

LaughTalk: Expressive 3D Talking Head Generation with Laughter

Kim Sung-Bin¹ Lee Hyun¹ Da Hye Hong²
 Suekyeong Nam³ Janghoon Ju³ Tae-Hyun Oh^{1,4,5}

¹Dept. of Electrical Engineering and ⁴Grad. School of Artificial Intelligence, POSTECH

²Sookmyung Women’s University ³KRAFTON

⁵Institute for Convergence Research and Education in Advanced Technology, Yonsei University

<https://laughtalk.github.io/>

Abstract

Laughter is a unique expression, essential to affirmative social interactions of humans. Although current 3D talking head generation methods produce convincing verbal articulations, they often fail to capture the vitality and subtleties of laughter and smiles despite their importance in social context. In this paper, we introduce a novel task to generate 3D talking heads capable of both articulate speech and authentic laughter. Our newly curated dataset comprises 2D laughing videos paired with pseudo-annotated and human-validated 3D FLAME parameters and vertices. Given our proposed dataset, we present a strong baseline with a two-stage training scheme: the model first learns to talk and then acquires the ability to express laughter. Extensive experiments demonstrate that our method performs favorably compared to existing approaches in both talking head generation and expressing laughter signals. We further explore potential applications on top of our proposed method for rigging realistic avatars.

1. Introduction

Speech-driven 3D facial animation has garnered increasing attention from both academic researchers and industries due to its practical applications. This field offers promising applications in content creation, including computer gaming, film production [24], and immersive interactions between humans and machines in virtual realities [41]. Following the practicality, recent advancements in deep learning techniques have yielded impressive outcomes in

Acknowledgment. This work was supported by IITP grant funded by Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub; No.RS-2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation; No.2022-0-00290, Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense; No.2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities).

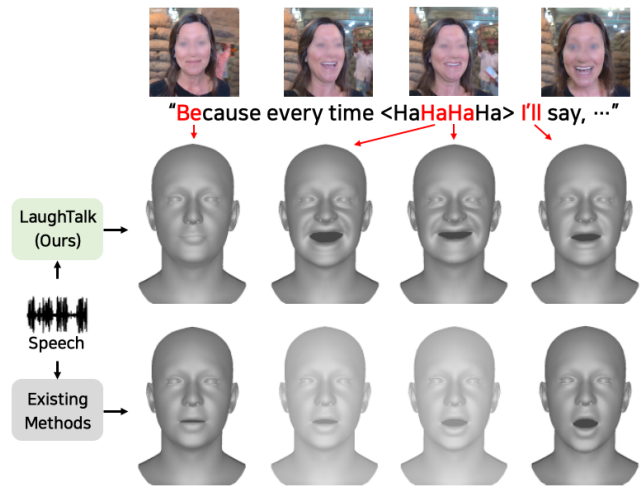


Figure 1. **Learning to laugh and talk.** We present a task for generating speech-driven 3D talking heads capable of conveying both verbal and non-verbal expressions. In contrast to existing methods that primarily focus on achieving accurate lip synchronization to verbal signals, our goal is to simultaneously articulate synchronized lip movements while also expressing synchronized laughter.

generating high-quality animated speech-driven 3D talking heads [8, 11, 21, 29, 34, 43].

However, these prior arts predominantly focus on achieving accurate lip synchronization to verbal signals in speech, often neglecting non-verbal signals, which play vital roles not only in human interactions [17, 25, 30] but also in human-computer interaction [22]. Among the non-verbal cues, laughter, in particular, holds significance as it is evoked to establish intimacy [32], rapport [1], trust [37], and create deep emotional exchanges [31, 36]. Furthermore, it serves as a powerful medium for conveying diverse social and emotional nuances beyond verbal expressions. Thus, learning to synthesize speech-driven 3D facial animations incorporating non-verbal expressions, including laughter, may open up the

development of intimate and empathetic machines capable of engaging in human-robot interactions.

Nevertheless, accomplishing the animation of 3D facial expressions for both speech and laughter presents intricate challenges. Firstly, the scarcity of comprehensive datasets encompassing both speech and laughter poses a hurdle. Existing 3D scan and speech-paired datasets, such as VOCASET [8] and BIWI [12], are derived from controlled lab environments, lacking the diversity and naturalness required for real-world scenarios. Secondly, verbal and non-verbal cues tend to be entangled within speech. This intermingling complicates the extraction of explicit verbal and non-verbal cues from speech, thus making it challenging to teach the 3D talking head to express them simultaneously. Addressing these challenges is pivotal for solving the 3D facial animation encompassing both speech and laughter.

In this work, we introduce a novel task to animate a 3D face from speech, conveying both talking and laughter. Our primary focus is learning the model to simultaneously articulate the synchronized accurate lip movements and synchronized laughter (Fig. 1). For this task, we collect the **LaughTalk dataset**, comprising in-the-wild 2D videos that feature diverse and natural speech along with laughter, and their corresponding 3DMM (FLAME) [23] parameters. Despite the abundance of extensive 2D video datasets [6,45,47], parsing videos containing both speech and laughter remains non-trivial. We devise a data curation pipeline capable of parsing laughing and talking videos while filtering out noisy ones, such as videos with non-active speakers, scene changes, and abrupt head movements. Since these 2D videos do not contain 3D information, we leverage 3D face reconstruction models [9, 14] and exemplar fine-tuning methods [20, 44] to generate reliable and robust pseudo ground truth FLAME parameters corresponding to the collected 2D videos.

Given this dataset, we design a baseline model for our task, called **LaughTalk**. LaughTalk adopts a two-stage training procedure, initially learning speech articulation and subsequently learning to express laughter. The model trained in the first stage (stage-1) entails extracting audio features from a pre-trained audio encoder [3] and training a Transformer [38] decoder to regress the FLAME parameters from the audio features. This training employs the subset from the LaughTalk dataset, $\text{LaughTalk}_{\text{MEAD}}$ ², which comprises neutral speech and facial movements, to focus solely on verbal cues. In the second stage, a separate model is trained to regress the residual FLAME parameters that the first stage model has not learned. The separate model used in the second stage (stage-2) has an identical architecture to the model used in the first stage, but its weights are not shared. Here, we employ another subset from the LaughTalk dataset, $\text{LaughTalk}_{\text{CELEB}}$ ³, containing both speech and laugh-

ter. Given that the stage-1 model has already learned to convey speech articulations, the stage-2 model focuses on learning to express non-verbal signals. Combining the FLAME parameters derived from both stage models facilitates the simultaneous representation of verbal and non-verbal cues within the 3D facial animation.

