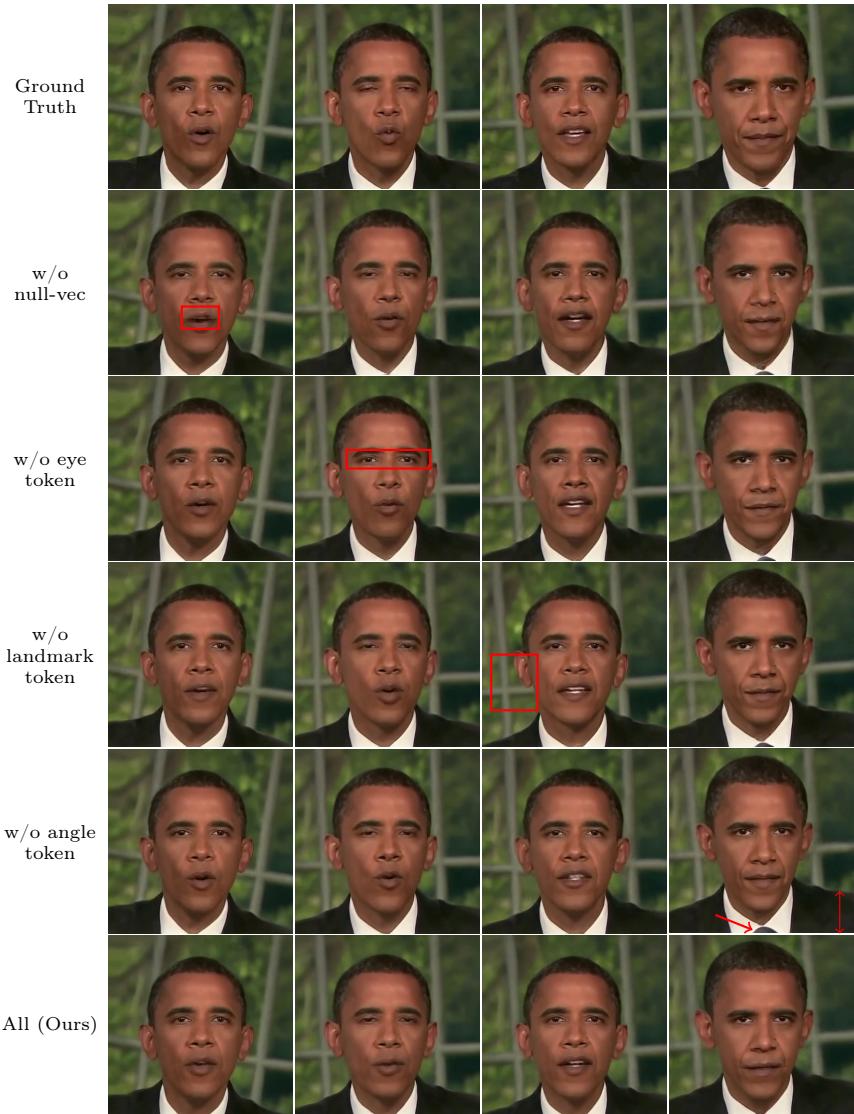
Fig. 11: Visualizations of triplanes. We visualize generated image results and their corresponding triplanes. Each set of three columns depicts the orthogonal planes of the triplane representation.
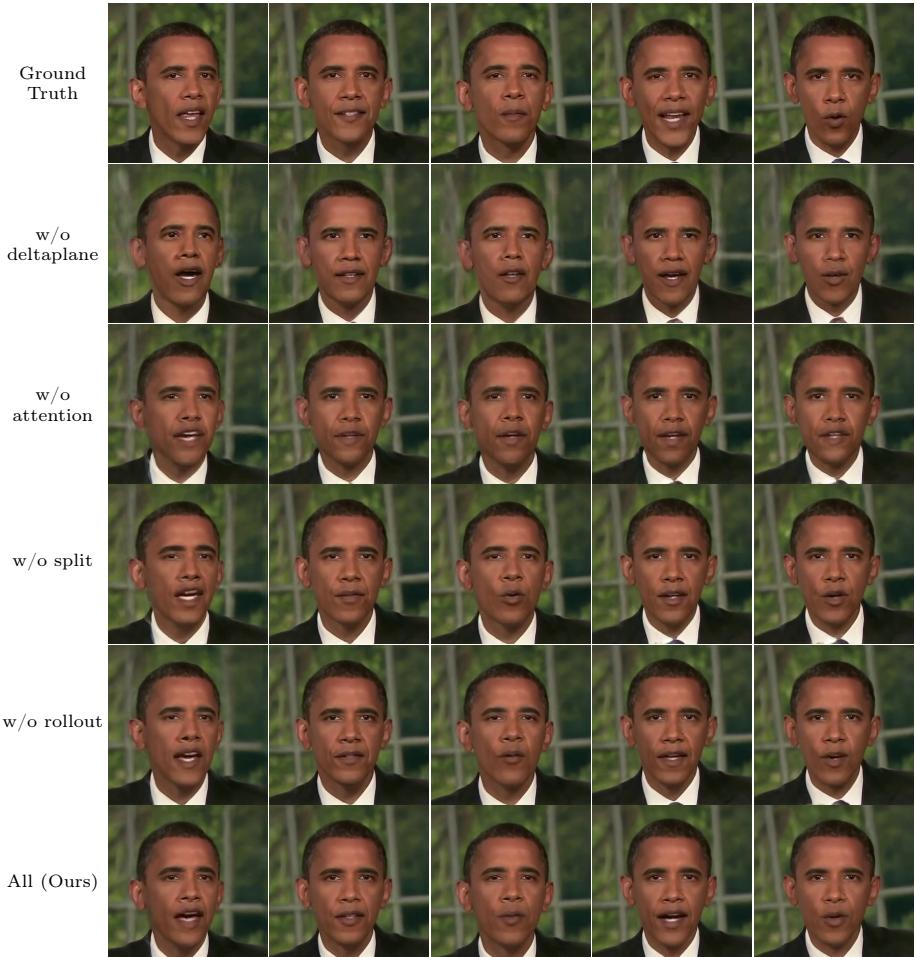
C.3 Ablation studies

