Utrecht University
Department of Information and Computing Science

Evaluation metrics for deep learning methods for 3D facial animation synthesis

ALKIVIADIS PAVLOU
2025930

Supervisors:
Dr. ZERRIN YUMAK
KAZI HAQUE MSc

Second Examiner:
Dr. ÖNAL ERTUGRUL

Contents

Abstract

The recent advancements in the research field of Audio-Driven Facial Animations have provided both new state-of-the-art approaches and more topics to be discovered in depth. One of the main factors of any new method is the ability to evaluate its achievements accurately and representatively. In Audio-Driven Facial Animations, this is achieved using evaluation metrics that either objectively or subjectively compare state-of-the-art models. As expected from an early-stage research field, the evaluation section has been following the typical path focusing on simple metrics in most cases. This study aims to clarify the process for future researchers through an in-depth analysis of evaluation metrics. Another important aspect of this study is the understanding of the different approaches deterministic and non-deterministic models need in terms of the evaluation of the results. Apprehending the nature of facial animations will also be examined by exploring ground truth datasets in various ways.

1 Introduction

The use of facial animations has been important and widespread in recent years with emerging research and industries aiming to create believable and efficient versions of them. Capturing minor details of facial expressions or emotions conveyed through animations can significantly impact the outcome of a production. The human connection that people can create towards a virtual character can be highly affected by the realism and the accuracy of the facial motions.

Often, particularly in higher-budget productions, the facial animations are obtained using facial capture equipment that is easily and accurately passed through. The gathering is usually efficient in terms of capturing the motions and parsing them to the virtual character. The accuracy of this method comes with an obvious disadvantage, which is the high budget needed for the motion and facial capturing equipment as well as the software to accurately encapsulate the animations. Low-budget productions do not have the opportunity to capture realistic facial animations due to the lack of hardware for motion capture. Additionally, the time and expertise needed for multiple takes, minor script changes, and the use of actors for different languages into which the production will be translated, further exacerbate time and budget constraints.

In many industries, smaller productions face challenges regarding accurate and efficient 3D facial animations. To address this, methods that use data solely collected for this purpose have emerged as the most useful and accurate way of producing these animations. The model mentioned makes the process faster and more accessible to anyone. Recent research and advancements in the field, have led to highly accurate and realistic facial animations. With further research, the results may soon reach the level of traditional motion capture processes. However, producing 3D facial animations using this method is not easy and constant research is required to improve the methods. Challenges of data collection still exist, which require similar hardware and software for training state-of-the-art models. Additionally, there are numerous branches of research on this topic that sometimes divide efforts, despite being closely connected.

The evaluation of models is a crucial factor in the field of audio-driven facial animations. However, evaluating these models can be a complex task due to the difficulty in determining what is more suitable. Unlike other tasks where metrics can be used for evaluation, audio-driven facial animations require a more careful approach. Creating accurate and realistic animations is challenging, which means the evaluation of these animations is equally difficult. This project

aims to address this gap in the field, by providing accurate and multi-purpose evaluation metrics. While there are commonly used evaluation metrics, they fall short in terms of providing a complete understanding of the strengths and weaknesses of a certain method. Evaluation metrics are not just for benchmarking, they also help in understanding the key components, advantages and possible improvements of a method.

The way the evaluation happens varies depending on the model that is evaluated. That leads to the first classification that will also be explored later in Section 3. There are different ways to evaluate audio-driven 3D facial animations, which will be our focus. Additionally, numerous other methods exist to evaluate other categories such as body animations and 2D talking heads which are also useful to our understanding. The body animation methods focus on estimating the realistic movement of a body and are a completely different area, though its literature could provide useful insight. 2D talking heads generally provide less complex results than 3D facial animations, as they usually use generative models on images to simulate lip sync and facial expressions, with that being done in two dimensions. The most in-depth way to capture facial expressions is the audio-driven 3D facial animations that work on a more complex spectrum using three dimensions, able to capture the smaller details.

Furthermore, there is another important separation that can be done for the 3D facial animation methods, that is between deterministic and non-deterministic models. Deterministic are the models that while using the same data will produce the same results every time, whereas non-deterministic are the opposite while using the same input. That is one huge factor that affects the whole methodology of the models and how they are evaluated, as the aim is completely different. With the use of non-deterministic models, the user expects a more diverse spectrum of results but, an accurate and close to ground truth result from deterministic models. This difference between the models and their goals necessitates different evaluation approaches, as the expectations and the results differ.

