DOI: 10.1111/cgf.70073
EUROGRAPHICS 2025 / A. Bousseau and A. Dai
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 2

"Wild West" of Evaluating Speech-Driven 3D Facial Animation Synthesis: A Benchmark Study

Kazi Injamamul Haque[1] and Alkiviadis Pavlou[1] and Zerrin Yumak[1]

[1]Utrecht University, The Netherlands

Abstract

Recent advancements in the field of audio-driven 3D facial animation have accelerated rapidly, with numerous papers being published in a short span of time. This surge in research has garnered significant attention from both academia and industry with its potential applications on digital humans. Various approaches, both deterministic and non-deterministic, have been explored based on foundational advancements in deep learning algorithms. However, there remains no consensus among researchers on standardized methods for evaluating these techniques. Additionally, rather than converging on a common set of datasets and objective metrics suited for specific methods, recent works exhibit considerable variation in experimental setups. This inconsistency complicates the research landscape, making it difficult to establish a streamlined evaluation process and rendering many cross-paper comparisons challenging. Moreover, the common practice of A/B testing in perceptual studies focus only on two common metrics and not sufficient for non-deterministic and emotion-enabled approaches. The lack of correlations between subjective and objective metrics points out that there is a need for critical analysis in this space. In this study, we address these issues by benchmarking state-of-the-art deterministic and non-deterministic models, utilizing a consistent experimental setup across a carefully curated set of objective metrics and datasets. We also conduct a perceptual user study to assess whether subjective perceptual metrics align with the objective metrics. Our findings indicate that model rankings do not necessarily generalize across datasets, and subjective metric ratings are not always consistent with their corresponding objective metrics. The supplementary video, edited code scripts for training on different datasets and documentation related to this benchmark study are made publicly available- https://galib360.github.io/face-benchmark-project/.

CCS Concepts

• ***Computing methodologies*** → ***Neural networks; Animation;*** • ***Human-centered computing*** → *User studies;*

1. Introduction

Speech-driven 3D facial animation synthesis for digital humans has been an active area of research both in academia and industry. Digital humans are no longer only being used for entertainment like video game and film production, but also gaining well-deserved traction for extended reality, virtual communication and human-computer interaction applications. There has been an influx of speech-driven 3D facial animation papers in recent years [CBL*19, XXZ*23, HY23, DCT*23, SHY23, WHY24, ZLZ*24, ALTT22, ATDN24, SWJ*24] and a lot more are still under review and available in the open repository Arxiv [TACT23, KCP*24]. While most work address this problem by proposing deterministic deep learning models [CBL*19, XXZ*23, HY23, DCT*23], recently there is an increasing number of non-deterministic approaches [SHY23, WHY24, ZLZ*24, SLY*24] that faithfully represent the stochastic characteristics of human motion as we observe in real-life. However, there has not been a consensus among researchers and practitioners on the objective and subjective metrics to be used in order to evaluate such models. Oftentimes non-deterministic methods are using metrics that are designed to evaluate deterministic models. There is a lack of consensus and comparative analysis among various objective metrics. Moreover, subjective evaluation metrics are often limited to lip-sync and overall realism neglecting in-depth evaluations when it comes to non-determinism and emotional variations. In this work we explore the "wild west" of evaluating speech-driven 3D facial animation synthesis approaches by looking at a variety of datasets and covering both deterministic, non-deterministic and emotion-control enabled models. We compare these models using an extensive and carefully curated set of objective metrics. Additionally, we carry out a perceptual user study to understand if/how the objective metrics are inline with the perceptual metric ratings that are inherently subjective by nature.

Closest to ours is the work of Yang et al. [YRC*24] that identifies the lack of systematic evaluation and benchmarking techniques in this area. Different from their work, we include deterministic, non-deterministic and emotion-control enabled methods as part of our comparison and provide benchmarking on three datasets. One

of the limitations of their work is the lack of comparisons over 3D vertex-based and emotion-enabled datasets. We include BIWI [FGR*10] and Multiface [WZA*22] as the 3D vertex-based methods and 3DMEAD as an emotionally-rich dataset based on the 2D video dataset MEAD [WWS*20]. Another difference in comparison to Yang et. al. [YRC*24] is the selection of objective evaluation metrics. Some of the metrics they introduce are specific to their methodology and their code is not openly available. Instead, we use a more balanced set of metrics that are open-source and readily accessible to researchers, focusing on both deterministic and non-deterministic aspects inline with previous work to allow for direct comparisons. In addition to lip vertex, mean vertex error and upper face dynamics deviation, we add the diversity metric. Two of the metrics are adopted from Yang et al. which are non-deterministic adaptations of lip vertex error. A final difference is the analysis of the relationship between subjective and objective metrics as these metrics do not always align with each other.

The main contributions of our work are enumerated as follows-

- Providing a comprehensive evaluation of most recent and prominent 3D facial animation synthesis methods across three different datasets objectively and subjectively.
- Analysis of deterministic, non-deterministic and emotion-control enabled 3D facial animation synthesis methods.
- Comparison of the objective and subjective evaluation metrics and analysis of the relationship between them.

2. Related Work

The field of speech-driven 3D facial animation has gained considerable momentum, thanks to the rapid advancements in deep learning research in recent years. Before data-driven approaches gained traction, 3D facial animation driven by speech followed procedural or rule-based frameworks based on phonemes and visemes representation [JAL23, CYvdS19, TMTM12]. Although animator-centric, these approaches often fall short in capturing the intricate temporal dynamics and expressive nuances required for lifelike animation. With the rise of deep learning, and advancements in sequence modeling approaches, many works tackled the speech-driven 3D facial animation generation task deterministically [KAL*17, TKY*17, CBL*19, FLS*22, RZW*21, XXZ*23, THA*23, HY23]. Although the lack of large-scale audio-4D dataset posed as a problem in the beginning for such data-driven approaches, significant improvements in vision based face tracking and 3D reconstruction methods allow to create large-scale datasets from videos, such as- [LBB*17, GKG*24, DBB22, EST*20, FFBB21]. However, as human motion, including facial motion is inherently non-deterministic and deterministic methods struggle to capture the natural variability of human expressions, more recent works have focused on proposing non-deterministic or probabilistic approaches that encourage diversity in the generated motions [SHY23, ATDN24, ZLZ*24, WHY24, YRC*24]. These methods utilize probabilistic frameworks to produce diverse yet realistic animations that better reflect the inherent stochastic nature of human facial expressions. Despite these recent advancements, emotional style control remains an underexplored yet a critical aspect of speech-driven expressive 3D facial animation. Capturing expressive variations in speech is essential for creating believable and emotionally rich digital humans. Several works [KAL*17, PWS*23, DCT*23, ZLZ*24, WHY24] have attempted to address this challenge by incorporating emotional cues into their models. However, the lack of standardized evaluation frameworks for assessing emotional expressiveness limits the comparability and effectiveness of these approaches, highlighting the need for more robust evaluation protocols.

In parallel, significant recent efforts have been devoted to generating 2D talking-head animation videos [JZW*22, SVH*24, ZCW*23, GYY*23, XZZ*23, GMW*23]. These works largely rely on vision-based techniques and output rendered RGB pixels, making them less suited for interactive applications requiring 3D animation, which is the primary focus of this work.

2.1. Deterministic Models

Deterministic algorithms produce the same output across multiple generations given identical set of inputs. The first end-to-end data-driven methods, such as- [TKY*17, ZXL*18, CBL*19, FLS*22, HY23, XXZ*23] deterministically generate 3D facial animation given speech input. Among these, researchers explored sliding window-based [TKY*17], recurrent neural network-based approaches [ZXL*18]. These methods however, rely on intermediary phoneme representations and only focus on the mouth region, neglecting the broader facial dynamics as a whole. Cudeiro et al. [CBL*19] propose VOCA that employs pre-trained DeepSpeech [HCC*14] for extracting audio features and the FLAME head model [LBB*17] for facial motion representation but it lacks expressivity for upper-face motion. Richard et al. [RZW*21] present MeshTalk, a speech-driven animation framework that learns a categorical latent space through cross-modality loss. By disentangling speech-correlated and uncorrelated information, MeshTalk more effectively models upper-face motions compared to VOCA. Despite the improvements, MeshTalk relies on a large amount of high-fidelity 3D facial data that poses a challenge in terms of scalability and accessibility. Fan et al. [FLS*22] propose FaceFormer, a transformer-based framework utilizing an autoregressive transformer architecture. Its encoder leverages Wav2Vec 2.0 [BZMA20], a self-supervised pretrained speech model, addressing the scarcity of large-scale 3D audio-visual data. Xing et al. [XXZ*23] highlight the limitations of prior works that treat cross-modal mapping as a regression task, which often leads to over-smoothed facial motions due to the regression-to-mean problem. To overcome this, they introduce CodeTalker, which integrates a transformer-based speech encoder with a cross-modal decoder that utilizes a learned discrete codebook. This approach achieves superior results compared to FaceFormer, VOCA, and MeshTalk. FaceXHuBERT [HY23] employs a pre-trained HuBERT speech model as the audio encoder, capturing both lexical and non-lexical speech information. A GRU-based motion decoder is used to synthesize facial animation with significantly reduced training time than the contemporary state-of-the-art models. While many models utillizes one-hot vectors for speaker identities, Thambiraja et al. [THA*23] propose Imitator that can learn speaker identity from a short speaking video.