Importance of each token selection. Fig. 12 provides a qualitative comparison from the ablation study of feature token selection (discussed in the main paper). The figure demonstrates how each feature token influences its corresponding region. We highlight the specific regions with red boxes, that illustrate each of the image quality degradation. Eye features are crucial for accurate eye movement prediction, while angle and landmark tokens are essential in controlling torso and background movement, respectively. See the difference in the eye-blink, background movement, or neck-tie region. The results obtained without employing null vectors show artifacts, particularly in the mouth region.

Deltaplane predictor. Fig. 13 showcases a qualitative evaluation of our delta-plane predictor's design choices discussed in the main paper. This evaluation highlights the critical role of each design element in optimizing both image quality and lip-synchronization accuracy. Our results establish the necessity of all design aspects in achieving these performance gains.

Effect of the personalized generator. To demonstrate the effectiveness of incorporating the personalized generator, we compare our full model with an ablation setting that does not utilize the generator fine-tuning method of VIVE3D [22]. The quantitative comparison in Tab. 11 illustrates the image quality enhancement achieved through our personalization method, as reflected in the higher metrics of our full model. In Fig. 14, we observed image quality degradation from the ablation result, especially for the noisy pixels around the nose or eye region.

Fig. 12: Ablation study on the use of each feature token. We assess the effectiveness of each feature token by alternatively turning them on and off.

Fig. 13: Ablation study on specific design selections of deltaplane predictor. We perform an ablation study on the deltaplane predictor design choice to demonstrate the impact of our method on modeling accurate facial reconstruction.

Table 11: Ablation study on the use of personalized generator.

Method	PSNR ↑	LPIPS ↓	FID ↓	Sync ↑	LMD ↓	AUE ↓	IDSIM ↑
Ground Truth	-	-	-	8.605	0	0	1
w/o $\mathcal{G}_{ID}$	26.031	0.060	17.860	6.498	3.237	1.802	0.879
w/ $\mathcal{G}_{ID}$	26.799	0.054	8.627	6.529	3.227	1.540	0.917

