

Figure A.1: Comparison of facial pose distribution of FFHQ dataset and training videos.

Appendix A. Motivation

The videos used for training NeRF-based 3D talking head models typically consist of long speech videos, where the speaker faces a static camera throughout. This setup results in minimal pose variation across the entire video, making it difficult to train a complete NeRF representation from scratch due to insufficient diversity in camera viewpoints.

To address this limitation, we leverage the learned 3D geometry of EG3D [3], which is trained on the FFHQ [7] dataset—a large collection of facial images with diverse pose distributions. In Fig. A.1, we compare the facial pose distributions of the FFHQ dataset and the training videos (Obama and Lieu) commonly used by NeRF-based baselines. As shown, the FFHQ dataset exhibits significantly greater pose diversity, covering a wide range of yaw and pitch angles. This broader distribution enables EG3D to provide robust generative priors, resulting in better image fidelity and geometry reconstruction at unseen poses compared to models trained solely on monocular videos.

Appendix B. Additional Implementation Details

In this section, we provide more details of the implementation of Talk3D. Our network requires approximately 8 hours to train and achieves an inference speed of 13

frames per second. We utilized a single Nvidia RTX 3090 for training and inference for our experiments.

Appendix B.1. Dataset

To perform audio-driven talking head synthesis, we require a few minutes of speaking portrait video paired with an audio track. Specifically, to compare with the state-of-the-art method, we directly employ datasets from AD-NeRF [6], comprising person-centric videos averaging 6,000 frames at 25 fps. Following the training methodology of previous NeRF-based works [6, 15, 13], we split the video into training and testing sets.

Appendix B.2. Pre-processing

We follow the same image cropping as VIVE3D [5]. They detect the 6 facial landmarks from every video frame, using an off-the-shelf detector [2] 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 [3] which employs Deep3DFace [4] for image cropping.

Appendix B.3. Augmenting feature extraction

For the audio feature extraction, we follow RAD-NeRF [13] which employ the pre-trained Wav2Vec [1] 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; obtained by concatenating the 2D coordinates of three individual landmarks) 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}_{CA} .

Appendix B.4. 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 upsampled 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:

$$\mathbf{E}_n^{\text{out}} = \mathcal{F}_{\text{CA}}(\mathbf{e}, \mathbf{t}_n). \quad (\text{B.1})$$

Given learnable parameters of cross-attention layer $\mathbf{w}_q, \mathbf{w}_k, \mathbf{w}_v$, the above process can be divided into the sub-processes as:

$$Q = \mathbf{e}\mathbf{w}_q, \quad K_n = \mathbf{t}_n\mathbf{w}_k, \quad V_n = \mathbf{t}_n\mathbf{w}_v, \quad (\text{B.2})$$

$$A_n = \text{softmax}(QK_n^T), \quad \mathbf{E}_n^{\text{out}} = A_n V_n, \quad (\text{B.3})$$

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 [3] with GFPGAN [14], 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.

Appendix C. Novel-view synthesis and depth information

We demonstrate the robustness of Talk3D by generating images from the extreme viewpoints, shown in Fig. C.2. We compare our method with the previous NeRF-based methods [6, 13, 8] 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 [13] and ER-NeRF [8] 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.