We validate the efficacy of LaughTalk by comparing it with existing 3D talking head models [8, 11, 29, 43]. For a fair comparison, we train existing models with our proposed dataset, also allowing them to learn to talk and laugh. We leverage a pre-trained emotion feature extractor [26] to assess whether the model conveys synchronized non-verbal cues and measure the lip vertex error to evaluate speech articulation. Our experiments demonstrate that LaughTalk not only excels in generating synced laughter but also exhibits favorable lip articulation compared to existing methods. Moreover, we showcase the practical application of LaughTalk, further underscoring its potential in real-world scenarios. Our main contributions are summarized as follows:

- Introducing a task to generate a 3D talking head that simultaneously expresses speech articulation and laughter.
- Collecting and curating the LaughTalk dataset, which includes 2D videos of speech and laughter along with their corresponding pseudo ground truth FLAME parameters.
- Proposing a baseline, LaughTalk, and a two-stage training scheme for learning to animate expressive 3D faces with verbal and non-verbal signals from speech.

2. Related work

Speech-driven 3D facial animation. There has been a lot of research into generating 3D talking heads for 3D gaming and virtual reality applications. Seminal works [4, 8, 10, 11, 18, 21, 28, 29, 35] have shown promising results in speech-driven 3D facial animation, particularly in generating verbal articulations in sync with input speech. FaceFormer [11] uses a Transformer-based model for the first time to autoregressively generate facial movements based on speech input. CodeTalker [43] learns a discrete codebook for generic facial motion and leverages a similar architecture from FaceFormer for animation synthesis. However, they mainly focus on the mouth movement, which is related to the verbal context of speech. For enhancing facial expression, Meshtalk [29] aims to capture upper face movements by designing separate latent codes for audio-related and non-audio-related movements, such as blinking and eyebrow raises. The aforementioned methods mainly focus on the verbal feature of speech or speech-uncorrelated movements but overlook addressing non-verbal expression arising from speech, which often carries the emotional and social context.

In more recent works, EMOTE [10] and EmoTalk [28] introduce emotional 3D talking heads. EMOTE requires explicit condition inputs to the model, such as emotion cate-

²The video clips are curated from MEAD [39].

³The video clips are curated from CelebV-HQ [47] and CelebV-Text [45].

gories (*e.g.*, angry, surprise, sad, etc.). EmoTalk alleviates this explicit conditioning by deriving emotion from speech but requires an artist-curated dataset for training. While these models can generate verbal articulation incorporated with emotion, they cannot capture non-verbal signals, such as laughter, from speech. In contrast, we focus on generating an expressive 3D talking head that embraces both non-verbal (*i.e.*, laughter) and verbal signals deriving directly from the speech and provide a new dataset suitable for this task.

Face video datasets. One of the challenges in speech-driven 3D facial animation is the lack of 3D paired datasets. BIWI [12], VOCASET [8], S3DFM [46], and Multiface [42] are publicly available datasets for this task. Although these datasets provide accurate 3D meshes scanned in lab environments, they are limited in size, diversity, speaking styles, and naturalness. On the other hand, many 2D face video datasets [2, 6, 7, 19, 27, 45, 47] are available. Most of them are curated from in-the-wild video sources, *e.g.*, YouTube, and have richness in size, diversity, and naturalness.

Our newly curated dataset is derived from existing 2D face video datasets. We annotate these videos with reliable pseudo 3DMM parameters, by leveraging the techniques from 3D face reconstruction methods [9, 13, 14]. We believe that utilizing extensive 2D facial video datasets with 3D reconstruction methods is advantageous for building a rich dataset that can be used for training 3D talking heads conveying a wide range of verbal and non-verbal expressions.

Laughter in non-verbal signals. Among diverse non-verbal signals, we especially focus on laughter, which plays a key role in social interactions, such as building rapport, expressing emotions, and creating deep emotional exchanges [31, 36]. Furthermore, laughter is a distinct human expression not found in other animals, which we often use to establish intimacy [32], grab attention [40], or build trust [37]. Understanding and synthesizing laughter is thus a crucial stepping stone toward creating expressive and intimate interactive agents that go beyond verbal-only communication.

3. Learning to laugh and talk

In this section, we provide a concise overview and preliminary for our proposed task (Sec. 3.1) and introduce the LaughTalk dataset comprising paired 2D video and 3DMM parameters (Sec. 3.2). Subsequently, we propose LaughTalk, a two-stage training baseline model capable of both laughing and talking synchronized to the given speech (Sec. 3.3).

3.1. Overview

Our goal is to synthesize a sequence of 3D face animations from given speech audio, encompassing both speech and laughter. In contrast to the explicit conditioning on global emotion labels, *e.g.*, [10], we drive the lips and expressions of 3D faces synchronized to the talking (verbal)



Figure 2. **LaughTalk Dataset.** We collect and curate 2D talking and laughing videos with corresponding pseudo 3DMM parameters. Here, we show the images sampled from the 2D videos and corresponding mesh images rendered from the 3DMM parameters.

and instantaneous laughing (non-verbal) cues in the speech. For instance, when the speech is delivered with a neutral tone, the 3D faces focus solely on animating the lips with neutral expressions. Conversely, during laughing in the speech, the 3D facial features would synchronously animate, depicting the characteristic upward shift of expression associated with laughter. As we take the pioneering step, we aim to build an appropriate 3D face dataset and baseline model for this task.