The evaluation metrics themselves can be identified in two other different categories: objective and subjective metrics. Objective metrics are quantitative measures, often derived from computational analyses, providing a numerical assessment of a model's performance. These metrics aim to capture specific aspects of facial animation, such as accuracy and efficiency. On the other hand, subjective metrics involve qualitative assessments, often gathered through user studies or surveys, reflecting the perceptual qualities of synthesized facial animations. The inclusion of both objective and subjective metrics in the evaluation process ensures a comprehensive understanding of a model's capabilities, addressing the nature of facial expressions and the connection of human perception. This dual approach is crucial for advancing the field, as it not only quantifies technical performance but also captures the subjective user experience, contributing to a more holistic and meaningful evaluation of audio-driven 3D facial animation synthesis models.

After examining the impact and importance of metrics, the need for evaluation that covers many areas and puts to the test multiple different factors becomes obvious. That is where the limitation of the existing literature comes in, as most of the recent studies, especially on 3D facial animations, do not use metrics that test different parts of the results. The challenge with most metrics is the production of one-dimensional results struggling to evaluate the strengths or the factors that affect negatively the models. Moreover, some of the important limitations also contain the lack of consistency between subjective and objective metrics in some cases as well as some lack of in-depth analysis of their correlation. The main focus of those metrics has been the

lip animations and their accuracy, with the evaluation of other parts being limited. Additionally, even if the lip animation is evaluated using a metric, sometimes there is a lack of in-depth analysis of important aspects such as the representation of correct lip closures. The metrics used for the whole face similarly lack identifying the effect on specific areas of the face that are usually influenced depending on the audio. A similar discussion was done for the Evaluation of Embodied Conversational Agents in the paper by Wolfert et al. [49]. This review identifies issues in previous evaluations such as the lack of systematic reporting methods for gesture generation and evaluation steps. The findings also included the need for a shared methodology for conducting systematic evaluations and proposed a comprehensive list of questionnaire dimensions and preferred tasks. Even if the nature of the 3D facial animations and gesture generation do not align, the similarities in the evaluation section could prove useful. As both topics cover the spectrum of interaction and human perception the evaluation needs to adhere to these needs.

On 3D facial animation, there has been an identification of those limitations in the paper by Yang et al. [56] which focuses on the disadvantages of using only accuracy metrics due to the need for more holistic evaluation for such models. The need for a more standardized benchmarking system as well as a plurality in the way the analysis of the evaluation is done is also stated. Furthermore, that paper also raises another important parameter that will be discussed, which is the nature of 3D facial animations in terms of determinism and non-determinism. With methods being proposed in both categories, especially recent advancements in non-deterministic 3D facial animations, it is required to understand the nature of 3D facial animations. There have been a lot of advancements in non-deterministic models of body animation [39] [35], but the applications for facial animations are still preliminary. Getting a grasp of the differences between these two categories, as well as understanding the different needs for their evaluation and the possible different applications is key to the research of 3D facial animations.

In conclusion, the main contributions of our research can be identified as follows.

1. **Provide an understanding of whether facial animation is non-deterministic by nature by developing a comprehensive analysis of the ground truth.**
2. **Provide a complete review of the existing evaluation metrics for audio-driven 3D facial animation synthesis by completing an inventory from the most useful ones across many different models and datasets.**
3. **Compare state-of-the-art deterministic and non-deterministic methods using existing objective and subjective metrics to provide a critical look at their advantages and limitations.**

1.1 Research Questions

After understanding the need for a comprehensive analysis of the existing evaluation metrics with the aim to identify the areas which need to be developed. The previously documented aspects must be taken into account and important factors such as deterministic methods or non-deterministic methods are key parameters in both the understanding as well as the progress in the field.

In conclusion, the main objectives of this project can be identified as follows.