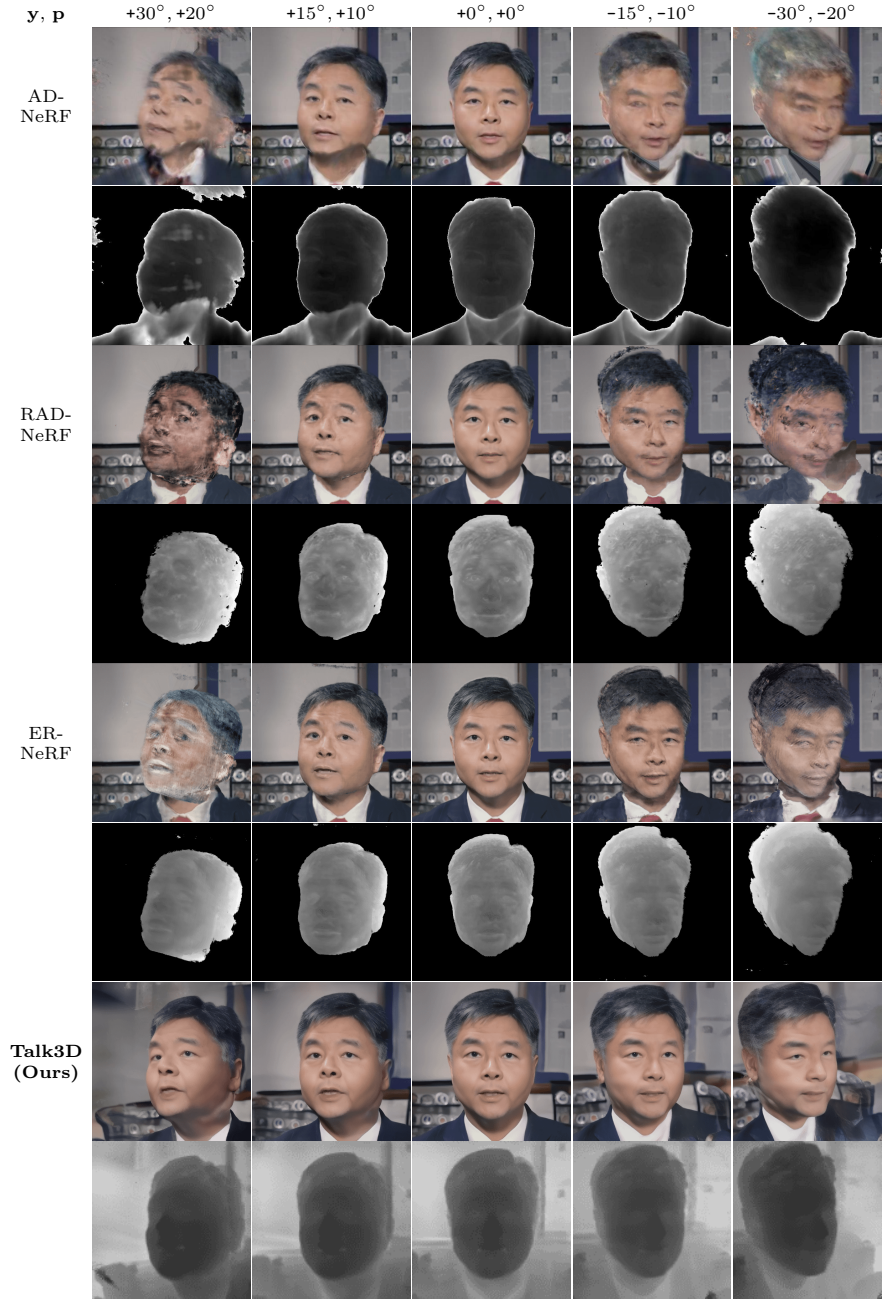


Figure C.2: **Visualization of synthesized portraits and depth map rendered from novel viewpoints.** We show a randomly selected frame from synthesized talking portraits (odd rows) and corresponding depth information (even rows) using different rendering viewpoints of yaw and pitch angles (y, p) with $15^\circ, 10^\circ$ intervals.

Methods	Testset A			Testset B		
	Sync \uparrow	LMD \downarrow	AUE \downarrow	Sync \uparrow	LMD \downarrow	AUE \downarrow
Ground Truth	7.850	0	0	6.976	0	0
AD-NeRF	5.670	7.378	4.736	5.076	5.542	3.711
RAD-NeRF	<u>6.532</u>	<u>5.848</u>	4.717	<u>5.472</u>	5.599	3.666
ER-NeRF	6.507	6.181	4.489	5.160	<u>5.374</u>	<u>3.519</u>
Talk3D (Ours)	6.827	5.352	<u>4.693</u>	5.780	4.814	3.132

Table C.1: **Quantitative comparison under the cross-driven setting.** We extract two audio clips from the demo of SynObama [12] to drive each method and compare the audio-lips synchronization and lips movement consistency.

Appendix D. Cross-driven synthesis

The cross-driven setting evaluates model generalization by synthesizing lip movements from entirely unrelated audio tracks not seen during training. We use two distinct audio clips extracted from SynObama [12] to drive each method, measuring lip-sync accuracy independent of training data.