Preliminary. We use FLAME [23], a parametric 3D head model, as the 3D human face representation for our collected dataset and the 3D talking head. Given the parameters of the face shape $\beta \in \mathbb{R}^{|\beta|}$, facial expression coefficients $\psi \in \mathbb{R}^{|\psi|}$, and pose $\theta \in \mathbb{R}^{3k+3}$ ($k=4$ joints), 3D face mesh with vertices $\mathbf{V} \in \mathbb{R}^{n_v \times 3}$ ($n_v=5023$) and 3D facial landmarks $\mathbf{J}^{3D} \in \mathbb{R}^{n_j \times 3}$ ($n_j=68$) can be acquired with a differentiable FLAME model M , where $[\mathbf{V}, \mathbf{J}^{3D}] = M(\beta, \psi, \theta)$.

3.2. LaughTalk dataset

We introduce the LaughTalk dataset, comprising talking and laughing facial video clips with corresponding pseudo-annotated FLAME parameters. We first curate 2D video clips from MEAD [39] for neutral talking (called LaughTalk_{MEAD}), and CelebV-HQ [47] and CelebV-Text [45] for capturing talking with laughter (called LaughTalk_{CELEB}). Then, we filter out the clips to select valid samples and finally obtain 3D pseudo-annotation from them. Examples of our curated dataset are shown in Fig. 2.

Data collection. CelebV-HQ and CelebV-Text have in-the-wild face videos with rich facial attributes, such as appearance, emotion, and action. Among the attributes, we query “laugh”, “smile”, “happy”, and “talk” to construct laughing and talking video clips for LaughTalk_{CELEB}. Similarly, MEAD provides the annotations of emotion attributes and their intensities. We query the “neutral” attribute in MEAD and collect neutral talking video clips for LaughTalk_{MEAD}.

Statistics of our LaughTalk dataset	
Number of Total Video Clips	943
Number of LaughTalk _{MEAD}	438
Number of LaughTalk _{CELEB}	505
Number of Train / Test of LaughTalk _{MEAD}	374 / 64
Number of Train / Test of LaughTalk _{CELEB}	455 / 50
Average Duration of Train set	3.5 sec.
Average Duration of Test set	5.6 sec.

Table 1. **Statistics of the LaughTalk dataset.** We trim training set video clips to 3.5 seconds but leave the test set with varying lengths for evaluation on diverse inputs. Thus, the average length of the test set video is longer than the training set.

Data filtering process. After the data collection, we filter out noisy samples to construct a valid and clean dataset. First, to ensure that video clips always contain laughter, we use a laugh detector [15] to filter out samples that do not have laughter for at least 3.5 seconds. This reduces false samples from incorrect attribute annotations in the original datasets. Second, our dataset must include talking faces. However, some video clips contain speech from outside the scene. Thus, we filter out the videos with non-active speakers using the active speaker detector [33], ensuring our dataset holds the facial video that synchronizes with speech. Third, we use the scene detector [5] to trim video clips at scene transitions, thereby preventing the inclusion of scene change videos. Only videos longer than 3.5 seconds are considered in this process. Lastly, we exclude video samples in which the individual’s face is not visible from the front, only a partial view of the face is shown, or there are abrupt head movements. We retain only clear, frontal facial shots.

Lifting 2D video to 3D. After acquiring the cleaned 2D in-the-wild videos, we reconstruct 3D faces synchronized with both the audio and facial movements from the video clips. However, existing 3D face reconstruction models [9, 13] have limitations for reconstructing temporally consistent and accurate 3D face meshes from videos. State-of-the-art face reconstruction models are typically trained exclusively on static 2D images. This results in limitations when extrapolating to faces in rare poses and produces jittered motion due to per-frame independent inference. To address these challenges, we employ an optimization method [44] that re-parameterizes 3D face meshes with neural network parameters, inspired by EFT [20]. We initialize the neural network with SPECTRE [14] and optimize it for each video clip, ensuring the acquisition of accurate and robust pseudo ground truth FLAME parameters. This approach is suitable for our purpose as it yields an accurate 3D face reconstruction result that best fits each video clip.

After all the processing, we have 943 video clips and corresponding pseudo-annotated FLAME parameters. We set separate training and test sets for each sub-dataset (LaughTalk_{MEAD}, LaughTalk_{CELEB}). The basic statistics for our dataset are summarized in Table 1.

3.3. Two-stage training baseline: LaughTalk

Since verbal and non-verbal signals are often intertwined within a single speech, teaching a 3D talking head model to animate both laughter and speech simultaneously is challenging. To address this, we approach the task by breaking it down into sub-problems. The schematic overview of our proposed baseline model, LaughTalk, is presented in Fig. 3.

LaughTalk undergoes a two-stage training strategy. First, the stage-1 model learns to talk (*i.e.*, verbal signals from speech). It focuses on learning to generate facial representations for lip movement synchronization with neutral speech videos (LaughTalk_{MEAD}). After the stage-1 model has acquired the ability to animate speech-related movements, we then progress to train the stage-2 model using LaughTalk_{CELEB}, aiming to animate both lip movements and facial expressions simultaneously. As we freeze the parameters of the stage-1 model, the stage-2 model focuses on learning to generate residual facial representations that are not learned by the stage-1 model. These residual aspects likely correspond to non-verbal cues present in the speech, such as cheek movements and facial expressions. By combining the residual representations with the output of the pre-trained stage-1 model, LaughTalk can simultaneously animate a 3D talking head with synchronized verbal and non-verbal signals. We now describe the brief task formulation, details of our model’s architecture, and the training objectives for each stage.

Task formulation. Let $\mathbf{F}_{1:T} = (\mathbf{f}_1, \dots, \mathbf{f}_T)$ be a temporal sequence of facial motions, with each frame \mathbf{f}_t denoting the facial representation. Here, we define the facial representation \mathbf{f}_t as the concatenation of the expression coefficient and jaw pose of the FLAME parameters, $\mathbf{f}_t = [\psi_t, \theta_t^{\text{jaw}}]$. Additionally, let $\mathbf{A}_{1:T} = (\mathbf{a}_1, \dots, \mathbf{a}_T)$ be a sequence of speech representations. Then, our goal is to sequentially synthesize the facial representations $\mathbf{F}_{1:T}$ from the corresponding $\mathbf{A}_{1:T}$. To visualize $\mathbf{F}_{1:T}$ as mesh vertices, these facial representations are fed into the FLAME model M along with arbitrary face shape parameter β , generating animated vertices as $\mathbf{V}_{1:T} = M(\beta, \mathbf{F}_{1:T})$.