1. **Is facial animation deterministic or non-deterministic by nature? How can we measure that using existing 2D and 3D facial motion datasets?**
2. **What are the evaluation metrics to assess the quality of generated 3D facial animations? What are the advantages and limitations of these metrics? How can we test them using state-of-the-art approaches to speech-driven 3D facial animation synthesis?**
3. **Are the current quantitative and qualitative evaluation metrics consistent with each other? How about different metrics in the same category? What are the different metrics that are meaningful for deterministic and non-deterministic approaches?**

Overall, the main goal of this project is to create a comprehensive analysis of the topic of evaluation for Audio-Driven facial animations through multiple individual paths. We will complete a metric inventory testing evaluation metrics across multiple models and datasets and compare their results. Secondly, we will explore the datasets using various data analysis methods to identify the nature of the facial animations in terms of their determinism. We will assess the existing quantitative metrics with the results of the qualitative evaluation to identify possible differences and the causes behind them.

2 Related Work

This section provides an overview of existing methods in 3D facial animation, 2D talking heads, and body animations. This exploration aims to establish a comprehensive understanding of the topic. As the main focus of the project is the evaluation metrics concerning 3D facial animations, that will be the aspect that will be covered more, but other methods are also going to be introduced as they are useful to our research. The spectrum of methods that will be covered is large and contains a plurality of implementations from the proposed methods. Recognizing that the project's focus lies not only on the methods themselves but also on their evaluation, it is beneficial to explore a variety of methods to enrich our study.

2.1 Audio-driven 3D facial animation

As research has increasingly focused on audio-driven 3D facial animations in recent years, the primary objective is to understand the reasons behind this shift. An easy answer could be the realism that is provided through this type of method as well as the accuracy in the representation of the facial features and expressions. Recent state-of-the-art models proved to have introduced new ways for the architecture as well as the training of the models. One of the first advancements in the area which is still considered a benchmark by many, while the results prove it was done by Cudeiro et al. with a method that became a setting stone for the recent progress in the field [7]. With a network using a facial template and audio as input, and mainly focusing on the encoder, conditioned on subject labels, uses DeepSpeech-extracted [15] speech features. Another important addition from the paper is the dataset called VOCASET which will be explored later on, which is a 4D face dataset with scans captured and synchronized audio from 12 speakers. The method of using multi-modality or cross-modality for the topic was mainly explored by Richard

et al. [36] resulting in the use of a categorical latent space that disentangles audio-correlated and audio-uncorrelated information based on a novel cross-modality loss. This process creates an accurate lip movement as well as taking into account uncorrelated parts such as eye blinks and eyebrow motion. Multi-modality has been the main point of exploration for other papers such as the paper from Han et al. [14] which creates an encoder that employs the talking head generation architecture using Wav2Lip [33] with an identity encoder, a speech encoder, and a face decoder to extract visual and textual information from speech. Another advancement was the use of a large-scale multi-modal dataset from Wu et al. [50] with the combination of the use of a network to consider both the regional and composite natures of facial animations.

One advancement in the field came from Thambiraja et al. [41] with a method that learns identity-specific details and produces novel facial expressions matching the identity-specific speaking style and facial factors of the target. The training is done with a style-agnostic transformer on a large facial expression dataset which they use as a prior for audio-driven facial expressions. An important addition is the novel loss function which is based on bilabial consonants to ensure plausible lip closures to achieve a more realistic result. Motion prior has been explored by Xing et al. [55] in a method that uses a codebook that is learned by self-reconstruction over real facial motions and thus embedded with realistic facial motion priors. They employ a temporal autoregressive model that synthesizes facial motions from the input speech signal. Another approach to 3D facial animations using an autoregressive model along with the cross-modal multi-head attention was introduced by Fan et al. [11] which aligns the audio-motion modalities, that offers abilities to generalize to longer audio sequences.

A significant work in this area has been the integration of emotions into the animations in different natures. One attempt was done by Karras et al. [20] with a deep neural network that learns a mapping from input waveforms to the 3D vertex coordinates of a face model and discovers a compact, latent code that disambiguates the variations in facial expression that cannot be explained by the audio alone. For the emotions, the method uses the latent code during inference as an intuitive control for the emotional state of the face puppet. One more recent attempt using emotions to capture the animations of the face was done by Danecek et al. [8]. In this paper, they introduce a novel deep perceptual emotion consistency loss during training, which helps ensure that the reconstructed 3D expression matches the expression depicted in the input image. The results seem to ensure the capturing of the full spectrum of facial expressions, such as subtle or extreme emotions. One of the most recent attempts in that area has been the paper from Peng et al. [8] which introduces the emotion disentangling encoder (EDE) to disentangle the emotion and content in the speech by cross-reconstructed speech signals with different emotion labels. Similarly, their decoder is driven by disentangled identity, emotional, and content embeddings. The more diverse facial movements captured are the key point of this method, especially with the use of facial blend shapes to reconstruct plausible 3D faces.