As shown in Table 3, Talk3D achieves the highest SyncNet scores and lowest landmark distance among NeRF-based methods across both test sets. This demonstrates superior generalization to novel audio inputs while maintaining precise lip synchronization. The method’s audio-driven deltaplane strategy combined with sync loss optimization enables accurate mapping from unseen phonemes to lip movements.

Settings	Methods	Wav2Lip [10]	PC-AVS [16]	AD-NeRF [6]	RAD-NeRF [13]	ER-NeRF [8]	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
audio-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 D.2: **User study results.** The rating is on a scale of 1-5, the higher the better. The top, second-best, and third-best results are shown in **red**, **orange**, and **yellow**, respectively.

Appendix E. 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 quantitative evaluation of the main study. We also include results from 2D talking head research, such as Wav2Lip [10] and PC-AVS [16]. Utilizing the mean opinion scores (MOS) rating protocol, participants first provided ratings for the generated videos of the novel-view synthesis setting, audio-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. D.2, 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.

Appendix F. Further Analysis

Appendix F.1. Analysis of attention

In Fig. F.3, 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 the canonical direction) captured by each of the specific conditioning tokens. The result shows that each of the conditioning tokens successfully disentangles

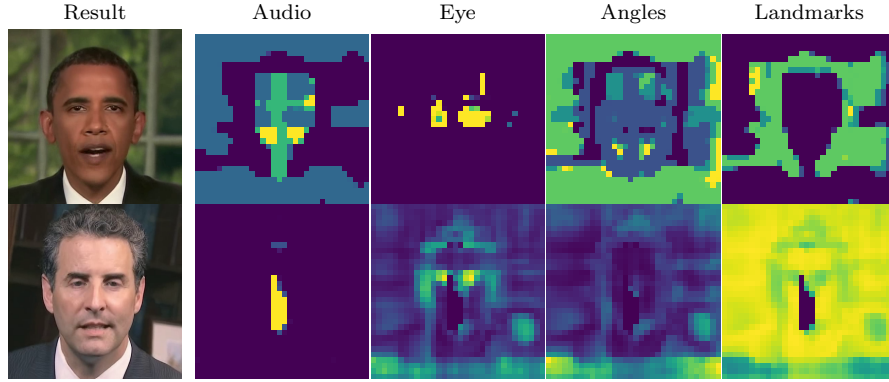


Figure F.3: **Visualizations of attention maps.** Our region attention module successfully captures the relation between diverse conditioning tokens and spatial regions.

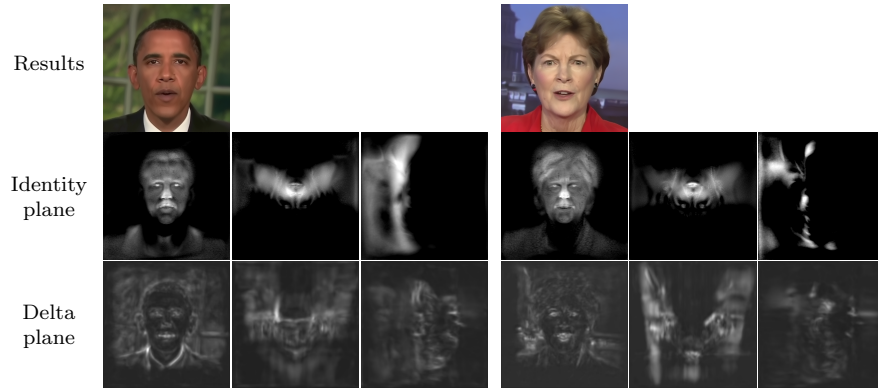


Figure F.4: **Visualizations of triplanes.** We visualize generated image results and their corresponding triplanes. Each set of three columns depicts the orthogonal planes of the triplane representation.

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.

Appendix F.2. Analysis of triplane

Fig. F.4 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.

Appendix F.3. Ablation Study

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

Appendix F.3.1. Use of the sync loss.

Due to the computationally expensive nature of NeRF limits full-image rendering during training time, prior NeRF-based works [6, 9, 13, 8] 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. F.3, 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.