For our benchmark experiments, FaceFormer [FLS*22], CodeTalker [XXZ*23], and FaceXHuBERT [HY23] were chosen as deterministic models for their performance and reproducibility.

2.2. Non-Deterministic Models

There has been a growing research interest in non-deterministic approaches that encourage diversity for speech-driven 3D facial animation task, aiming to properly reflect lifelike diverse facial motion that we as humans generate in our daily life. Primarily appeared in many body motion synthesis papers [TRG*22, ANBH23, AZL, YLL*23, CDA*24], non-determinism quickly got adopted and explored by researchers for facial animation generation [NJH*22, YRC*24, SHY23, ZLZ*24, ATDN24, SLY*24]. Some works further explored non-deterministic techniques for holistic animation generation task [YLL*23,LZB*24,CLW*24,NRB*24]. Probabilistic or non-deterministic deep learning methods proposed in recent years oftentimes encompass VQ-VAE or diffusion techniques incorporated in the model architecture. Learning to Listen [NJH*22], although for listeners' motion generation in a dyadic setting, was on of the first works that incorporated non-deterministic output for facial animation using transformer based VQ-VAE. Holistic model EMAGE [LZB*24] also adopts a VQ-VAE architecture and employs a diversity metric inspired from [LKP*21] to evaluate their model's results. Inspired from several body motion generation works based on diffusion techniques, Stan et al. [SHY23] proposed FaceDiffuser that utilizes diffusion mechanism to encourage diversity in the output. The paper introduces its own diversity metric that measures variation across identities that enabled fair comparisons with contemporary deterministic models. Sun et al. [SLY*24] propose DiffPoseTalk leveraging a diffusion based transformer network which further includes head rotation for enhanced realism in the output animation. Yang et al. [YRC*24] proposes a probabilistic method based on residual vector quantized codebook together with a benchmarking framework to evaluate non-deterministic methods. The authors proposed several new objective metrics that are well suited for evaluating non-deterministic approaches. Utilizing a dataset based on neural parametric head models (NPHMs) [GKG*24], Aneja et al. [ATDN24] propose FaceTalk that follows a transformer based diffusion model architecture. The authors draw inspiration from [RPZK23] to evaluated diversity in the generated motion. Holistic model Audio2Photoreal [NRB*24] also employs diffusion mechanism similar to the above mentioned papers. While these works showcase their advantage in synthesizing diverse animation outputs, lack the ability to control emotional expressivity in the synthesized animation.

In this study, we selected FaceDiffuser [SHY23] as a representative non-deterministic model based on its diffusion architecture. Additionally, we include a non-deterministic version of CodeTalker [XXZ*23], which uses VQ-VAE and is derived from its official open-source implementation, enabling direct comparison with its deterministic counterpart. Holistic models such as EMAGE and Audio2PhotoReal were excluded from our benchmark as their primary focus is on body animation, with insufficient attention to facial animation details.

2.3. Models with Emotional Style Control

Several works focused on generating emotion-controllable audio-driven 2D talking-head videos [GYY*23, XZZ*23, GMW*23]. While some methods, such as- [XZZ*23], use 3DMM parameters as an intermediary face representation, they heavily depend on vision-based image analysis, requiring large amounts of RGB frames or videos during training. This reliance makes them less suitable for interactive 3D applications, where the output must operate directly in the 3D space.

In 3D space, some works such as- [KAL*17, PWS*23, DCT*23,ZLZ*24,WHY24] have been proposed over recent years. [KAL*17] trains a CNN-based neural network to generate facial animation with emotional expressivity using a small, yet high-quality dataset captured with a commercial 4D capture system. However, the emotion control is not explicit and the small scale dataset used contributes towards the lack of generalizability. By utilizing a semantically emotion annotated dataset 3D-ETF, Peng et al. [PWS*23] proposed EmoTalk introducing an emotion-disentangling encoder that separates emotion and speech content through a cross-reconstruction loss. The authors constructed 3D-ETF dataset by reconstructing two separate video datasets-RAVDESS [LR18] and HDTF [ZLDF21]. Similarly, deterministic method EMOTE [DCT*23] reconstructs emotion annotated MEAD [WWS*20] video dataset and employs a VAE based transformer network in a 2-stage training process. The dataset features annotations in terms of 8 basic emotion classes and 3 levels of intensities. The authors first train a motion prior model using a temporal VAE and in the later stage, introduce audio and emotion conditioned training that allow to infer facial animation in a non-autoregressive manner. EMOTE does not conduct an explicit objective evaluation against EmoTalk but qualitative and visual comparison suggest its superior visual quality. Unlike categorical emotion control, Zhao et al. [ZLZ*24] more recently proposed Media2Face that allows for text-based emotion control using CLIP [RKH*21] text embeddings in their training, encouraging a more granular approach for emotion control. The authors constructed their own dataset M2F-D and follow a 2-stage training approach. The first stage incorporates a VAE to learn a latent space for facial geometry and facial expression that decouples identity and expression. This latent space is employed to construct the M2F-D dataset by extracting high-quality expressions and accurate head poses from a large array of videos from three video datasets - RAVDESS, HDTF and MEAD. In the second stage, a transformer-based diffusion model is trained so that it can generate the final animations non-deterministically. While Media2Face achieves high visual quality, its dataset and codebase were not publicly available at the time of writing, preventing direct comparisons. Furthermore, the authors do not provide an evaluation pertaining to diverse output generation. Song et al. [SWJ*24] present another two-stage framework combining an emotion-enabled VQ-VAE with a latent diffusion model. Their approach outperforms EMOTE but falls short of FaceDiffuser in performance. Similar to [ZLZ*24], the authors do not evaluate their model in terms of the ability to generate diverse results. ProbTalk3D [WHY24] introduces a non-deterministic, emotion-controllable framework trained on the 3DMEAD dataset. Rather than directly utilizing high-dimensional vertex-based data, it leverages low-dimensional FLAME parameters. ProbTalk3D employs a 2-stage VQ-VAE framework and demonstrates superior performance compared to recent deterministic and non-deterministic methods. For our benchmark experiments, we selected ProbTalk3D due to its non-deterministic and emotion-controllable design.

3. Methodology

In this section, we outline the experimental setup for our benchmarking process. The following subsections introduce the selected datasets for the experiment, followed by brief descriptions of the models used in the benchmarking process. Subsequently, we present the curated set of objective metrics employed for evaluation. The methodology for the perceptual user study is described in Section 3.5. The models chosen for this experiment are first trained on the selected datasets, with identical dataset splits maintained throughout. The objective metrics are then computed, and the benchmarking table is constructed based on the results. The models are ranked according to their performance. Additionally, for the emotion-enabled 3DMEAD dataset, we conducted a perceptual user study to gather insights into how users perceive lip synchronization and animation realism across generated animations from different models.

The experimental setup consists of the following steps:

- Dataset splits for the datasets presented in Section 3.1 are defined and are kept consistent throughout the benchmarking experiment.
- A set of objective metrics, detailed in Section 3.4, is used for the quantitative evaluation of the trained models.
- The speech-driven models, described in Sections 3.2 and 3.3, are trained on the datasets, and the corresponding evaluation metrics are computed on unseen test sets.
- A perceptual user study, detailed in Section 3.5, is conducted using the generated animations from the trained models to evaluate the results subjectively.
- Ranking tables are constructed based on the obtained results, with each column representing the evaluation metrics and each row representing the respective rankings in ascending order, as shown in Tables 5 and 6.
- The results are presented in Section 4, followed by a discussion of the findings in Section 5.

3.1. Datasets

We employ three audio-visual 3D datasets- BIWI [FGR*10], Multiface [WZA*22] and 3DMEAD [WWS*20, DCT*23]. Two of these datasets are commonly used 3D vertex based datasets however they do not have enough emotional variation. 3DMEAD dataset is constructed from 2D videos containing emotionally-rich expressions. All three were used for objective benchmarking, whereas 3DMEAD is further used for the perceptual user study.