Stage-1: learning to talk. The stage-1 model is designed to learn to animate the 3D faces, primarily capturing speech-related signals. This model comprises the Verbal Encoder E_v , which extracts speech-related features from the input audio, and the Transformer Decoder D_v , which takes the speech-related features and generates sequences of facial representations in an autoregressive manner (Stage-1 in Fig. 3).

Following Faceformer [11], we employ wav2vec 2.0 [3] for the Verbal Encoder. This includes the audio feature extractor and multi-layer transformer encoder. The audio feature extractor utilizes a temporal convolutional network (TCN) to convert the raw waveform of speech into feature vectors. Then, the Transformer Encoder transforms the au-

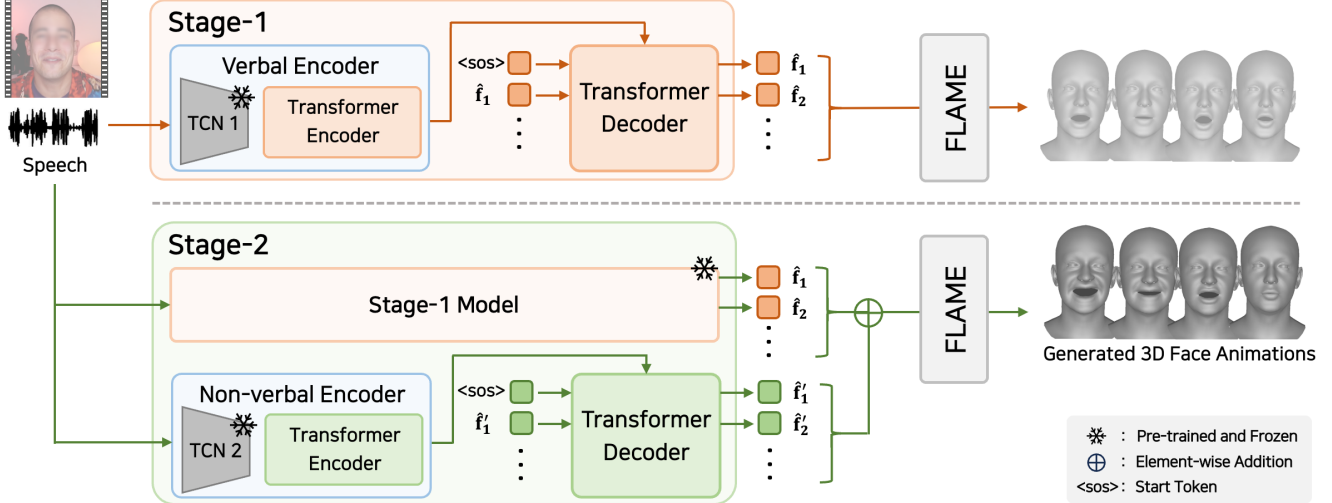


Figure 3. **LaughTalk architecture.** The stage-1 model extracts verbal features from the input speech and generates facial motion representations in an autoregressive manner. Simultaneously, the stage-2 model extracts non-verbal features from the same speech and generates residual facial motion representations. These two sets of representations are element-wise added and subsequently fed into the FLAME model to synthesize 3D face animations.

dio features into speech representations. The Transformer Decoder is equipped with causal self-attention to learn the dependencies within the context of previous facial representations and employs cross-modal attention to align the audio and facial representations. Formally, this process can be written as: $\hat{\mathbf{f}}_t = D_v(E_v(\mathbf{A}_{1:T}), \hat{\mathbf{F}}_{1:t-1})$, where $\hat{\mathbf{f}}_t$ is the currently predicted facial representation, and $\hat{\mathbf{F}}_{1:t-1}$ is the past predicted sequences. After predicting all the sequences $\hat{\mathbf{F}}_{1:T}$, we feed these along with arbitrary shape parameter β to the FLAME model M to convert them to the mesh vertices and 3D landmarks, $[\hat{\mathbf{V}}_{1:T}, \hat{\mathbf{J}}_{1:T}^{3D}] = M(\beta, \hat{\mathbf{F}}_{1:T})$.

We train the Transformer Encoder and Decoder of the stage-1 model while keeping TCN frozen with the pre-trained weights of wav2vec 2.0. To ensure that the model learns to animate the facial movements by mainly focusing on the verbal signals, we train it using LaughTalk_{MEAD}, which contains speech delivered with neutral tones. The training objective for the stage-1 model is:

$$L_{\text{stage-1}} = \lambda_{\text{exp}} \|\psi_{1:T} - \hat{\psi}_{1:T}\|_2 + \lambda_{\text{lmk}} \|\mathbf{J}_{1:T}^{3D} - \hat{\mathbf{J}}_{1:T}^{3D}\|_2, \quad (1)$$

where $\hat{\psi}_{1:T}$ is the expression parameters from the predicted facial representations $\hat{\mathbf{F}}_{1:T}$, $\hat{\mathbf{J}}_{1:T}^{3D}$ is the predicted 3D facial landmarks, and $\{\lambda_*\}$ denotes the weights for each loss term.

Stage-2: learning to laugh. The stage-2 model incorporates the pre-trained stage-1 model and is designed to learn to animate 3D faces with both speech and laughter. As we freeze the stage-1 model, the stage-2 model is induced to focus on learning to generate residual facial representations that the stage-1 model has not learned. We postulate that such residual representations likely correspond to non-verbal cues within speech. In other words, the stage-2 model

learns to predict the residual facial representations $\hat{\mathbf{F}}'_{1:T}$, which would make the 3D face express both verbal and non-verbal signals, when combined with the outputs of the stage-1 model $\hat{\mathbf{F}}_{1:T}$ (Stage-2 in Fig. 3).

