-Supplementary Material-
LaughTalk: Expressive 3D Talking Head Generation with Laughter

The contents in this supplementary material are as follows: A. Training details for counterpart methods (Sec. A), B. Details of the emotional feature distance (Sec. B), and C. Details of the user study (Sec. C). We recommend viewing the supplementary video, which showcases generated 3D face animations from speech and laughter.

A. Training details for counterpart methods

As the pre-trained models of existing methods are not trained to capture the laughing expression, we retrain the existing methods [1, 3, 5] with our proposed LaughTalk dataset to ensure fair qualitative and quantitative comparison. Here, we provide the training details of the existing methods.

CodeTalker. CodeTalker [5] comprises two stages: the first stage involves learning generic motion through discrete tokens of a codebook, and the second stage focuses on generating 3D talking heads using these discrete tokens. Initially, we attempted to train only the second stage model with the LaughTalk dataset, while utilizing the pre-trained first stage model. However, due to the first stage model's lack of exposure to diverse and expressive talking heads, the model still failed to generate laughing expressions in the second stage. Therefore, we trained the first stage model on our dataset, encompassing diverse laughing expressions, and subsequently trained the second stage model with the same dataset. We followed the training scheme of the official code[1].

FaceFormer and VOCA. For FaceFormer [3] and VOCA [1], we initiated pre-training using LaughTalk$_{\text{MEAD}}$, which consists of neutral speech and corresponding 3D faces. Subsequently, fine-tuning was conducted on both models using LaughTalk$_{\text{CELEB}}$, featuring laughing and speech data. Notably, attempts to train these models using the entire LaughTalk dataset resulted in mode-collapsed outputs. The training process adhered to the official code of each method[2].

B. Details of the emotional feature distance

As discussed in the main paper, relying solely on measuring the lip vertex error (LVE) is insufficient to accurately assess facial movement synchronization to laughter. To address this limitation, we introduce Emotional Feature Distance (EFD) as a perceptual metric for evaluating laughter synchronization (Fig. S1). To compute the EFD, we utilize an off-the-shelf emotion recognition model, AffectNet [4]. Using this model, we measure the average feature distance between sequences of images rendered from the generated mesh vertices and the image frames sourced from the ground truth 2D video.

However, one challenge to note is the fixed pose inherent in existing 3D talking head generation methods (ours included), especially when compared to the diverse head movements present in the ground truth video frames. This discrepancy in head movement may result in misalignment between the generated meshes and the corresponding ground truth images, leading to less meaningful metric evaluations.

To mitigate this issue, we employ the iterative closest point (ICP) algorithm to align the generated meshes with the ground truth images (Fig. S1 (a)). Specifically, we begin by reconstructing a face mesh for each ground truth image using EMOCA [2]. The ICP algorithm then computes the rigid transformation matrix between the generated mesh and the mesh reconstructed from the ground truth images. This process is facilitated by the known correspondence between the vertex indices of the two meshes. The rigid transformation is subsequently applied to the generated mesh vertices, aligning them with the mesh of the ground truth image. We then proceed to texturize the aligned mesh with the texture map of the ground truth image and overlay it on top of the ground truth image (Fig. S1 (b)). Lastly, we feed both the rendered meshes and the ground truth images to the AffectNet and measure the l_2 distance between the extracted features, thus obtaining the EFD (Fig. S1 (c)).

C. Details of the user study

We conduct a user study to assess the performance of our method compared to the existing methods from a human perception standpoint. Our user study questionnaire interface is illustrated in Fig. S2. During the study, participants watched two generated 3D talking head videos and responded to four questions, without any time constraints. The user study comprises a total of 15 sets, each consisting

[1] https://github.com/Doubiiu/CodeTalker.

[2] https://github.com/EvelynFan/FaceFormer. https://github.com/TimoBolkart/voca.