BIWI [FGR*10]: The BIWI dataset comprises $14 \times 40 \times 2$ sequences, where each sequence combines audio with corresponding facial animations. Fourteen subjects were tasked with reading and expressing 40 different sentences, each performed twice: once with a neutral expression and once with an emotional one. Each sequence lasts approximately 5 seconds on average and was recorded at 25 frames per second (fps). The dataset provides high-resolution face meshes, consisting of 23,370 3D vertices, though only the front of the head is captured. In line with prior work [FLS*22, XXZ*23, SHY23], we utilize only the emotional subset of sequences following the guidelines in [SHY23]. For training, six subjects (three female and three male) were chosen, each providing 32 spoken sentences, resulting in a total of 192 sequences that make up the BIWI-Train dataset. The remaining 8 sentences from these subjects were divided, with 4 used for validation (24 sequences total) and 4 for testing (24 sequences total).

Multiface [WZA*22]: The publicly available version of the dataset consists of 13 subjects, selected from a larger pool of 250 individuals used for training Meshtalk [RZW*21]. Each subject provides 50 spoken sentences, and the sequences are organized by both subject and sentence, allowing for consistent training approaches such as one-hot embeddings. The dataset creators ensured the sentences were phonetically balanced to support generalization across a wide range of phonemes. Each animation sequence features 3D face meshes captured at 30 frames per second (fps), with each frame representing the complete 3D face of the actor. The meshes contain 6,172 3D vertices, detailing features such as eyelids, neck, and various hairstyles. We adopt the dataset split for Multiface as outlined in FaceDiffuser [SHY23].

3DMEAD [WWS*20]: The 3DMEAD dataset is a 3D reconstruction of the 2D audio-visual MEAD dataset [WWS*20]. The 3D reconstruction process was performed using DECA [FFBB21] and MICA [ZBT22]. 3DMEAD was first introduced in the EMOTE paper [DCT*23] for a deterministic method and later was used in non-deterministic methods [WHY24] and [SWJ*24] enabling emotion-control. The 3DMEAD dataset consists of 3D reconstructions of 47 subjects speaking in English, expressing eight emotions at three intensity levels. The emotions include neutral, happy, sad, surprised, fear, disgust, anger, and contempt. Apart from the neutral emotion, each category is represented at three intensity levels: weak, medium, and strong. Each subject contributes 30 short sentences for the seven basic emotions, each expressed at the three intensity levels, along with additional 40 sentences reflecting neutral expressions.

We select the 3DMEAD dataset for our experiments due to its relatively large-scale, high-quality facial animation data, which includes a broad spectrum of emotional expressions. The motion data is sampled at 25 frames per second (fps), and each frame is represented using the FLAME [LBB*17] 3D Morphable Model (3DMM) parameters $\beta, \theta, \psi \in \mathbb{R}^{406}$, where $\beta \in \mathbb{R}^{300}$ corresponds to face shape, $\theta_{jaw} \in \mathbb{R}^3$ represents the jawbone's Euler angle rotation (x, y, z), $\theta_{global} \in \mathbb{R}^3$ denotes the global head pose, and $\psi \in \mathbb{R}^{100}$ represents the expression parameters. Consistent with EMOTE and ProbTalk3D, we focus on utilizing only $\psi, \theta_{jaw} \in \mathbb{R}^{53}$ for model training, where $\psi \in \mathbb{R}^{50}$ accounts for the first 50 of the 100 expression parameters. In contrast to the dataset split used in EMOTE which is only suitable for subjective evaluation, we adopt the split proposed in ProbTalk3D which facilitates both objective and subjective evaluations.

3.2. Benchmark Models: Deterministic

In this subsection, we summarize the deterministic speech-driven 3D facial animation generation models that we used for the benchmarking process. For a detailed understanding of these models, we refer to the corresponding papers.

Dataset	Subjects	Sequences	Emotions	Intensities	Total Sequences	Hours	Frames
BIWI	14	40	binary	1	1120	$\approx$ 1.55	139.5 K
Multiface	13	50	1	1	562	0.67	65 K
3DMEAD	47	30 (emotion) + 40 (neutral)	8	3	21115	$\approx$ 26	2.3 M

Table 1: *Summary of the datasets used for benchmarking speech-driven 3D facial animation task.*

Dataset	**BIWI**	**3DMEAD**	**Multiface**
Training Set	6 subjects 32 sequences per subject Total = 192 sequences	32 subjects 24 sequences per subject (emotional) 32 sequences per subject (neutral) Total = 17098 sequences	9 subjects 40 sequences per subject Total = 360 sequences
Validation Set	6 seen subjects 4 sequences per subject Total = 24 sequences	32 seen subjects 3 sequences per subject (emotional) 4 sequences per subject (neutral) Total = 2108 sequences	9 seen subjects 5 sequences per subject Total = 45 sequences
Test Set	6 seen subjects 4 sequences per subject Total = 24 sequences	32 seen subjects 3 sequences per subject (emotional) 4 sequences per subject (neutral) Total = 1909 sequences	9 seen subjects 5 sequences per subject Total = 45 sequences

Table 2: *Dataset Splits. The training and validation sets are used during training while the test sets were used to compute objective metrics.*

3.2.1. FaceFormer

FaceFormer [FLS*22] introduces an encoder-decoder architecture that leverages pretrained Wav2Vec2.0 [BZMA20] as the audio encoder and a transformer-based autoregressive decoder to predict animation sequences. By incorporating long-term audio context, FaceFormer addresses limitations of traditional models that rely on short audio windows, often leading to inaccuracies in lip synchronization and facial expressions during extended speech. The model employs a biased causal multi-head self-attention mechanism to align audio input with motion data and maintain temporal consistency. This design enables FaceFormer to generate coherent 3D facial animations with accurate lip synchronization and expressive facial motions. Additionally, it tackles the scarcity of 3D audio-visual data by integrating self-supervised pretraining and attention mechanisms. Extensive experiments and user studies demonstrate its superiority over state-of-the-art models in realism and lip-sync.

3.2.2. CodeTalker

Xing et al. [XXZ*23] tackle the challenge of generating realistic animations from audio input using a two-stage training framework with a VQ-VAE. Similar to FaceFormer [FLS*22], it incorporates the pretrained Wav2Vec2.0 model as the audio encoder. To address the regression-to-mean problem, which limits expressiveness in prior methods, CodeTalker formulates the task as a code query in a discrete motion prior space rather than a direct regression. The VQ-VAE learns a discrete codebook of facial motion primitives, reducing ambiguity in cross-modal audio-to-motion learning. A speech-conditioned autoregressive model then synthesizes 3D facial motions sequentially, leveraging the learned motion space to generate lip-synchronized and expressive animations. CodeTalker's speaker-agnostic motion prior enhances generalization across speakers and expressions. Both qualitative and quantitative comparative evaluations, along with a user study, demonstrate its superiority in generating 3D facial animation from speech.

3.2.3. FaceXHuBERT

FaceXHuBERT [HY23] is an expressive 3D facial animation synthesis model based on an end-to-end encoder-decoder architecture. The model leverages HuBERT [HBT*21] as the audio encoder, capturing both lexical and non-lexical speech features, allowing it to generalize across diverse speech inputs without requiring a large-scale paired dataset. The audio embeddings from HuBERT are processed through an *Input Representation Adjustment* module to ensure alignment of audio features with 3D facial animation data. The decoder employs a lightweight two-layer GRU architecture, which generates vertex displacements for 3D facial mesh sequences. Additional inputs, including speaker identity and binary emotion labels (neutral or expressive), enhance the emotional expressiveness of the animations. This design reduces network complexity and training time compared to transformer-based decoders while achieving high-quality lip-sync accuracy and realistic expressions. FaceXHuBERT outperforms models like FaceFormer and CodeTalker in objective evaluations and achieves improved training efficiency, maintaining superior accuracy in capturing subtle emotional facial motions.

3.3. Benchmark Models: Non-Deterministic

Here, we briefly describe the non-deterministic speech-driven 3D facial animation generation models that were used for the benchmark experiment. For further details about the models, we refer readers to the corresponding papers. We select one VQ-VAE based model which is a non-deterministic modification of CodeTalker and one diffusion-based FaceDiffuser [SHY23]. We also add the emotion-control enabled model ProbTalk3D [WHY24].

3.3.1. CodeTalker-ND

Although CodeTalker originally produces deterministic results, we adopted their implementation and made use of the multinomial sampling method during the codebook entry retrieval instead of retrieving the best matched entry. We include this non-deterministic version of the original model in our experiment, namely CodeTalker-ND.