Fig. 14: Visualization of the effect of personalized generator.
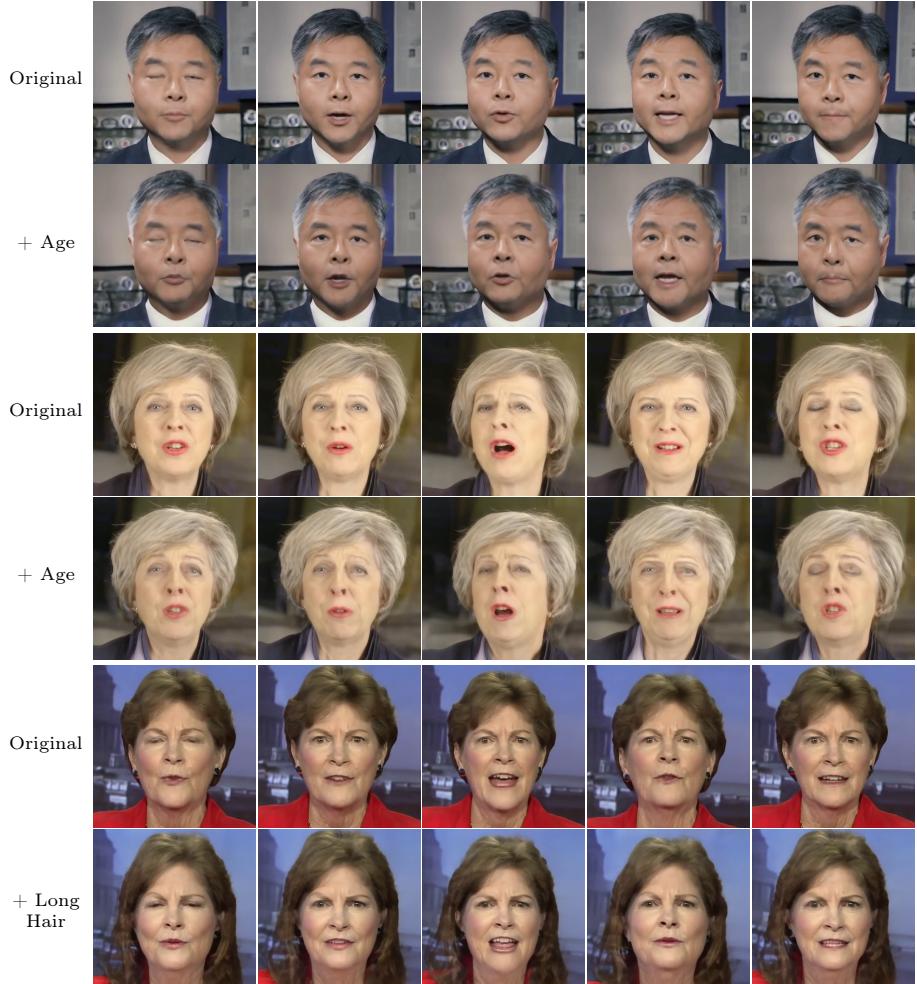
D   Facial Editing

In this section, we introduce an additional feature of our model that distinguishes it from other NeRF-based methods: facial attribute manipulation. Talk3D is built on pre-trained EG3D [8] and thus inherits the rich and diverse latent space of the generative models. The latent space of EG3D enables semantic editing by adding pre-defined style vectors to the input latent code. We exploit Inter-FaceGAN [47,48] to find several style vectors $\mathbf{w}_{\text{edit}}$ which represent the semantic editing directions within the EG3D latent space. However, naively applying InterFaceGAN to our methodology is not feasible, since our approach directly predicts the triplane representation instead of a latent code. So we slightly alter the methodology of InterFaceGAN by simply replacing the identity triplane with the edited triplane $\mathbf{P}_{\text{edit}}$. Specifically, for given personalized generator $\mathcal{G}_{\text{ID}}$ and identity latent code $\mathbf{w}_{\text{ID}}$, we first construct the edited triplane $\mathbf{P}_{\text{edit}}$ as:

$$\mathbf{P}_{\text{edit}} = \mathcal{G}_{\text{ID}}(\mathbf{w}_{\text{ID}} + \mathbf{w}_{\text{edit}}; \theta_{\mathcal{G}}^*). \tag{10}$$

Then we replace the identity triplane to generate edited image I_n^{edit} as:

$$I_n^{\text{edit}} = \mathcal{R}(\mathbf{P}_{\text{edit}} + \Delta\mathbf{P}_n, \pi_n; \theta_{\mathcal{R}}^*). \tag{11}$$

In Fig. 15, we visualize the results of editing several attributes, including age and hair length. The process demonstrates consistent manipulation across attributes like age and hair length, without disrupting lip synchronization.

Fig. 15: Facial attribute manipulation results.

E   Supplementary Video

To comprehensively visualize the efficacy of our proposed method in the domain of talking facial video synthesis, we have compiled a supplementary video that encapsulates each result shown in the main paper and the supplementary materials. This video showcases not only the facial animations generated by our method but also includes detailed comparisons with other relevant techniques. The video also features diverse outcomes, highlighting our approach's versatility across different languages and robustness of rendering performance at extreme viewpoints. Additionally, the video incorporates an ablation study that dissects the contributions of individual components within our methodology.

