

Towards High-fidelity 3D Talking Avatar with Personalized Dynamic Texture

Supplementary Material

In the supplementary material, we provide further information about our work, including more details about our implementation (Sec. A), the setup of our data capture system (Sec. B), the specific design of our user study (Sec. C) with additional experiments (Sec. D), and discussions on ethical impacts (Sec. E). We also present a short description of our supplementary video (Sec. F). The code and dataset will be publicly released after publication.

A. Method Details

A.1. Implementation Details

The training of TexTalker consists of two parts: (1) training two codebooks that store the animation primitive of facial motion and wrinkle, respectively, and 2) training the motion-wrinkle latent diffusion model (LDM) based on the learned latent spaces. We provide further details about the implementation of the two parts.

Animation Primitive Learning. We utilize Face3D [3] to map vertex offsets to UV space, generating motion maps. These maps are then normalized for better learning. For the wrinkle map, we use sigmoid to transform the value range to 0-1 after computing the pixel-wise ratio with the neutral texture. Taking the motion autoencoder for example, following VQGAN [1], our encoder \mathcal{E}_f and generator \mathcal{G}_f respectively consists of 5 resizing layers and 12 residual blocks. A single attention layer is applied on the lowest resolution. The codebook size is set to 1024 and the dimension of each item is set to 16, balancing the expressiveness and computational costs. We use the patch-based discriminator \mathcal{D}_f as in [4]. For better training stability, the discriminator is introduced after 40K iterations. All setting remains the same for the wrinkle autoencoder \mathcal{E}_w , \mathcal{G}_w and \mathcal{D}_w .

Motion-Wrinkle LDM. The latent diffusion model consists of an eight-layer transformer decoder with eight attention heads. The diffusion timestep \mathbf{n} , clean sample $\mathbf{X}_{-T_p:0}^{r_0}$ from the previous window, and current noisy sample $\mathbf{X}_{0:T_w}^m$ are first embedded into a dimension of 1024 and then concatenated together before denoising. A positional embedding layer is then applied to the features. We align the length of audio features with visual frames through linear interpolation. Following [8], the learnable start features are used to generate the initial window. In addition, considering that the window length is not fixed during inference, we randomly truncate the input samples during training to improve the robustness.

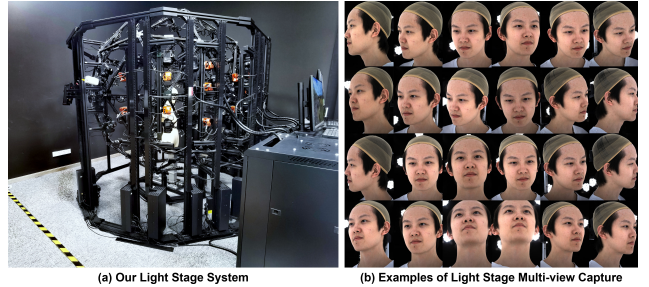


Figure A. The diagram of our capture system.

A.2. Inference Speed

We conduct inference with an NVidia 3090 GPU. Our inference stage involves two steps: (1) using LDM to generate motion and wrinkle latent codes from audio, and (2) generating motion maps and wrinkle maps separately using generator \mathcal{G}_f and \mathcal{G}_w . Our LDM can generate motion and wrinkle latent features at 24 FPS and the generators can reconstruct maps at 38 FPS.

B. Capture System Setup

As is shown in Fig. A, our LightStage system for 4D data capture consists of 24 time-consistent dynamic cameras capable of capturing multi-view videos at 60FPS with a resolution of 4096×3000 pixels. Each camera is hardware-controlled with a time error of less than 1 microsecond. The 24 cameras are precisely calibrated to obtain accurate intrinsic and extrinsic parameters. To capture real facial wrinkles while avoiding environmental influences, we employ multiple surrounding light sources for uniform lightning. We use a dual-channel 48000Hz microphone synchronized with the cameras to record talking audio.

We capture a 1-minute-long video for each of the 100 subjects. All subjects are required to prepare the text content for talking in advance. The content is reviewed by on-site staffs, which should have a variety of syllables and not contain any sensitive information, personal privacy, or insulting and discriminatory content. During filming, subjects are constrained to chairs while talking, maintaining relatively static body movements, which aligns with the facial capture requirements in industrial production processes. The subjects spontaneously recite the lengthy text to simulate natural speech conditions.

Instructions:

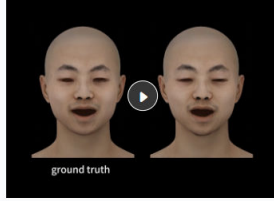
The study consists of 30 questions, each shows a short video (duration ~20s) of two facial animations. Observe the animations carefully and answer the questions.

- Focus on the **facial details**, including shadows, wrinkles, and texture changes.
- Rate the animation **on the right** in terms of its **realism** and **consistency** from 1 to 5, the higher the the better.
- Ensure your **sound** is turned on for the best experience.

This survey will take approximately **5-10 minutes** to complete. Thank you for your participation!

***consistency**: to what extent the facial texture change is visually aligned with the underlying facial motion.

Q1 Referring to ground truth, please carefully observe the changes in facial shadows and wrinkles in the two animations.



Q2 Referring to the ground truth, how realistic do you think the animation on the right is?

- ☐ 5
☐ 4
☐ 3
☐ 2
☐ 1

Q3 Referring to the ground truth, how consistent do you think the facial changes in the animation on the right is?

- ☐ 5
☐ 4
☐ 3
☐ 2
☐ 1

Figure B. Demonstration of the user study interface design.