3.3.2. FaceDiffuser

Stan et al. [SHY23] introduced FaceDiffuser, a non-deterministic model that incorporates a diffusion mechanism to generate varied facial animations. Unlike deterministic methods, which produce identical outputs for the same speech input, FaceDiffuser models the natural variability of facial expressions. It uses HuBERT, a pre-trained self-supervised speech encoder, to process audio input and applies a denoising diffusion process to learn and synthesize expressive facial animations. The GRU-based decoder predicts either 3D vertex displacements or blendshape parameters or rig control values, making the model compatible with both vertex-based datasets and blendshape animation rigs. FaceDiffuser supports animation generation under diverse speech conditions, such as noisy environments or multiple speakers, enhancing its robustness. Extensive evaluations show that it matches or outperforms state-of-the-art models like FaceFormer and CodeTalker, while introducing variability in the generated outputs.

3.3.3. ProbTalk3D

ProbTalk3D [WHY24] is an emotion-controllable speech-driven 3D facial animation model trained on the 3DMEAD dataset. The model employs a two-stage VQ-VAE architecture to generate diverse yet accurate animations with variability in expression and emotional intensity. In Stage 1, a VQ-VAE-based motion autoencoder learns a discrete latent representation of facial motion using a codebook. In Stage 2, speech features from HuBERT audio encoder are fused with an emotion-controlled style embedding (i.e. subject identity, emotion class, and intensity) and decoded into facial motion. Probabilistic sampling during inference allows the model to produce diverse outputs for the same input set. Evaluations show that ProbTalk3D outperforms state-of-the-art models like FaceDiffuser and CodeTalker in diversity metrics while maintaining competitive vertex error metrics. Qualitative comparisons and user study further highlight its ability to generate expressive and visually realistic diverse facial animations.

3.4. Benchmark Objective Metrics

The quantitative metrics that we used to benchmark the aforementioned models are described in this subsection. We use the most commonly used objective metrics found in recent relevant literature: lip vertex error (LVE), mean vertex error (MVE) and upper face dynamics deviation (FDD). These three metrics are designed for deterministic models and they provide insights about accuracy of the models. However, a different set of objective metrics are required to properly evaluate non-deterministic models, *diversity* being an important metric that appears in non-deterministic or probabilistic models. Unfortunately, there has not been a common diversity metric that the different papers are using. For example, [SHY23] defines diversity in terms of conditioning on different speaking subjects to be able to fairly compare with deterministic models. While [WHY24, SLY*24] define diversity by generating multiple samples using the same inputs and by computing the average differences of the generated samples inspired from [RPZK23] that used diversity for evaluating non-deterministically generated body animations. We adopt the diversity metric as used in [WHY24]. Additionally, following [YRC*24], we incorporate mean estimate error (MEE) and coverage error (CE) in our experiments to quantitatively evaluate non-deterministic models.

3.4.1. LVE

The Lip Vertex Error (LVE) measures the maximum $\mathcal{L}_2$ error between the vertices in the lip or mouth region of a predicted frame and the corresponding ground truth frame. This error is then averaged across all generated frames. Given that the facial topology is identical for both the 3DMEAD and VOCASET datasets, we utilize the same lip mask as applied in [XXZ*23, SHY23].

For N predicted frames from the test audio, let x^i_{lip} represent the vertices of the lip region in the ground truth frame, and $\hat{x}^i_{lip}$ denote the corresponding vertices in the predicted frame. The LVE is then computed as:

$$\text{LVE} = \frac{1}{N}\sum_{i=1}^{N} \max |x^i_{lip} - \hat{x}^i_{lip}|_2 \tag{1}$$

3.4.2. MVE

The Mean Vertex Error (MVE) is used to calculate the average Euclidean distance between the predicted frames and the corresponding ground truth frames across the entire test set. Let N denote the total number of frames generated for all test-set audio inputs. We define x_i as the ground truth of the i-th frame and $\hat{x}_i$ as the predicted frame. The MVE is computed using the following equation:

$$\text{MVE} = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \hat{x}_i\| \tag{2}$$

3.4.3. FDD

The Upper Face Dynamic Deviation (UFDD), introduced in [XXZ*23], quantifies the variation in facial dynamics of motion sequences relative to the ground truth. This metric provides an indication of how closely the standard deviation—or the variation in upper face motion—of the generated sequences for the test-set audio aligns with the variation observed in the ground truth sequences.

3.4.4. Diversity

To evaluate diversity, we adopt the definition proposed in [RPZK23], initially introduced for assessing synthesized body motions and later adapted for the evaluation of facial animations in works such as DiffPoseTalk [SLY*24], 3DiFACE [TACT23], and ProbTalk3D [WHY24]. Our goal is to quantify the diversity of generated facial animations under identical input conditions.

Given A audio inputs, we generate 10 facial animation samples for each input, all guided by the same control signals. For the i-th audio input, two random subsets are sampled from the set of generated animations, each consisting of B samples. For instance, for 3DMEAD dataset, $A = 1909$ (corresponding to the test-set audio inputs), and $B = 5$ (i.e., 10 generated samples are randomly divided into two subsets, S_1 and S_2, each containing 5 samples).

We then calculate the average Euclidean distance between the j-th sample within the two subsets for each audio input. This process is repeated for all A audio inputs, and the overall mean value is computed to represent the final diversity metric. The formal definition of this diversity measure is as follows:

$$\text{Diversity} = \frac{1}{A \times B} \sum_{i=1}^{A} \sum_{j=1}^{B} \|(\hat{x}_{i,j} \in S_1) - (\hat{x}_{i,j} \in S_2)\|_2 \quad (3)$$

3.4.5. MEE

The Mean Estimate Error (MEE), introduced in [YRC*24], is proposed to evaluate how closely the mean of a sampling distribution approximates the ground truth. To compute MEE, we generate a set of samples $S = \hat{x_1}, \hat{x_2}, \ldots, \hat{x_{10}}$. For each test audio, 10 motion sequences are generated, and the mean $E(\hat{x})$ of these 10 samples is calculated.

MEE is then computed for all test sequences using Eq. 4, and the average is taken across the entire set of test sequences. A lower MEE value suggests that the model is more successful in producing ground truth-aligned lip movements. This metric is particularly suitable for probabilistic and non-deterministic models, as it evaluates a set of generated samples rather than a single instance.

$$\text{MEE} = \text{LVE}(x, E(\hat{x})) \quad (4)$$

3.4.6. CE

Coverage Error (CE) assesses how closely the sampling distribution of a probabilistic model aligns with the ground truth, as described in [YRC*24]. To compute CE for a single test sequence, we generate a set S consisting of 10 samples, similarly to the procedure for MEE. The minimum Lip Vertex Error (LVE) between the ground truth and the generated samples is then calculated using Eq. 5.

The mean CE across all test sequences is computed to provide the final CE value. A lower CE indicates that the probabilistic model's predictions better encompass the ground truth samples in terms of lip motion.

$$\text{CE} = \min_{\hat{x} \in S} \text{LVE}(x, \hat{x}) \quad (5)$$

3.5. Perceptual User Study

Our perceptual user study aims to gather subjective metrics in which users evaluates the results based on two key aspects: realism and lip-sync that were rated independently. Realism reflects the overall facial movement accuracy while lip-sync is concerned with the alignment between lip movements and speech. We define

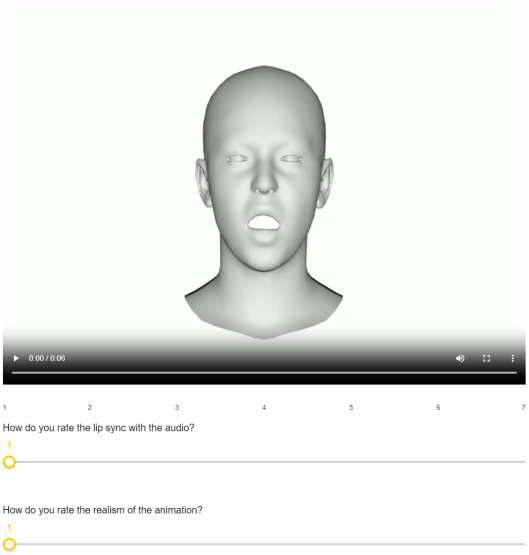


Figure 1: *Perceptual user study UI for each individual animation viewed and rated by the users in terms of lip-sync and realism.*

an experiment assumption that lip-sync is related to objective metrics LVE, MEE, CE (these metrics calculate the vertex error for the lip region) whereas realism relates to FDD, MVE, Diversity (these metrics provide insights about upper face and full face accuracy). For the user study, we exclusively utilized the 3DMEAD dataset due its expressive richness. In conventional perception studies for 3D facial animations, A/B testing is typically employed to directly compare a proposed model with both state-of-the-art models and the ground truth. However, given that our objective was not to demonstrate the superiority of a specific model, we opted for an individual rating methodology. In this approach, participants rated rendered stimuli independently without simultaneous comparison, thus providing a more nuanced evaluation without a forced choice.