F   Broader Impact

F.1   Ethical considerations

In developing Talk3D, we hope to advance applications in digital humans, video production, and human-computer interaction assistance by generating highly realistic talking portraits with accurate lip-audio synchronization. However, we recognize the ethical considerations surrounding the misuse of such technology for malicious purposes. The photorealistic nature of the generated portrait videos makes it challenging for individuals to distinguish between authentic and synthetic content. To address this concern, we emphasize the importance of informing users about the authenticity of videos and recognize the ongoing challenges in discriminating synthesized high-fidelity portraits from recent NeRF-based methods. To contribute to the responsible development and deployment of talking portrait synthesis, we commit to sharing our generated results with deepfake detection communities and supporting the enhancement of detection mechanisms. Additionally, we advocate for protective measures, such as incorporating digital watermarks in real portrait speech videos, to mitigate potential misuse. Furthermore, we highlight the need to consider regulatory frameworks that govern the use of deepfake techniques to prevent unintended negative consequences when synthetic content is shared on social media platforms. Both policymakers and the public must be informed about the potential risks associated with deepfakes, fostering a cautious and responsible approach to their creation and utilization.

F.2   Limitations and future work

Compared to earlier NeRF-based works, our Talk3D excels in high-fidelity talking portrait synthesis, especially when rendered from the extreme viewpoint, thanks to the rich generative prior of EG3D [8]. However, we also inherit its shortcomings, as our method does not generalize well outside of photorealistic images of human faces, compared to other talking face synthesis works such as MakeItTalk [76] and SadTalker [73], which can handle cartoon characters or stylized caricatures. Also, our model's current reliance on GAN inversion introduces

technical complexities that impact data preparation. Specifically, it demands precise alignment and cropping of video frames, extending preprocessing time. Additionally, incomplete coverage of essential facial regions can lead to visual artifacts in the form of blurriness or distortion. Future development will focus on overcoming these limitations to increase our method's adaptability and ensure consistent performance across a wider range of training data.

References

1. An, S., Xu, H., Shi, Y., Song, G., Ogras, U.Y., Luo, L.: Panohead: Geometry-aware 3d full-head synthesis in 360deg. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20950–20959 (June 2023)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. NeurIPS **33**, 12449–12460 (2020)
3. Bai, Y., Fan, Y., Wang, X., Zhang, Y., Sun, J., Yuan, C., Shan, Y.: High-fidelity facial avatar reconstruction from monocular video with generative priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4541–4551 (June 2023)
4. Baltrušaitis, T., Mahmoud, M., Robinson, P.: Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). vol. 6, pp. 1–6. IEEE (2015)
5. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018)
6. Bhattarai, A.R., Nießner, M., Sevastopolsky, A.: Triplanenet: An encoder for eg3d inversion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3055–3065 (2024)
7. 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: Proceedings of the IEEE international conference on computer vision (2017)
8. 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)
9. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
10. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII. pp. 333–350. Springer (2022)
11. Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII 15. pp. 538–553. Springer (2018)
12. 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)

13. Chen, X., Deng, Y., Wang, B.: Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
14. Christos Doukas, M., Zafeiriou, S., Sharmanska, V.: Headgan: Video-and-audio-driven talking head synthesis. arXiv (2020)
15. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 87–103. Springer (2017)
16. 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)
17. Das, D., Biswas, S., Sinha, S., Bhowmick, B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In: ECCV (2020)
18. Deng, Y., Yang, J., Xiang, J., Tong, X.: GRAM: Generative radiance manifolds for 3D-aware image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
19. 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 (2019)
20. DeVries, T., Bautista, M.A., Srivastava, N., Taylor, G.W., Susskind, J.M.: Unconstrained scene generation with locally conditioned radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14304–14313 (2021)
21. Ekman, P., Friesen, W.V.: Facial Action Coding System: Manual. Palo Alto: Consulting Psychologists Press (1978)
22. Frühstück, A., Sarafianos, N., Xu, Y., Wonka, P., Tung, T.: Vive3d: Viewpoint-independent video editing using 3d-aware gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4446–4455 (2023)
23. 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)
24. Gu, J., Liu, L., Wang, P., Theobalt, C.: StyleNeRF: A style-based 3D aware generator for high-resolution image synthesis. In: Proceedings of the International Conference on Learning Representations (2022)
25. 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)
26. 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)
27. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: Adaptive curriculum learning loss for deep face recognition. In: CVPR (2020)
28. Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: Synthesising talking faces from audio. International Journal of Computer Vision **127**, 1767–1779 (2019)
29. Jiang, B., Guo, Z., Yang, Y.: Meta-auxiliary network for 3d gan inversion. arXiv preprint arXiv:2305.10884 (2023)