C. User Study Details

Fig. B shows the designed user study interface. The study consists of 30 questions, each presenting the participants with a random 20-second-long clip of the ground truth animation alongside the corresponding fragments generated by an anonymous method. The order of appearance of different methods is disrupted. For each video, participants are instructed to rate the anonymous animation on a scale of 1-5 based on the ground truth, with higher ratings indicating better performance. Each question consists of two sub-items: “Referring to the ground truth, how realistic do you think the animation on the right is?” and “Referring to ground truth, how consistent do you think the facial changes in the animation on the right are?” To ensure the participants are fully engaged in the user study, only when they watch the entire video and rate all the items can they proceed to the next question, otherwise a warning message would pop up. Only when all questions are answered will they be included in the final results.

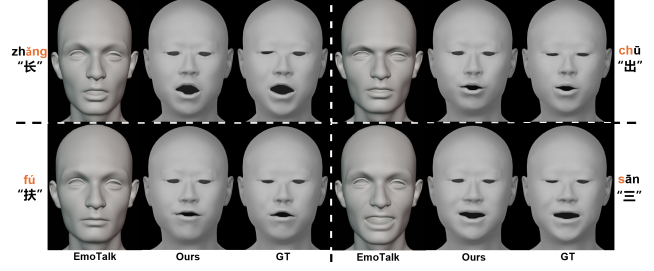


Figure C. Comparison with EmoTalk on unseen ID and speech.

Audio Encoder	Motion			Texture		
	LVE↓ 10 ⁻² mm	MVE↓ 10 ⁻² mm	FDD↓ 10 ⁻³ mm	PSNR↑ dB	SSIM↑ -	LPIPS↓ -
MFCC	2.13	2.79	1.57	43.92	0.984	0.0106
Wav2Vec2-CN	2.10	2.85	1.23	43.64	0.984	0.0105
HuBERT	1.57	2.43	1.21	44.02	0.985	0.0102
HuBERT-CN (Ours)	1.49	2.34	1.20	44.13	0.985	0.0101

Table A. Ablation study on audio encoders. “CN” means model pretrained on Chinese data.

D. Additional Experiments

D.1. Additional Comparison with EmoTalk [6]

We further qualitatively compare our method with the blendshape-based method EmoTalk, using the style from the unseen identity. As is shown in Fig. C, our method performs better than EmoTalk. Besides, EmoTalk requires artists to produce blendshapes manually and does not involve dynamic textures.

D.2. Additional Ablation Study of Audio Encoders

We present an additional ablation study to show the influence of different audio encoders. Specifically, we quantitatively compare the metrics of models using MFCC, Wav2Vec2-CN¹, HuBERT and HuBERT-CN², where “CN” means encoder trained on Chinese data. Compared to other audio encoders, HuBERT can significantly improve generation quality. Meanwhile, encoders trained specifically on Chinese data can achieve slightly better results compared to the conventional version.

D.3. Generalization Discussion

As discussed in the limitation, our TexTalk4D dataset is mainly composed of Asian youths and only includes Mandarin speech. To study the generalization ability of the trained model, we directly test on unseen races, as is shown in Fig. D. Although our dataset is limited in diversity, the model generalizes well to other languages, races, and ages. Please refer to the sup. video for dynamic results.

¹<https://huggingface.co/TencentGameMate/chinese-wav2vec2-base>

²<https://huggingface.co/TencentGameMate/chinese-hubert-base>

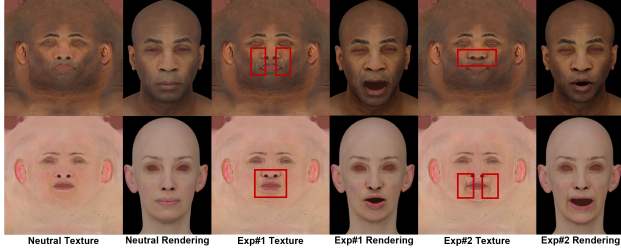


Figure D. Results on unseen races with styles from unseen IDs.

E. Ethics Discussion

In our dataset, all subjects have signed agreements authorizing us to use the collected data for research. We are committed to privacy protection to prevent the misuse of the collected data for criminal purposes. Specifically, we will only disclose the reconstructed assets and not the original captures. The dataset will only be available to researchers with professional titles who apply with an official email address. Further, all subjects reserve the right to revoke the authorization, and we will stop using the relevant data, but the research already conducted will not be affected.

F. Video Dynamic Results

To better demonstrate our work, we provide additional dynamic experimental results in the supplementary video. Specifically, we showcase several dynamic samples from our TexTalk4D dataset. For a more comprehensive evaluation of our method, we also provide more qualitative comparison results with the competitors [2, 5, 7–9] regarding facial motion and texture generation, highlighting the superiority of our approach over previous works. Additionally, the video also shows the dynamic results of our method in the disentangled control of speaking and wrinkling styles. It proves that the proposed pivot-based style injection method can effectively capture complex styles and is useful for achieving highly personalized dynamic facial animation. Finally, we present the facial animations driven by different languages.

References

- [1] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 1
- [2] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *CVPR*, pages 18770–18780, 2022. 3
- [3] Yao Feng. face3d: Python tools for processing 3d face. 1
- [4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 1
- [5] Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li. Dynamic facial asset and rig generation from a single scan. *TOG*, 39(6):215–1, 2020. 3
- [6] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *ICCV*, pages 20687–20697, 2023. 2
- [7] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–11, 2023. 3
- [8] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Mingjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *TOG*, pages 1–9, 2024. 1
- [9] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *CVPR*, pages 12780–12790, 2023. 3