We generated four random predictions for each motion from all models trained on the 3DMEAD dataset. These results, along with the ground truth, were then rendered for user evaluation. To minimize potential biases, a single random sample from the four predictions was presented to the user. Additionally, the emotions and models were intermixed to reduce any favoritism that could arise from familiarity or preference. Participants rated each motion on a scale from 1 to 7, where 7 represented the highest score in terms of realism and lip-sync accuracy. The user interface for this rating process is illustrated in Figure 1.

Due to the large number of motions that required evaluation—six models plus ground truth for each emotion, resulting in a total of 56 motions—we divided the study into two segments. Users were randomly assigned to groups and presented with a subset consisting of four randomly selected emotions for all models. This strategy ensured that the study did not become overly time-consuming, thus maintaining participant engagement and the quality of their responses. Participants were recruited via Prolific, ensuring appropriate monetary compensation. In total, we gathered data from 45 paid participants and 16 volunteers, amounting to 61 study partic-

Model	LVE (↓) $\times 10^{-4}$	FDD (↓) $\times 10^{-5}$	MVE (↓) $\times 10^{-3}$	Diversity (↑) $\times 10^{-3}$	MEE (↓) $\times 10^{-4}$	CE (↓) $\times 10^{-4}$
Dataset: BIWI						
FaceDiffuser [SHY23]	4.946	4.430	6.810	0.002	**4.951**	**4.993**
CodeTalker-ND [XXZ*23]	6.333	5.186	7.357	**0.323**	6.670	6.180
CodeTalker [XXZ*23]	4.791	4.117	**6.013**	-	-	-
FaceFormer [FLS*22]	4.928	4.627	7.135	-	-	-
FaceXHuBERT [HY23]	**4.729**	**3.901**	6.295	-	-	-
Dataset: Multiface						
FaceDiffuser [SHY23]	**5.766**	5.160	**6.739**	0.001	**5.767**	**5.766**
CodeTalker-ND [XXZ*23]	20.262	9.350	14.288	**0.012**	21.622	20.059
CodeTalker [XXZ*23]	16.940	5.431	10.945	-	-	-
FaceFormer [FLS*22]	13.405	6.934	8.446	-	-	-
FaceXHuBERT [HY23]	18.001	**5.002**	9.167	-	-	-

Table 3: *Objective Evaluation Metrics for the Models on BIWI and multiface datasets.*

ipants. After the study responses are collected, the ratings for the subjective metrics were then aggregated.

4. Results

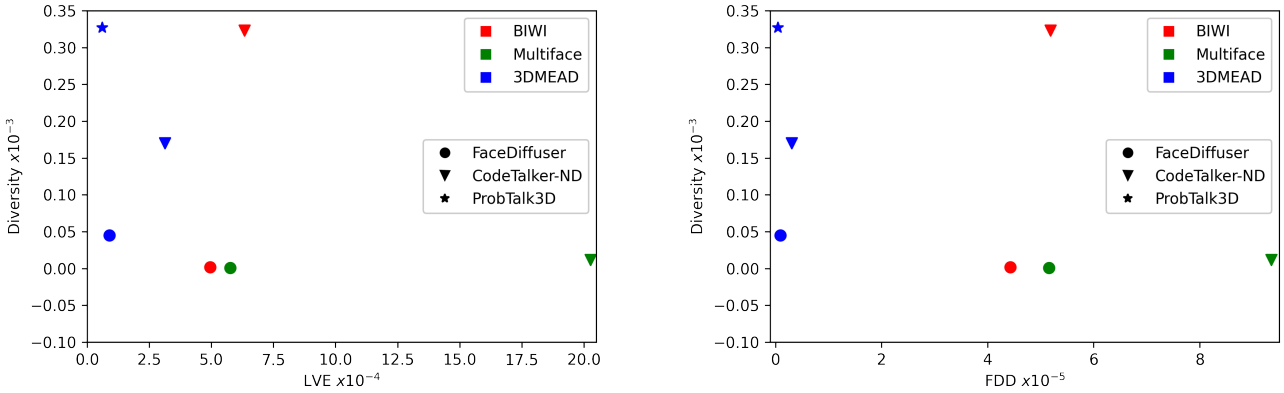
We evaluated the performance of the selected speech-driven 3D facial animation models on three datasets: BIWI, Multiface and 3DMEAD. The objective metrics used include- Lip Vertex Error (LVE), Upper Face Dynamic Deviation (FDD), Mean Vertex Error (MVE), Diversity, Mean Estimate Error (MEE), and Coverage Error (CE). Additionally for 3DMEAD dataset, a user study was conducted to assess subjective metrics, specifically lip-sync accuracy and realism, rated on a 7-point scale. As presented in Tables 3 and 4, a lower value indicates better performance for LVE, FDD, MVE, MEE, and CE, while higher values are preferred for Diversity, Lip-sync, and Realism. The mean ratings of perceptual analysis for each model and the ground truth are presented in Table 4. Furthermore, Figure 2 shows the objective metric plots for the non-deterministic models. We present below the results analysis per dataset.

BIWI Dataset For BIWI dataset, FaceXHuBERT demonstrates the best performance in terms of LVE and FDD, achieving the lowest values for both metrics. CodeTalker produced the lowest MVE and second lowest FDD, demonstrating good performance for full face and upper face vertex accuracy. For non-deterministic models, FaceDiffuser outperforms CodeTalker-ND in terms of MEE and CE. On the other hand, CodeTalker-ND, scores higher in diversity metric in comparison to FaceDiffuser demonstrating a trade-off between accuracy and diversity.

Multiface Dataset On Multiface dataset, FaceDiffuser excelled by achieving the lowest values for LVE, MVE, MEE, and CE, demonstrating its good generalization across both the BIWI and Multiface datasets. FaceXHuBERT achieved the best FDD score but struggled with LVE and MVE compared to FaceDiffuser. CodeTalker-ND led in diversity but recorded significantly higher in all vertex error values, reflecting reduced accuracy in lip synchronization and vertex prediction, showing a similar accuracy-diversity trade-off observed for BIWI dataset results.

3DMEAD Dataset ProbTalk3D emerged as the best-performing model in terms of objective metrics on this dataset, while FaceXHuBERT excelled in perceptual analysis. ProbTalk3D achieved the lowest values for LVE, FDD, MVE, MEE, and CE, and also recorded the highest Diversity score, highlighting its ability to generate the most varied outputs among the models. FaceDiffuser delivered competitive results in LVE and FDD while also achieving relatively high scores in the user study for both lip-sync and realism. FaceXHuBERT led in subjective metrics, achieving the highest scores for lip-sync and realism, outperforming most models in user-based evaluations, although it struggled with objective metrics. Similar to the other datasets, CodeTalker-ND showed a higher Diversity score than FaceDiffuser but ranked lower in all vertex error computations, again showing an accuracy-diversity trade-off. FaceFormer showed moderate performance across objective metrics and achieved decent scores for lip-sync and realism. CodeTalker, while performing better than most models in LVE and MVE, achieved the lowest subjective scores for lip-sync and realism. Finally, the Ground Truth animations scored the highest overall for both lip-sync and realism in subjective evaluations.