The architecture of the stage-2 model is identical to that of the stage-1, comprising the Non-verbal Encoder E_n and the Transformer Decoder D_n . Different from stage-1, the stage-2 model uses wav2vec 2.0 for the Non-verbal Encoder, initialized with the pre-trained weights of the emotion recognition model⁴.

We train the Transformer Encoder and Decoder on LaughTalk_{CELEB}, which contains in-the-wild audio samples with both laughter and speech, while keeping TCN and the stage-1 model frozen. The training objective for the stage-2 model is:

$$L_{\text{stage2}} = \|\mathbf{V}_{1:T} - \hat{\mathbf{V}}'_{1:T}\|_2, \quad (2)$$

where $\mathbf{V}_{1:T}$ and $\hat{\mathbf{V}}'_{1:T}$ are the sequence of the ground truth and predicted vertices. The predicted vertices are made by summing the FLAME parameters from the pre-trained stage-1 model and the residual part of the stage-2 model. Formally, we denote $\hat{\mathbf{V}}'_{1:T}$ as $\hat{\mathbf{V}}'_{1:T} = M(\beta, \hat{\mathbf{F}}_{1:T} + \hat{\mathbf{F}}'_{1:T})$.

To dissect the impact of the residual representations learned within the stage-2 model, we convert the facial representation generated by each model into mesh vertices. Figure 4 visually illustrates the outputs of the stage-1, residual, and the final stage-2 model. While the outputs of the stage-1 model are geared towards animating speech-related facial motions, the residual outputs contribute to producing more

⁴<https://huggingface.co/harshit345/xlsr-wav2vec-speech-emotion-recognition>.

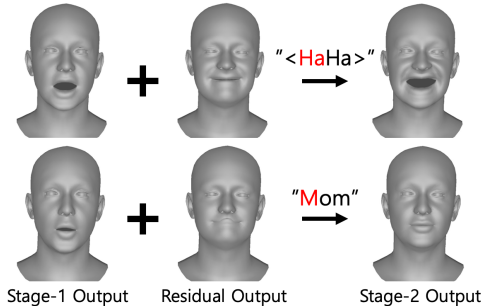


Figure 4. **Visualization of each model’s output.** We visualize the output meshes synthesized by the models from each stage. The stage-1 model outputs the lip movements for the input speech, while the stage-2 model complements the expression of non-verbal signals, and contributes to more accurate lip movements.

expressive 3D faces. These residual representations serve the dual purpose of conveying the non-verbal laughter signal while also enhancing lip articulations. These findings support the efficacy of the two-stage training approach, successfully animating a 3D talking head with speech and laughter simultaneously.

Training details. We train the stage-1 model on a single GeForce RTX 3090 for 100 epochs with early stopping. We use the Adam optimizer, and set the batch size to 1 and the learning rate to $2 \cdot 10^{-4}$. After training, we freeze the stage-1 model and train the stage-2 model with the same experiment setup. For both stages, we randomly sample 100 frames from the input data for training.

4. Experiments

We evaluate the performance of LaughTalk using our proposed dataset, as detailed in Sec. 3.2. We begin with a quantitative assessment to measure the accuracy of both lip articulation and laughter synchronization in relation to the input speech (Sec. 4.1). We then conduct qualitative comparisons against existing methods (Sec. 4.2). Most importantly, we conduct user studies to validate the human perceptual experience of the generated 3D talking head, encompassing both verbal and non-verbal signals (Sec. 4.3). Finally, we show ablation studies to verify our design choices of the model training (Sec. 4.4).

4.1. Quantitative evaluation

To evaluate the lip synchronization to the speech, we measure the lip vertex error (LVE) metric proposed in MeshTalk [29]. The LVE computes the average ℓ_2 error between the lip regions of the generated mesh vertices and the ground truth from the test set. For each frame, the LVE is defined as the maximum ℓ_2 error across all lip vertices⁵.

⁵We use the lip vertex indexes which are provided at <https://flame.is.tue.mpg.de/>.

Method	LaughTalk _{MEAD}	LaughTalk _{CELEB}	
	LVE (↓)	LVE (↓)	EFD (↓)
VOCA [8]	2.45	2.17	17.71
FaceFormer [11]	2.15	<u>1.81</u>	<u>17.51</u>
CodeTalker [43]	2.81	2.22	19.80
LaughTalk (Ours)	<u>2.27</u>	1.76	17.37

Table 2. **Quantitative comparison to existing methods.** We compare LaughTalk (Ours) with existing methods trained on the LaughTalk dataset. The results show that LaughTalk performs favorably in the lip vertex error (LVE), while outperforming other methods in the emotion feature distances (EFD). The measurement scale for LVE is $\times 10^{-4}$ mm scale. We highlight the best results in **bold** and underline the second best among all the methods.

However, solely measuring the ℓ_2 error of lip vertices is insufficient to assess the facial movement synchronization to the laughter, as there is no one-to-one mapping between input audio and the generated 3D faces. To address this, we introduce the emotional feature distance (EFD) as an additional metric to assess laughter synchronization. This metric leverages AffectNet [26], an emotion recognition model. Specifically, using AffectNet, we extract the emotion features from both the sequence of rendered images and the frames from ground truth 2D video, then compute the frame-by-frame feature ℓ_2 distance. We average these feature distances to determine the EFD. Our rationale behind this approach is that achieving a lower feature distance may suggest closer temporal and semantic alignments of facial expressions between the generated mesh and the ground truth 2D video, indicating improved synchronization.

We conduct a quantitative comparison of our approach against three state-of-the-art methods, VOCA [8], FaceFormer [11], and Codetalker [43]. To ensure a fair comparison, we retrain these models using the mesh vertices from our LaughTalk dataset. Table 2 summarizes the LVE and EFD over the test set of the LaughTalk dataset. Notably, LaughTalk performs favorably compared to the other methods across all test sets. Particularly for the LaughTalk_{CELEB} test set, encompassing both speech and laughter, our approach demonstrates superior performance in terms of both LVE and EFD metrics. This result highlights the effectiveness of our proposed method, LaughTalk, in achieving accurate synchronization of both lip movements and laughter with the corresponding speech.