30. 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)
31. Ko, J., Cho, K., Choi, D., Ryoo, K., Kim, S.: 3d gan inversion with pose optimization. WACV (2023)
32. Li, J., Zhang, J., Bai, X., Zhou, J., Gu, L.: Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. arXiv preprint arXiv:2307.09323 (2023)
33. Lin, C., Lindell, D., Chan, E., Wetzstein, G.: 3d gan inversion for controllable portrait image animation. In: ECCV Workshop on Learning to Generate 3D Shapes and Scenes (2022)
34. Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII. pp. 106–125. Springer (2022)
35. Lu, Y., Chai, J., Cao, X.: Live speech portraits: Real-time photorealistic talking-head animation. ACM Trans. Graph. **40**(6) (dec 2021)
36. 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)
37. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
38. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
39. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11453–11464 (2021)
40. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: StyleSDF: High-resolution 3D-consistent image and geometry generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2022)
41. 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)
42. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Trans. Graph. (2021)
43. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: GRAF: Generative radiance fields for 3D-aware image synthesis. In: Proceedings of the Advances in Neural Information Processing Systems (2020)
44. Schwarz, K., Sauer, A., Niemeyer, M., Liao, Y., Geiger, A.: VoxGRAF: Fast 3D-aware image synthesis with sparse voxel grids. In: Proceedings of the Advances in Neural Information Processing Systems (2022)
45. Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., Lu, J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII. pp. 666–682. Springer (2022)
46. 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. pp. 666–682. Springer (2022)

47. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
48. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE transactions on pattern analysis and machine intelligence (2020)
49. 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 **17**, 585–598 (2022)
50. Sun, J., Wang, X., Wang, L., Li, X., Zhang, Y., Zhang, H., Liu, Y.: Next3d: Generative neural texture rasterization for 3d-aware head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20991–21002 (2023)
51. 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)
52. Tang, J., Wang, K., Zhou, H., Chen, X., He, D., Hu, T., Liu, J., Zeng, G., Wang, J.: Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368 (2022)
53. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 716–731. Springer (2020)
54. Trevithick, A., Chan, M., Stengel, M., Chan, E., Liu, C., Yu, Z., Khamis, S., Chandraker, M., Ramamoorthi, R., Nagano, K.: Real-time radiance fields for single-image portrait view synthesis. ACM Transactions on Graphics (TOG) **42**(4), 1–15 (2023)
55. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI. pp. 700–717. Springer (2020)
56. 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)
57. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9168–9178 (2021)
58. Wiles, O., Koepke, A.S., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII 15. pp. 690–706. Springer (2018)
59. Xiang, J., Yang, J., Deng, Y., Tong, X.: GRAM-HD: 3D-consistent image generation at high resolution with generative radiance manifolds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
60. Xie, J., Ouyang, H., Piao, J., Lei, C., Chen, Q.: High-fidelity 3d gan inversion by pseudo-multi-view optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 321–331 (2023)
61. Xu, H., Song, G., Jiang, Z., Zhang, J., Shi, Y., Liu, J., Ma, W., Feng, J., Luo, L.: Omniavatar: Geometry-guided controllable 3d head synthesis. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12814–12824 (2023)

62. Xu, Y., Shu, Z., Smith, C., Huang, J.B., Oh, S.W.: In-n-out: Face video inversion and editing with volumetric decomposition. arXiv preprint arXiv: 2302.04871 (2023)

63. Xue, Y., Li, Y., Singh, K.K., Lee, Y.J.: Giraffe HD: A high-resolution 3D-aware generative model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18440–18449 (2022)

64. 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)

65. Ye, Z., He, J., Jiang, Z., Huang, R., Huang, J., Liu, J., Ren, Y., Yin, X., Ma, Z., Zhao, Z.: Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. arXiv preprint arXiv:2305.00787 (2023)

66. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In: The Eleventh International Conference on Learning Representations (2022)

67. Yin, F., Zhang, Y., Wang, X., Wang, T., Li, X., Gong, Y., Fan, Y., Cun, X., Shan, Y., Oztireli, C., et al.: 3d gan inversion with facial symmetry prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 342–351 (2023)

68. Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. arXiv preprint arXiv:2112.05131 (2021)

69. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision **129**, 3051–3068 (2021)

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

71. Yuan, Z., Zhu, Y., Li, Y., Liu, H., Yuan, C.: Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2437–2447 (October 2023)

72. 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. pp. 586–595 (2018)

73. 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8652–8661 (2023)

74. 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 (2021)

75. 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)

76. 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)