Accuracy-Diversity Trade-off As reported in [WHY24] and according to the results obtained in our study, for non-deterministic approaches, a trade-off exists between accuracy and diversity. As models produce diverse outputs, the vertex errors increase, as presented in Figure 2a for lip vertex error (with the exception of ProbTalk3D trained on 3DMEAD). In Figure 2b, we can observe a similar trade-off pattern where models with lower FDD (i.e. FaceDiffuser) achieve lower diversity than CodeTalker-ND. The accuracy related metrics compare the generated animation with the ground truth data while the diversity metric computes how different each generated sample is given the same input to the trained model. A randomly generated sample from a non-deterministic method could produce visually appealing animation but might also lead to a higher error value when compared to the single ground truth sample available in the dataset. Datasets used in this research field contain only one recorded sample for each piece of textual content to be spoken. This limits diversity analysis and appropriate evaluation of



(a) *LVE (X-axis) and Diversity (Y-axis)*

(b) *FDD (X-axis) and Diversity (Y-axis)*

Figure 2: *Objective metrics plots showing the metric values trained with the non-deterministic benchmark models and corresponding datasets. Different colors represent datasets while different shapes represent models. Figure 2a shows the plot of LVE and Diversity whereas Figure 2b, FDD and Diversity.*

Dataset: 3DMEAD								
Model	LVE (↓) $\times 10^{-4}$	FDD (↓) $\times 10^{-5}$	MVE (↓) $\times 10^{-3}$	Diversity (↑) $\times 10^{-3}$	MEE (↓) $\times 10^{-4}$	CE (↓) $\times 10^{-4}$	Lip-sync (↑) (7-point scale rating)	Realism (↑)
FaceDiffuser [SHY23]	0.894	0.091	1.324	0.045	0.885	0.878	3.861	3.194
CodeTalker-ND [XXZ*23]	3.133	0.309	3.478	0.170	3.159	3.020	3.753	3.030
ProbTalk3D [WHY24]	**0.604**	**0.041**	**0.724**	**0.327**	**0.555**	**0.523**	3.671	3.118
CodeTalker [XXZ*23]	1.598	0.204	1.712	-	-	-	3.574	2.992
FaceFormer [FLS*22]	2.027	0.066	2.855	-	-	-	3.858	3.335
FaceXHuBERT [HY23]	2.969	0.085	3.143	-	-	-	**4.020**	**3.345**
Ground truth	-	-	-	-	-	-	4.035	3.375

Table 4: *Objective Evaluation Metrics together with user study results for the Models on the 3DMEAD Dataset*

generative probabilistic models. As models generate increasingly diverse outputs, vertex errors also rise, as they are compared with just one ground truth sample, which is a limitation imposed by the datasets.

5. Discussion

Based on the results obtained from our experiment, we created ranking tables for the three datasets as presented in Tables 5 and 6. These ranking tables reveal that only the objective metrics specifically designed for non-deterministic approaches (i.e. Diversity, MEE and CE) demonstrate consistency across three datasets. The other objective metrics (i.e. LVE, FDD and MVE) and the subjective metrics (i.e lip-sync and realism) do not demonstrate consistent rankings across the datasets showing that the model performance do not generalize when the dataset changes. Furthermore, the assumption of subjective metrics- (i) Lip-sync being related to LVE, MEE, CE and (ii) Realism being related to FDD, MVE, Diversity, does not hold as observed in Table 6.

One of the critical revelations in our study is the lack of generalizability of the model performances in terms of the objective metrics across different datasets. Although certain models demonstrate strong performance on specific datasets, their success does not necessarily translate to others, which indicates that current methods are tailored to specific data distributions. This raises concerns about the robustness and adaptability of the models when applied to new or unseen data. It underscores the need for further investigations into how well models generalize across a broader range of datasets, particularly when tasked with diverse facial expressions, speech styles, and cultural variations that real-world applications demand. Another key observation is the absence of a direct relationship between objective and subjective metrics, pointing to a disconnect between technical accuracy and human perception. For example, FaceXHuBERT, which ranks poorly in Lip Vertex Error (LVE) and Mean Vertex Error (MVE) for 3DMEAD, surprisingly achieves the highest scores in the perceptual user study in terms of both lip-sync and realism. This inconsistency suggests that while objective metrics like LVE and MVE are valuable for quantifying aspects such as geometric accuracy, they may not fully capture the nuances that users perceive as contributing to the quality of facial animations. This highlights the need for a deeper understanding of what objective metrics truly represent and how they relate to user experiences.

Dataset: BIWI						
Rank	LVE	FDD	MVE	Diversity	MEE	CE
1	FaceXHuBERT	FaceXHuBERT	CodeTalker	CodeTalker-ND	FaceDiffuser	FaceDiffuser
2	CodeTalker	CodeTalker	FaceXHuBERT	FaceDiffuser	CodeTalker-ND	CodeTalker-ND
3	FaceFormer	FaceDiffuser	FaceDiffuser	-	-	-
4	FaceDiffuser	FaceFormer	FaceFormer	-	-	-
5	CodeTalker-ND	CodeTalker-ND	CodeTalker-ND	-	-	-
Dataset: Multiface						
1	FaceDiffuser	FaceXHuBERT	FaceDiffuser	CodeTalker-ND	FaceDiffuser	FaceDiffuser
2	FaceFormer	FaceDiffuser	FaceFormer	FaceDiffuser	CodeTalker-ND	CodeTalker-ND
3	CodeTalker	CodeTalker	FaceXHuBERT	-	-	-
4	FaceXHuBERT	FaceFormer	CodeTalker	-	-	-
5	CodeTalker-ND	CodeTalker-ND	CodeTalker-ND	-	-	-

Table 5: *Ranking of the models per objective evaluation metrics in columns for BIWI and Multiface datasets.*

Dataset: 3DMEAD								
Rank	LVE	FDD	MVE	Diversity	MEE	CE	Lip-sync	Realism
1	ProbTalk3D	ProbTalk3D	ProbTalk3D	ProbTalk3D	ProbTalk3D	ProbTalk3D	FaceXHuBERT	FaceXHuBERT
2	FaceDiffuser	FaceFormer	FaceDiffuser	CodeTalker-ND	FaceDiffuser	FaceDiffuser	FaceDiffuser	FaceFormer
3	CodeTalker	FaceXHuBERT	CodeTalker	FaceDiffuser	CodeTalker-ND	CodeTalker-ND	FaceFormer	FaceDiffuser
4	FaceFormer	FaceDiffuser	FaceFormer	-	-	-	CodeTalker-ND	ProbTalk3D
5	FaceXHuBERT	CodeTalker	FaceXHuBERT	-	-	-	ProbTalk3D	CodeTalker-ND
6	Codetalker-ND	CodeTalker-ND	CodeTalker-ND	-	-	-	CodeTalker	CodeTalker

Table 6: *Ranking of the models per evaluation metrics in columns for 3DMEAD dataset.*

5.1. Current Limitations

Our study highlights several critical limitations in the field of speech-driven 3D facial animation generation. The first major challenge lies in the limitations of publicly available datasets. Existing datasets often lack sufficient variation and diversity. Additionally, the quality of data in these datasets is often suboptimal, with many relying on pseudo ground truth data reconstructed from 2D videos that is subject to vision based reconstruction loss. Furthermore, current datasets provide limited support for multi-language coverage, restricting the applicability of models across diverse multi-cultural contexts in 3D interactive applications such as social XR.

The second set of challenges pertains to the objective evaluation metrics. Metrics such as vertex error serve only as proxy measurements and fail to comprehensively capture perceptual nuances. These metrics lack identity- and emotion-specific evaluations, leaving key aspects of animation quality unaddressed. Additionally, there remains a disconnect between objective and subjective evaluations, undermining the reliability of current assessment frameworks.

Thirdly, subjective evaluation methodologies face their own set of shortcomings. Existing approaches often oversimplify evaluation questions, focusing narrowly on realism and lip-sync accuracy while neglecting critical aspects such as emotion and identity perception. Sample sizes for perceptual studies are often limited, with insufficient representation of varied demographic groups. Moreover, perceptual evaluations frequently exclude textured animations, which are essential for assessing visual realism in practical applications. Finally, there is a lack of ecologically valid evaluations that consider the contextual requirements of target applications, such as those in gaming or virtual reality environments.

Furthermore, many models struggle to generalize effectively to unseen subjects or lack mechanisms for facial motion retargeting, which is crucial for interactive 3D applications. Moreover, existing approaches offer limited controllability and editability of generated animations. Current evaluations of the methods often overlook these broader practical aspects, instead focus predominantly on machine learning performance.

5.2. Suggestions

Addressing these challenges requires a more standardized approach within the field. Currently, the lack of consensus on how to evaluate speech-driven 3D facial animation models has led to inconsistent evaluation methodologies across studies, making it difficult to compare results. To facilitate fairer comparisons and ensure more robust evaluations, we urge the research community to work towards establishing a standardized evaluation procedure that can be widely adopted. At the very least, this should involve the use of common benchmark datasets, dataset splits and a carefully selected set of objective metrics that reflect the diverse aspects of model performance. Moreover, it may be necessary to develop a suite of metrics tailored to different types of models or approaches rather than applying all available metrics indiscriminately, as this can dilute meaningful insights into the strengths and weaknesses of each method. For instance, [YRC*24] discusses about the limitations

of the maximal lip vertex error (LVE) for evaluating probabilistic methods and proposed MEE and CE instead.