Appendix F.3.2. Feature token selection.

We also investigate the significance of using augmented conditions, such as eye blink, head rotation, and facial landmarks. In Tab. F.4, we measure the impact of each feature on image fidelity by turning them on and off in turn. The lower Sync and AUE scores are caused by the feature entanglement between lip movement and other scene variations, which degrades lip-sync accuracy. Furthermore, PSNR and SSIM show that the absence of each token impacts the reconstruction of scene variations such as proper eye closure or torso movement. We also show detailed visualizations in the appendix.

Appendix F.3.3. Deltaplane predictor design.

We further ablate design choices of deltaplane predictor. Tab. F.5 shows four different design choices, which are predicting \mathbf{w} latent vector instead of deltaplane(w/o

Table F.3: **Ablation study** on the importance of using the sync loss function.

Method	PSNR \uparrow	LPIPS \downarrow	LMD \downarrow	AUE \downarrow	Sync \uparrow
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

Table F.4: Ablation study on use of each feature token.

Method	PSNR \uparrow	LPIPS \downarrow	LMD \downarrow	AUE \downarrow	Sync \uparrow
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	26.195	0.059	3.392	1.719	5.498
w/o angle	26.152	0.060	3.313	1.920	6.508
All (Ours)	26.799	0.054	3.227	1.540	6.529

Table F.5: **Ablation study** on specific design selections for deltaplane prediction.

Method	PSNR \uparrow	LPIPS \downarrow	LMD \downarrow	AUE \downarrow	Sync \uparrow
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

deltaplane), replacing attention module to affine layer(w/o attention), merging split-convolution layer into a single convolution(w/o split), and removing roll-out method. Our model exhibits superior image generation quality compared to other design choices, highlighting the effectiveness of our model architecture. Notably, the results for the *w/o deltaplane* configuration—where facial movements are modeled by traversing EG3D’s latent space rather than using the delta plane—show significantly weaker lip synchronization and overall reconstruction accuracy. This underscores the importance of directly predicting the delta plane to capture precise audio-driven dynamics while preserving structural integrity.