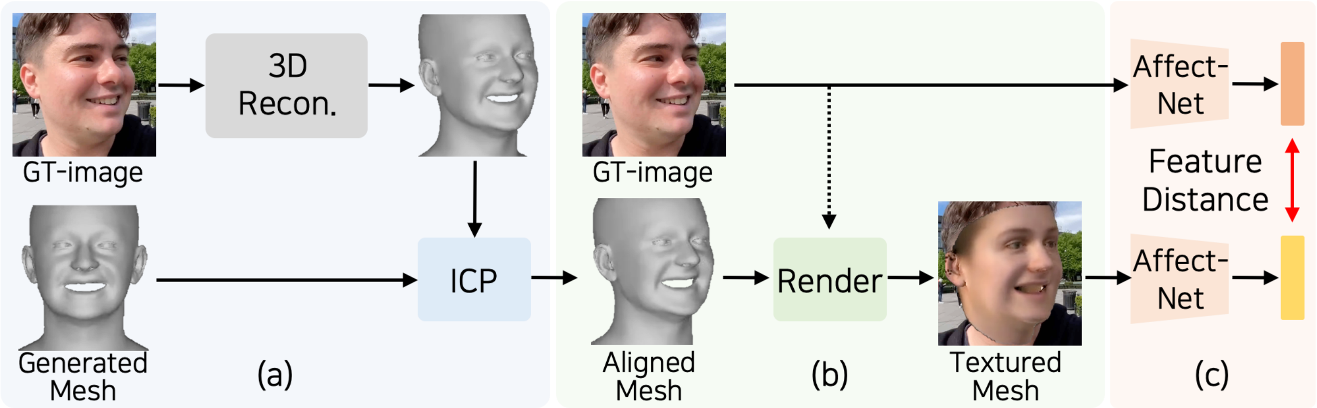
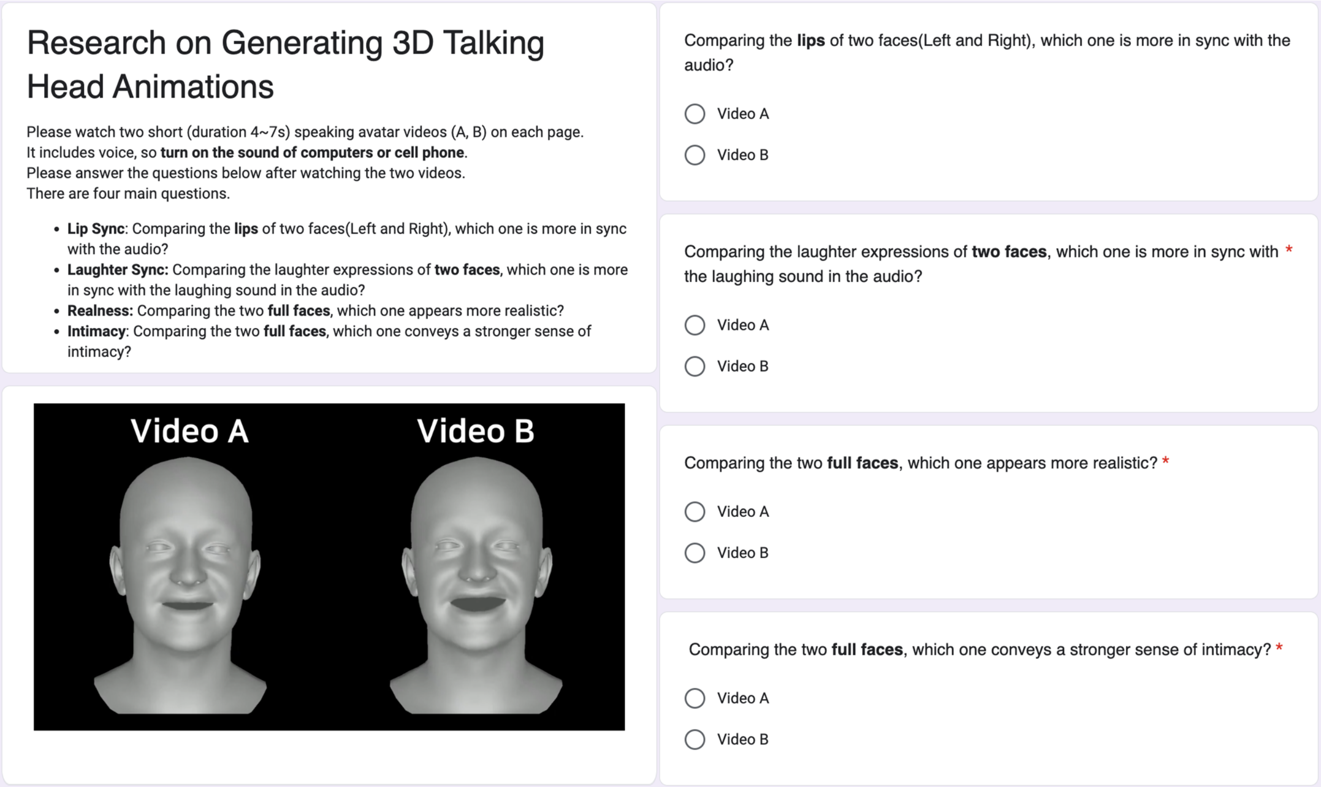


Figure S1. **Measuring emotional feature distance.** We present the Emotional Feature Distance (EFD), a perceptual metric designed to evaluate the synchronization of facial movements with laughter. To calculate this metric, first, (a) we align the speech-driven generated mesh vertices with the mesh reconstructed from the ground truth image of the original video. Next, (b) we render the generated mesh and texturize it using the texture map extracted from the ground-truth image. Finally, (c) we feed both the ground truth image and the textured mesh into AffectNet [4] and measure the ℓ_2 feature distance.



Research on Generating 3D Talking Head Animations

Please watch two short (duration 4~7s) speaking avatar videos (A, B) on each page.
It includes voice, so **turn on the sound of computers or cell phone**.
Please answer the questions below after watching the two videos.
There are four main questions.

- **Lip Sync**: Comparing the **lips** of two faces(Left and Right), which one is more in sync with the audio?
- **Laughter Sync:** Comparing the laughter expressions of **two faces**, which one is more in sync with the laughing sound in the audio?
- **Realness:** Comparing the two **full faces**, which one appears more realistic?
- **Intimacy**: Comparing the two **full faces**, which one conveys a stronger sense of intimacy?

Comparing the **lips** of two faces(Left and Right), which one is more in sync with the audio?

○ Video A

○ Video B

Comparing the laughter expressions of **two faces**, which one is more in sync with the laughing sound in the audio? *

○ Video A

○ Video B

Comparing the two **full faces**, which one appears more realistic? *

○ Video A

○ Video B

Comparing the two **full faces**, which one conveys a stronger sense of intimacy? *

○ Video A

○ Video B

Figure S2. **Example of a user study experiments.** Each page contains a pair of generated 3D talking head videos for comparative analysis accompanied by four questions designed to assess the performance of our model.

of 2 videos and featuring four questions in each set. Our study includes 50 participants, encompassing individuals both within and outside the research field. The questions we ask to the participants are as follows:

- Lip Sync: Comparing the lips of two faces (Left and Right), which one is more in sync with the audio?
- Laughter Sync: Comparing the laughter expressions of two faces, which one is more in sync with the laughing sound in the audio?

- Realness: Comparing the two full faces, which one appears more realistic?
- Intimacy: Comparing the two full faces, which one conveys a stronger sense of intimacy?

References

[1] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017.

[5] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.