Regarding the datasets, we recommend the use of publicly available datasets— BIWI, Multiface, and 3DMEAD due to their complementary strengths in speech-driven 3D facial animation research. The BIWI dataset offers high-resolution face mesh topology, which ensures detailed facial expressivity and provides a well-balanced set of samples covering both neutral and expressive conditions. Although the full Multiface dataset comprising 250 subjects is not publicly accessible, the available subset of 13 subjects is sufficient for research purposes. 3DMEAD, on the other hand, presents a large-scale dataset reconstructed from 2D videos of the MEAD dataset, featuring pseudo ground truth animation data annotated emotions. A key advantage of 3DMEAD is its compatibility with both low-dimensional parametric data and high-dimensional vertex-based data, as it follows the FLAME 3D morphable model (3DMM), facilitating seamless conversion between morphable model parameters and 3D vertices.

We further encourage the community to develop and construct a dataset designed specifically to facilitate robust diversity analysis. Human motion, including facial motion, is inherently non-deterministic, yet existing datasets typically contain only a single random sample of each uttered textual content. This limitation prevents a comprehensive evaluation of non-deterministic methods in terms of their ability to generate diverse animations. A dataset comprising multiple samples of the same text being uttered would enable a deeper analysis of the diversity within the ground truth data distribution and support the establishment of objective metrics tailored to evaluate such methods effectively. Additionally, we urge new and future publications to make their respective newly constructed datasets publicly available for research purposes.

To ensure fairness and reproducibility in evaluations using objective metrics, we recommend utilizing commonly adopted metrics from recent publications, particularly those with publicly available open-source implementations. Furthermore, we strongly encourage future works to provide open-source implementations for any newly proposed metrics. This will enable the community to adopt such metrics effectively and avoid potential issues arising from re-implementation. For deterministic methods, metrics such as vertex errors (LVE, MVE) and FDD should be used in conjunction with qualitative evaluations of the generated animations, as they should not serve as the sole criteria for model assessment. In contrast, for non-deterministic or probabilistic models, as employed in our study, metrics like MEE, CE, and Diversity are more appropriate, as they account for multiple generations rather than evaluating a single random output.

Furthermore, the process of conducting perceptual user studies demands greater scrutiny. The current methodologies, often based on A/B testing, is subject to critics for their lack of reliability and consistency, especially when user preferences can be influenced by random or subjective factors. A more rigorous and scientifically grounded approach to perceptual studies is needed to obtain reliable user feedback on facial animation quality. To achieve this, we believe collaboration between technical researchers (from both machine learning and computer graphics) and experts in Human-Computer Interaction (HCI) could be highly beneficial. HCI researchers bring a wealth of knowledge on designing user studies and understanding human perception, which can help refine the methods used to evaluate speech-driven 3D facial animations. By teaming up, these interdisciplinary collaborations can lead to the discovery of more robust, reproducible, and insightful approaches for conducting perceptual studies, ultimately improving the evaluation process for such technical works.

Future research can benefit from adopting the benchmark framework established in this study to enable fairer comparisons and to gain deeper insights into the alignment between objective and subjective evaluations. Standardized benchmarking can ensure transparency and robustness in assessing model performance across different approaches. Additionally, organizing dataset challenges at major graphics and machine learning conferences, or establishing online leaderboards, could further drive progress by encouraging not only academic but also industry participation and fostering collaboration. These initiatives would potentially promote the adoption of standardized datasets and evaluation protocols.

6. Conclusion

In this work, we presented a comprehensive evaluation of both deterministic and non-deterministic models for speech-driven 3D facial animation using a curated set of objective and subjective metrics. Our study aimed to address the growing need for standardized evaluation methodologies by benchmarking several state-of-the-art models across multiple datasets, including BIWI, Multiface, and 3DMEAD. The results demonstrate a clear disparity in model performance across datasets, underscoring the challenges of generalizability in this domain. While some models performed well in subjective user evaluations, they did not always align with objective metrics, highlighting the need for metrics that better capture human perception of animation quality. Additionally, different from the recent benchmarking work [YRC*24], we included objective metrics that appear in majority of recent publications in our benchmark experiment and further investigate the relation between objective and subjective metrics via a perceptual user study. For the data-driven task of synthesizing facial animation, we believe perceptual user studies to be an essential complement to the objective metrics, bridging the gap between technical accuracy and user experience. The findings of our study reveal the necessity of establishing a standardized framework for evaluating speech-driven 3D facial animation synthesis models. This includes- using a common set of datasets, consistent dataset splits and metrics tailored to specifics of both deterministic and non-deterministic approaches. In conclusion, by standardizing the evaluation process and by bridging the gap between objective and subjective assessment, future research on speech-driven 3D facial animation synthesis can achieve more consistent, fairly comparable and real world applicable outcomes.

References

[ALTT22] AYLAGAS M. V., LEON H. A., TEYE M., TOLLMAR K.: Voice2face: Audio-driven facial and tongue rig animations with cvaes. In *EUROGRAPHICS SYMPOSIUM ON COMPUTER ANIMATION (SCA 2022* (2022). 1

[ANBH23] ALEXANDERSON S., NAGY R., BESKOW J., HENTER G. E.: Listen, denoise, action! audio-driven motion synthesis with

diffusion models. *ACM Trans. Graph. 42*, 4 (2023), 1–20. doi:10.1145/3592458. 3

[ATDN24] ANEJA S., THIES J., DAI A., NIESSNER M.: Facetalk: Audio-driven motion diffusion for neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 21263–21273. 1, 2, 3

[AZL] AO T., ZHANG Z., LIU L.: Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Trans. Graph.* doi:10.1145/3592097. 3

[BZMA20] BAEVSKI A., ZHOU Y., MOHAMED A., AULI M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems 33* (2020), 12449–12460. 2, 5

[CBL*19] CUDEIRO D., BOLKART T., LAIDLAW C., RANJAN A., BLACK M. J.: Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10101–10111. 1, 2

[CDA*24] CHHATRE K., DANĚČEK R., ATHANASIOU N., BECHERINI G., PETERS C., BLACK M. J., BOLKART T.: AMUSE: Emotional speech-driven 3D body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 1942–1953. URL: https://amuse.is.tue.mpg.de. 3

[CLW*24] CHEN J., LIU Y., WANG J., ZENG A., LI Y., CHEN Q.: Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 7352–7361. 3

[CYvdS19] CHARALAMBOUS C., YUMAK Z., VAN DER STAPPEN A.: Audio-driven emotional speech animation for interactive virtual characters. *Computer Animation and Virtual Worlds 30* (2019). 2

[DBB22] DANECEK R., BLACK M. J., BOLKART T.: EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 20311–20322. 2

[DCT*23] DANĚČEK R., CHHATRE K., TRIPATHI S., WEN Y., BLACK M., BOLKART T.: Emotional speech-driven animation with content-emotion disentanglement. ACM. URL: https://emote.is.tue.mpg.de/index.html, doi:10.1145/3610548.3618183. 1, 2, 3, 4

[EST*20] EGGER B., SMITH W. A. P., TEWARI A., WUHRER S., ZOLLHOEFER M., BEELER T., BERNARD F., BOLKART T., KORTYLEWSKI A., ROMDHANI S., THEOBALT C., BLANZ V., VETTER T.: 3d morphable face models—past, present, and future. *ACM Trans. Graph. 39*, 5 (jun 2020). URL: https://doi.org/10.1145/3395208, doi:10.1145/3395208. 2

[FFBB21] FENG Y., FENG H., BLACK M. J., BOLKART T.: Learning an animatable detailed 3D face model from in-the-wild images. vol. 40. URL: https://doi.org/10.1145/3450626.3459936. 2, 4

[FGR*10] FANELLI G., GALL J., ROMSDORFER H., WEISE T., VAN GOOL L.: A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia 12*, 6 (2010), 591–598. 2, 4

[FLS*22] FAN Y., LIN Z., SAITO J., WANG W., KOMURA T.: Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18770–18780. 2, 4, 5, 8, 9

[GKG*24] GIEBENHAIN S., KIRSCHSTEIN T., GEORGOPOULOS M., RÜNZ M., AGAPITO L., NIESSNER M.: Mononphm: Dynamic head reconstruction from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 10747–10758. 2, 3

[GMW*23] GURURANI S., MALLYA A., WANG T.-C., VALLE R., LIU M.-Y.: Space: Speech-driven portrait animation with controllable expression. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct. 2023), IEEE, p. 20857–20866. URL: http://dx.doi.org/10.1109/ICCV51070.2023.01912, doi:10.1109/iccv51070.2023.01912. 2, 3

[GYY*23] GAN Y., YANG Z., YUE X., SUN L., YANG Y.: Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2023), pp. 22634–22645. 2, 3

[HBT*21] HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R., MOHAMED A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 29* (2021), 3451–3460. 5