4.2. Qualitative evaluation

Since evaluating speech-driven 3D facial movements based on only quantitative metrics might not capture the full scope of their quality, we visualize the synthesized meshes for more comprehensive evaluation. In Fig. 5, we present a visual comparison between the meshes generated by our approach and existing methods.

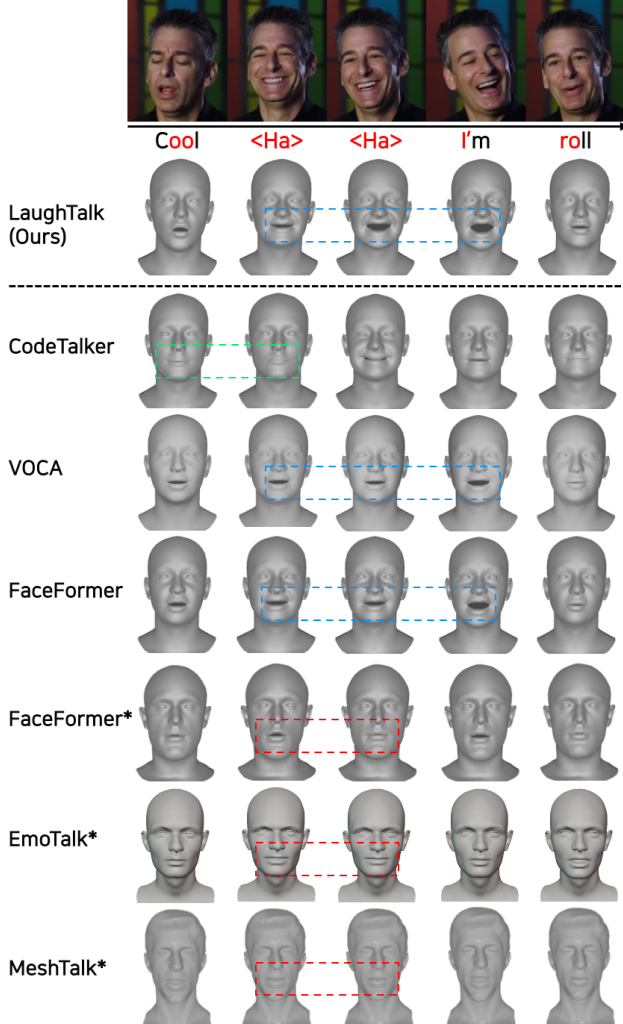


Figure 5. **Qualitative comparison to existing methods.** Each row shows specific frames synthesized by LaughTalk (Ours) and counterpart methods. The models presented in sequence, Ours, CodeTalker, VOCA, and FaceFormer, are trained with our proposed LaughTalk dataset. The models with superscript *, FaceFormer*, EmoTalk*, and MeshTalk*, are the official pre-trained models.

The top three rows display results from the CodeTalker, VOCA, and FaceFormer models trained on the LaughTalk dataset, where we observe subtle smiles synchronized with the laughter in the speech. However, CodeTalker struggles to accurately animate lip movements aligned with the verbal signal (green dotted line in Fig. 5), while VOCA and FaceFormer exhibit less expressive results compared to our approach (blue dotted line in Fig. 5). The bottom three rows showcase the results from models using their pre-trained weights. These primarily exhibit facial movements aligned with the verbal cues, yet neglect the laughter signal, as highlighted by the red dotted lines in Fig. 5. This demonstrates that our two-stage training method enables the model to produce both verbal and non-verbal signals simultaneously.

Competitors (vs. B)	Lip Sync	Laughter Sync	Realism	Intimacy
vs. FaceFormer	88.00	82.67	86.00	84.67
vs. CodeTalker	78.00	78.00	82.00	80.00
vs. FaceFormer*	80.67	93.33	81.33	84.67
vs. CodeTalker*	67.33	92.00	79.33	82.00
vs. EmoTalk*	72.67	86.67	66.67	81.33

Table 3. **User study results.** We employ A vs. B testing and report the percentage (%) of responses where A (Ours) is preferred over B. A higher percentage indicates better performance of our method. The user study is conducted to evaluate the generated meshes in four aspects: lip sync, laughter sync, realness, and intimacy. The models with superscript *, FaceFormer*, CodeTalker*, and EmoTalk*, are the official pre-trained models.

4.3. User study

We provide the user study to evaluate our proposed model, LaughTalk. Given that the human perception system has evolved to effectively understand subtle facial motions and capture lip articulation, employing user studies stands as a reliable measure for assessing the quality of speech-driven facial animation. We generate fifteen 3D talking head videos using our method (A) and other methods (B) from the test split of LaughTalkCELEB and design a user questionnaire based on A vs. B testing. This questionnaire prompts users to choose between two samples based on four distinct aspects: lip synchronization, laughter synchronization, realism, and intimacy. Notably, the ‘‘Intimacy’’ aspect assesses which generated facial animation elicits a stronger sense of intimacy for human-computer interaction [22]. A total of 50 participants take part in this user study.

We compare our model with existing methods trained on the LaughTalk dataset and official pre-trained models. The user study results are summarized in Table 3, indicating that the participants favor the generated results of LaughTalk over counterpart methods. Particularly in the ‘‘Laughter Sync’’ and ‘‘Intimacy’’ evaluations, our model gets significant preference over the competing methods. We believe this preference is attributed to two key factors. First, our two-stage training design accurately articulates lip movements in the first stage and learns residual features in the second stage to enhance expressiveness and convey non-verbal signals. Second, our model’s ability to achieve accurate lip synchronization while conveying non-verbal signals potentially results in users feeling greater intimacy towards the generated meshes, underscoring the importance of effectively conveying non-verbal signals in 3D talking head models.