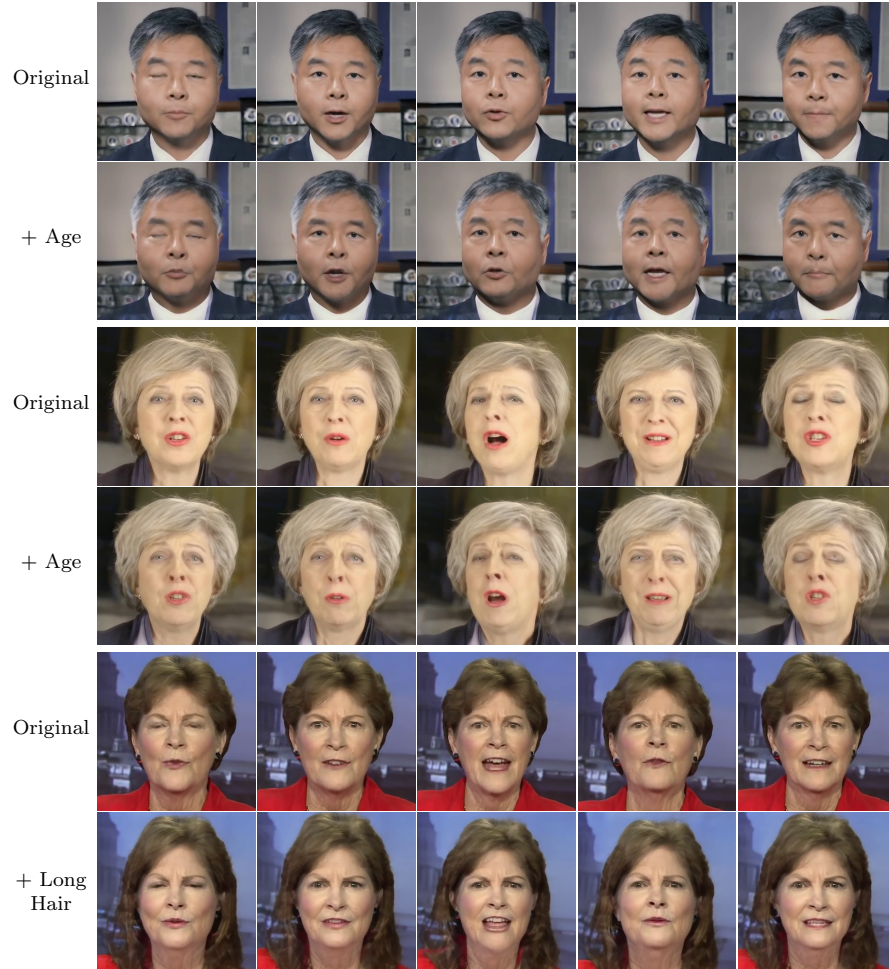


Figure G.5: **Facial attribute manipulation results.**

Appendix G. Application: 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 [3] 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 InterFaceGAN [11] to find several style vectors \mathbf{w}_{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}_{edit} . Specifically, for given personalized generator \mathcal{G}_{ID} and identity latent code \mathbf{w}_{ID} , we first construct the edited triplane \mathbf{P}_{edit} as:

$$\mathbf{P}_{\text{edit}} = \mathcal{G}_{\text{ID}}(\mathbf{w}_{\text{ID}} + \mathbf{w}_{\text{edit}}; \theta_{\mathcal{G}}^*). \quad (\text{G.1})$$

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

$$I_n^{\text{edit}} = \mathcal{R}(\mathbf{P}_{\text{edit}} + \Delta\mathbf{P}_n, \pi_n; \theta_{\mathcal{R}}^*). \quad (\text{G.2})$$

In Fig. G.5, 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.

Appendix H. Broader Impact

Appendix H.1. Ethical considerations

Talk3D aims to advance applications in digital humans, video production, and human-computer interaction by generating realistic talking portraits with accurate lip-audio synchronization. However, the technology’s potential misuse for malicious purposes raises ethical concerns, particularly in distinguishing authentic from synthetic content. To mitigate misuse, we advocate for measures like digital watermarks, collaboration with deepfake detection communities, and regulatory frameworks to foster responsible use and inform policymakers and the public about associated risks.

Appendix H.2. Limitations and future work

While Talk3D demonstrates strong performance in high-fidelity talking portrait synthesis by leveraging the generative prior of EG3D [3], it has several limitations that warrant further exploration:

1. **Preprocessing Challenges with GAN Inversion:** The reliance on GAN inversion introduces preprocessing challenges, such as the need for precise alignment

and cropping of input videos. Errors in these steps can lead to visual artifacts, particularly around the neck and background regions. Future work will explore more robust inversion techniques that reduce preprocessing dependency.

2. **Limited Temporal Consistency:** Although Talk3D achieves accurate lip synchronization and disentangles audio-driven dynamics from unrelated variations, it does not explicitly incorporate temporal enhancement. Adding lightweight temporal consistency mechanisms could improve smoothness and coherence across video frames.
3. **Background Decoupling:** Talk3D inherits EG3D’s limitation of not separating foreground and background during training or inference. This can result in artifacts when handling complex backgrounds. Future efforts will focus on developing methods to decouple background dynamics while maintaining computational efficiency.

By addressing these limitations, we aim to improve Talk3D’s adaptability and ensure consistent performance across diverse datasets, including those with varying visual styles, complex backgrounds, and dynamic temporal requirements.

References

- [1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 33:12449–12460, 2020.
- [2] Adrian Bulat and Georgios Tzimiropoulos. 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.
- [3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.
- [4] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 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.
- [5] Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. Vive3d: Viewpoint-independent video editing using 3d-aware gans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4446–4455, 2023.
 - [6] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5784–5794, 2021.
 - [7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019.
 - [8] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 7568–7578, 2023.
 - [9] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. 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, pages 106–125. Springer, 2022.
 - [10] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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, pages 484–492, 2020.
 - [11] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.
 - [12] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG), 36(4):1–13, 2017.

- [13] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368, 2022.
- [14] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9168–9178, 2021.
- [15] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In The Eleventh International Conference on Learning Representations, 2022.
- [16] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4176–4186, 2021.