[HCC*14] HANNUN A., CASE C., CASPER J., CATANZARO B., DIAMOS G., ELSEN E., PRENGER R., SATHEESH S., SENGUPTA S., COATES A., ET AL.: Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014). 2

[HY23] HAQUE K. I., YUMAK Z.: Facexhubert: Text-less speech-driven e(x)pressive 3d facial animation synthesis using self-supervised speech representation learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23)* (New York, NY, USA, 2023), ACM. URL: https://doi.org/10.1145/3577190.3614157, doi:10.1145/3577190.3614157. 1, 2, 5, 8, 9

[JAL23] Jali research, 2023. https://jaliresearch.com/. 2

[JZW*22] JI X., ZHOU H., WANG K., WU Q., WU W., XU F., CAO X.: Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings* (New York, NY, USA, 2022), SIGGRAPH '22, Association for Computing Machinery. URL: https://doi.org/10.1145/3528233.3530745, doi:10.1145/3528233.3530745. 2

[KAL*17] KARRAS T., AILA T., LAINE S., HERVA A., LEHTINEN J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph. 36*, 4 (jul 2017). URL: https://doi.org/10.1145/3072959.3073658, doi:10.1145/3072959.3073658. 2, 3

[KCP*24] KIM J., CHO J., PARK J., HWANG S., KIM D. E., KIM G., YU Y.: Deeptalk: Dynamic emotion embedding for probabilistic speech-driven 3d face animation, 2024. URL: https://arxiv.org/abs/2408.06010, arXiv:2408.06010.

[LBB*17] LI T., BOLKART T., BLACK M. J., LI H., ROMERO J.: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph. 36*, 6 (nov 2017). URL: https://doi.org/10.1145/3130800.3130813, doi:10.1145/3130800.3130813. 2, 4

[LKP*21] LI J., KANG D., PEI W., ZHE X., ZHANG Y., HE Z., BAO L.: Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 11273–11282. doi:10.1109/ICCV48922.2021.01110. 3

[LR18] LIVINGSTONE S. R., RUSSO F. A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one 13*, 5 (2018), e0196391. 3

[LZB*24] LIU H., ZHU Z., BECHERINI G., PENG Y., SU M., ZHOU Y., ZHE X., IWAMOTO N., ZHENG B., BLACK M. J.: EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (June 2024). 3

[NJH*22] NG E., JOO H., HU L., LI H., , DARRELL T., KANAZAWA A., GINOSAR S.: Learning to listen: Modeling non-deterministic dyadic facial motion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022). 3

[NRB*24] NG E., ROMERO J., BAGAUTDINOV T., BAI S., DARRELL T., KANAZAWA A., RICHARD A.: From audio to photoreal embodiment: Synthesizing humans in conversations. In *IEEE Conference on Computer Vision and Pattern Recognition* (2024). 3

[PWS*23] PENG Z., WU H., SONG Z., XU H., ZHU X., LIU H., HE J., FAN Z.: Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision* (2023). 2, 3

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G., SUTSKEVER I.: Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (18–24 Jul 2021), Meila M., Zhang T., (Eds.), vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html. 3

[RPZK23] REN Z., PAN Z., ZHOU X., KANG L.: Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), IEEE, pp. 1–5. 3, 6

[RZW*21] RICHARD A., ZOLLHÖFER M., WEN Y., DE LA TORRE F., SHEIKH Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 1173–1182. 2, 4

[SHY23] STAN S., HAQUE K. I., YUMAK Z.: Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), November 15–17, 2023, Rennes, France* (New York, NY, USA, 2023), ACM. URL: https://doi.org/10.1145/3623264.3624447, doi:10.1145/3623264.3624447. 1, 2, 3, 4, 5, 6, 8, 9

[SLY*24] SUN Z., LV T., YE S., LIN M., SHENG J., WEN Y.-H., YU M., LIU Y.-J.: Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Trans. Graph. 43*, 4 (jul 2024). URL: https://doi.org/10.1145/3658221, doi:10.1145/3658221. 1, 3, 6

[SVH*24] STYPUŁKOWSKI M., VOUGIOUKAS K., HE S., ZIĘBA M., PETRIDIS S., PANTIC M.: Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (January 2024), pp. 5091–5100. 2

[SWJ*24] SONG W., WANG X., JIANG Y., LI S., HAO A., HOU X., QIN H.: Expressive 3d facial animation generation based on local-to-global latent diffusion. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–11. doi:10.1109/TVCG.2024.3456213. 1, 3, 4

[TACT23] THAMBIRAJA B., ALIAKBARIAN S., COSKER D., THIES J.: 3diface: Diffusion-based speech-driven 3d facial animation and editing, 2023. URL: https://arxiv.org/abs/2312.00870, arXiv:2312.00870. 1, 6

[THA*23] THAMBIRAJA B., HABIBIE I., ALIAKBARIAN S., COSKER D., THEOBALT C., THIES J.: Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 20621–20631. 2

[TKY*17] TAYLOR S., KIM T., YUE Y., MAHLER M., KRAHE J., RODRIGUEZ A. G., HODGINS J., MATTHEWS I.: A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 1–11. 2

[TMTM12] TAYLOR S. L., MAHLER M., THEOBALD B.-J., MATTHEWS I.: Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Goslar, DEU, 2012), SCA '12, Eurographics Association, p. 275–284. 2

[TRG*22] TEVET G., RAAB S., GORDON B., SHAFIR Y., COHEN-OR D., BERMANO A. H.: Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022). 3

[WHY24] WU S., HAQUE K. I., YUMAK Z.: Probtalk3d: Non-deterministic emotion controllable speech-driven 3d facial animation synthesis using vq-vae. In *The 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games (MIG '24), November 21–23, 2024, Arlington, VA, USA* (New York, NY, USA, 2024), ACM. URL: https://doi.org/10.1145/3677388.3696320, doi:10.1145/3677388.3696320. 1, 2, 3, 4, 5, 6, 8, 9

[WWS*20] WANG K., WU Q., SONG L., YANG Z., WU W., QIAN C., HE R., QIAO Y., LOY C. C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV* (August 2020). 2, 3, 4

[WZA*22] WUU C.-H., ZHENG N., ARDISSON S., BALI R., BELKO D., BROCKMEYER E., EVANS L., GODISART T., HA H., HYPES A., ET AL.: Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243* (2022). 2, 4

[XXZ*23] XING J., XIA M., ZHANG Y., CUN X., WANG J., WONG T.-T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. *arXiv preprint arXiv:2301.02379* (2023). 1, 2, 3, 4, 5, 6, 8, 9

[XZZ*23] XU C., ZHU J., ZHANG J., HAN Y., CHU W., TAI Y., WANG C., XIE Z., LIU Y.: High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), IEEE, p. 6609–6619. URL: http://dx.doi.org/10.1109/CVPR52729.2023.00639, doi:10.1109/cvpr52729.2023.00639. 2, 3

[YLL*23] YI H., LIANG H., LIU Y., CAO Q., WEN Y., BOLKART T., TAO D., BLACK M. J.: Generating holistic 3d human motion from speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 469–480. 3

[YRC*24] YANG K. D., RANJAN A., CHANG J.-H. R., VEMULAPALLI R., TUZEL O.: Probabilistic speech-driven 3d facial motion synthesis: New benchmarks methods and applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 27294–27303. 1, 2, 3, 6, 7, 10, 11

[ZBT22] ZIELONKA W., BOLKART T., THIES J.: Towards metrical reconstruction of human faces. In *Computer Vision – ECCV 2022* (Cham, Oct. 2022), vol. 13 of *Lecture Notes in Computer Science, 13673*, Springer, pp. 250–269. doi:10.1007/978-3-031-19778-9_15. 4

[ZCW*23] ZHANG W., CUN X., WANG X., ZHANG Y., SHEN X., GUO Y., SHAN Y., WANG F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 8652–8661. 2

[ZLDF21] ZHANG Z., LI L., DING Y., FAN C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 3661–3670. 3

[ZLZ*24] ZHAO Q., LONG P., ZHANG Q., QIN D., LIANG H., ZHANG L., ZHANG Y., YU J., XU L.: Media2face: Co-speech facial animation generation with multi-modality guidance. In *ACM SIGGRAPH 2024 Conference Papers* (New York, NY, USA, 2024), SIGGRAPH '24, Association for Computing Machinery. URL: https://doi.org/10.1145/3641519.3657413, doi:10.1145/3641519.3657413. 1, 2, 3

[ZXL*18] ZHOU Y., XU Z., LANDRETH C., KALOGERAKIS E., MAJI S., SINGH K.: Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG) 37*, 4 (2018), 1–10. 2