4.4. Ablation study

We conduct a series of ablation studies to validate our design choices and assess the effectiveness of our proposed dataset, as summarized in Table 4. Note that, except for our

Configurations	LaughTalk _{MEAD}	LaughTalk _{CELEB}	
	LVE (↓)	LVE (↓)	EFD (↓)
w/o LaughTalk _{CELEB}	2.61	4.65	20.33
w/o LaughTalk _{MEAD}	2.54	2.24	18.56
w/o two-stage training	2.43	1.86	18.42
LaughTalk (Ours)	2.27	1.76	17.37

Table 4. **Ablation studies for our design choices.** We evaluate on test set of LaughTalk_{MEAD} and LaughTalk_{CELEB}. While the first two configurations differ in terms of the training set, the subsequent configuration serves as an ablation for the two-stage training strategy. The measurement scale for LVE is $\times 10^{-4}$ mm scale.

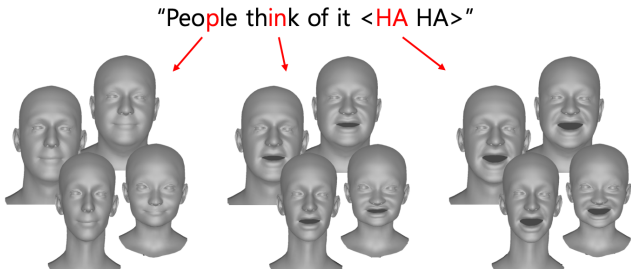


Figure 6. **User controllable identities.** As we use the FLAME parameters for facial representations, our model can synthesize 3D talking heads with diverse identities by changing the β parameter.

final LaughTalk model, all the ablation models are trained only in stage-1 with Eq. (1). Regarding the dataset aspect, using LaughTalk_{CELEB} yields significant improvement in both LVE and EFD metrics. This indicates that the data from expressive in-the-wild videos helps the model learn to animate both talking and laughing. While utilizing the complete LaughTalk dataset could potentially yield even better performance, training with all the data at once might be complex due to the intertwined verbal and non-verbal signals. Thus, as proposed, a two-stage training framework addresses these challenges and yields the best overall results in the evaluated metrics.

5. Applications

Unlike existing methods that utilize mesh vertex representations for 3D talking head models [8, 11, 21, 29, 43], our approach employs FLAME parameters to represent facial motion. While the prior arts are limited to a fixed number of pre-defined identity templates for 3D talking heads, the use of FLAME parameters in our method offers more distinctive and user-friendly control. Moreover, these parameters often serve as control signals for rigging neural avatars. This opens up interesting and practical applications as follows.

User controllable identities. Given the flexibility of our model in defining the shape parameter β , we can easily manipulate or even input β extracted from a reference image.

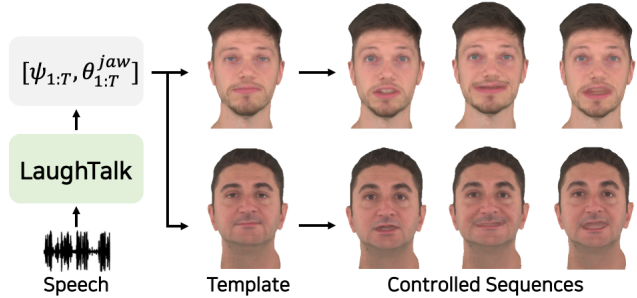


Figure 7. **Rigging neural avatars.** Our model can be used for rigging the photorealistic neural avatar from speech. Here, $\psi_{1:T}$ and $\theta_{1:T}^{jaw}$ denote the sequence of expression and jaw pose parameters predicted by our model from input speech. These predicted parameters serve as control signals for NHA [16].

As shown in Fig. 6, with the same input speech, our model can synthesize 3D talking heads with diverse identities. Combined with the 3D reconstruction methods [13, 48], we can finely control the identity of the 3D talking head by users.

Rigging neural avatars. In addition to synthesizing 3D talking meshes, our model also offers FLAME parameters for avatar rigging. Recent studies have shown that the FLAME parameters can effectively serve as control signals for rigging photorealistic neural avatars [16, 49]. Figure 7 showcases several examples produced by combining our approach with NHA [16]. The expression and jaw pose parameters that our model predicts are directly fed into NHA to control the avatar. The results clearly illustrate that our method serves as a conduit for rigging photorealistic avatars with diverse speech inputs.

6. Conclusion

In this work, we introduce a novel task of animating a speech-driven 3D face, capable of expressing both verbal and non-verbal signals. We especially focus on laughter as a non-verbal signal, given its importance in social interactions. To approach this task, we curate the LaughTalk dataset, consisting of diverse in-the-wild 2D videos paired with pseudo ground truth 3D data. Furthermore, we propose LaughTalk, a two-stage training baseline that can synthesize lip articulation and facial motion, synchronized to the input speech and laughter. Our extensive experiments show the efficacy of our approach, demonstrating strong performance in lip and laughter synchronization with speech, and evoking a sense of intimacy by accurately reflecting the laughter signal in the 3D talking head model. We would like to note that our proposed design choice and learning approach are independent of the specific non-verbal signal. As we primarily tackle the prominent non-verbal signal of laughter, future research could explore extending our model to convey other essential non-verbal cues, such as crying or shouting, thereby broadening its application scope.

References

- [1] Viveka Adelswärd. Laughter and dialogue: The social significance of laughter in institutional discourse. *Nordic Journal of Linguistics*, 1989. 1
- [2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 3
- [3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 4
- [4] Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 2005. 2
- [5] Brandon Castellano. Pyscenedetect. <https://github.com/Breakthrough/PySceneDetect>. 4
- [6] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2018. 2, 3
- [7] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3
- [8] 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. 1, 2, 3, 6, 8, 11
- [9] 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. 2, 3, 4, 11
- [10] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael J Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2023. 2, 3
- [11] 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. 1, 2, 4, 6, 8, 11
- [12] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 2010. 2, 3
- [13] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 2021. 3, 4, 8
- [14] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Spectre: Visual speech-informed perceptual 3d facial expression reconstruction from videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 2, 3, 4
- [15] Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Borman. Robust laughter detection in noisy environments. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2021. 4
- [16] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8
- [17] Robert A Hinde. *Non-verbal communication*. Cambridge University Press, 1972. 1
- [18] Ahmed Hussen Abdelaziz, Barry-John Theobald, Paul Dixon, Reinhard Knothe, Nicholas Apostoloff, and Sachin Kajareker. Modality dropout for improved performance-driven talking faces. In *Proceedings of the International Conference on Multimodal Interaction*, 2020. 2
- [19] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [20] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *International Conference on 3D Vision (3DV)*, 2021. 2, 4
- [21] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 2017. 1, 2, 8
- [22] Billy Lee. Nonverbal intimacy as a benchmark for human-robot interaction. *Interaction Studies*, 2007. 1, 7
- [23] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 2017. 2, 3
- [24] Chen Liu. An analysis of the current and future state of 3d facial animation techniques and systems. *Simon Fraser University*, 2009. 1
- [25] Albert Mehrabian. *Nonverbal communication*. Routledge, 2017. 1
- [26] 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. 2, 6, 11, 12
- [27] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2017. 3
- [28] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 2
- [29] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6, 8
- [30] Disa A Sauter, Frank Eisner, Andrew J Calder, and Sophie K Scott. Perceptual cues in nonverbal vocal expressions of

- emotion. *Quarterly journal of experimental psychology*, 2010. [1](#)
- [31] Sophie K Scott, Nadine Lavan, Sinead Chen, and Carolyn McGettigan. The social life of laughter. *Trends in cognitive sciences*, 2014. [1](#), [3](#)
- [32] David Stauffer. Let the good times roll: Building a fun culture. *Harvard Management Update*, 1999. [1](#), [3](#)
- [33] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *ACM International Conference on Multimedia (MM)*, 2021. [4](#)
- [34] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *IEEE International Conference on Computer Vision (ICCV)*, 2022. [1](#)
- [35] Guanzhong Tian, Yi Yuan, and Yong Liu. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2019. [2](#)
- [36] Linda Tickle-Degnen and Robert Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1990. [1](#), [3](#)
- [37] Robert A Vartabedian and Laurel Klinger Vartabedian. Humor in the workplace: A communication challenge. *ERIC*, 1993. [1](#), [3](#)
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [2](#)
- [39] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision (ECCV)*, 2020. [2](#), [3](#)
- [40] Melissa B Wanzer, Ann B Frymier, and Jeffrey Irwin. An explanation of the relationship between instructor humor and student learning: Instructional humor processing theory. *Communication education*, 2010. [3](#)
- [41] Isabell Wohlgenannt, Alexander Simons, and Stefan Stieglitz. Virtual reality. *Business & Information Systems Engineering*, 2020. [1](#)
- [42] Cheng-hsin Wu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, et al. Multi-face: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. [3](#)
- [43] 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. [1](#), [2](#), [6](#), [8](#), [11](#)
- [44] Kim Youwang, Lee Hyun, Kim Sung-Bin, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. A large-scale 3d face mesh video dataset via neural re-parameterized optimization. *arXiv preprint arXiv:2310.03205*, 2023. [2](#), [4](#)
- [45] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [2](#), [3](#)
- [46] Jie Zhang and Robert B Fisher. 3d visual passcode: Speech-driven 3d facial dynamics for biometrics. *Signal processing*, 2019. [3](#)
- [47] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European Conference on Computer Vision (ECCV)*, 2022. [2](#), [3](#)
- [48] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*, 2022. [8](#)
- [49] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [8](#)

Appendix

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 [8, 11, 43] with our proposed LaughTalk dataset to ensure fair qualitative and quantitative comparison. Here, we provide the training details of the existing methods.

CodeTalker. CodeTalker [43] 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 code⁶.

FaceFormer and VOCA. For FaceFormer [11] and VOCA [8], we initiated pre-training using LaughTalk_{MEAD}, which consists of neutral speech and corresponding 3D faces. Subsequently, fine-tuning was conducted on both models using LaughTalk_{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 method⁷.

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 [26]. 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 [9]. 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 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?

⁶<https://github.com/Doubiiu/CodeTalker>.

⁷<https://github.com/EvelynFan/FaceFormer>.

<https://github.com/TimoBolkart/voca>.

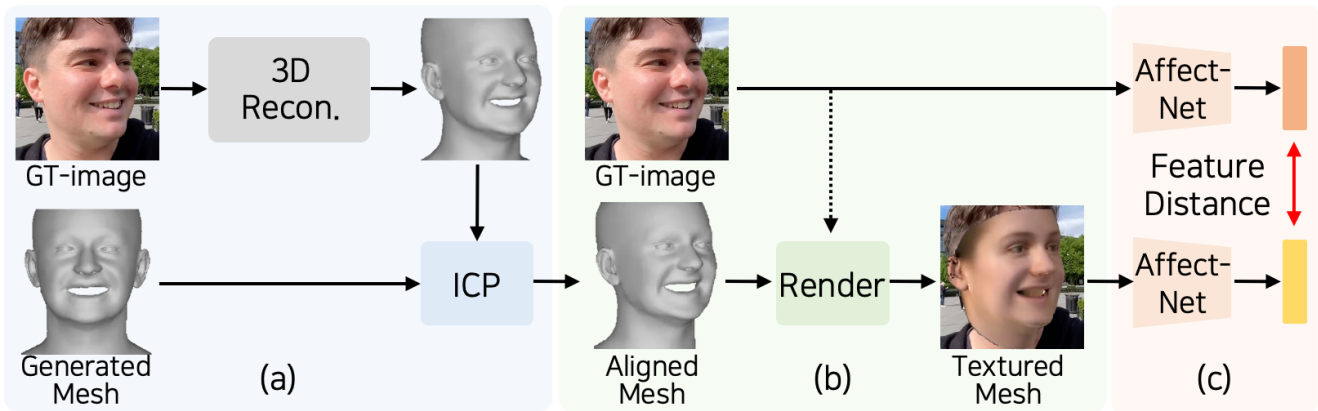



Figure S1. **Measuring emotional feature distance.** We present the Emotional Feature Distance (EFD), a perceptual metric designed to evaluate the synchronization of facial movements with laughter. To calculate this metric, first, (a) we align the speech-driven generated mesh vertices with the mesh reconstructed 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 [26] 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?

Video A



Video B



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.