A process that was first introduced by Sohl-Dickstein et al. [37] is the method of diffusion. The method involves a process inspired by non-equilibrium statistical physics that is applied to model complex data sets in machine learning. The authors propose an approach that involves iteratively applying a forward diffusion process to systematically and slowly destroy structure in a data distribution. This process makes the data more flexible while maintaining tractability for learning, sampling, inference, and evaluation. Subsequently, a reverse diffusion process is learned to restore structure in the data. The outcome is a generative model that is both highly flexible and computationally tractable. This approach is particularly useful for deep generative

models with a large number of layers or time steps, enabling rapid learning, sampling, and evaluation of probabilities, as well as the computation of conditional and posterior probabilities under the learned model. This has been applied to human motion by Tevet et al. [39] with their method using a transformer-based architecture, they also introduce the noised ground truth motions as an additional input to the network. Their method succeeds in generating non-deterministic animations at inference time, which is essentially the most important factor for using diffusion on motion estimation. The general notion and use of diffusion in vision have been thoroughly explained in a survey from Croitoru et al. [6]. Diffusion has been employed recently in many 3D facial animation methods as it adds to the non-deterministic spectrum of models. Park et al. [28] used diffusion for their method with a focus on achieving one-to-many using speech to achieve better results not only on lip sync but also on other facial attributes. In their paper, they showcase the alignment of lip motion by exploiting audio-mesh synchronization and masked conditioning, while also modelling identity and pose in addition to facial motions so that it can generate 3D face animation without requiring a reference identity mesh and produce natural head poses. Another recent attempt has been the work from Thambiraja et al. [40] on the topic of using diffusion with the proposal of a lightweight audio-conditioned diffusion model. Their model showcases important advantages such as training on a small 3D motion dataset, maintaining expressive lip motion output and the ability to be fine-tuned for specific subjects, requiring only a short video of the person. One of the most important works on the topic of using diffusion for 3D facial animation has been done by Stan et al. [38], with the FaceDiffuser model. Using the HuBERT model [18], to encode the input, FaceDiffuser achieves being the first to use solely diffusion method for the task of speech-driven 3D facial animation synthesis and producing a non-deterministic model. One other paper that suggests a method using HuBERT is FacexHuBERT by Haque and Yumak [16], which uses the self-supervised HuBERT model in training to incorporate both lexical and non-lexical information in the audio without using a large lexicon. Additionally, to achieve the identification of subtle facial expressions they use binary emotion conditions and speaker identity to guide the training.

In addition, the approach proposed by Aylagas et al. [45] uses a model that aims to separate the process into two parts. One part is the conditional VAE that generates mesh animations from speech and the other is mapping the animations to rig controller space. Key factors that the paper proposes also are the automated method for speech style control, a method to train a model with data from different quality levels and the method for tongue animation. Additionally, the paper from Pham et al. [31] uses a method that learns the latent representations of time-varying contextual information and effective states within the speech. That results in a more accurate representation of the motions depending on the context of the speech as it estimates the emotional intensity of the speaker. Another approach was implemented in the paper by Zhao et al. [59] and their paper presidents VividTalker, a framework that focuses on head pose and natural facial details. They use disentanglement for the facial animation into the head pose and mouth movement and encode them separately into discrete latent spaces and they generate the attributes through an autoregressive process leveraging a window-based Transformer architecture. One of the latest approaches on the topic and one of the most advanced has been done by Yang et al. [56], a paper that can be helpful in the notion of evaluation as well. In that paper they use a probabilistic model to learn a residual vector-quantized codebook for the motions and train a two-stage, probabilistic auto-regressive model to predict these codes. The paper also discusses the limitations of existing work in terms of non-deterministic models and existing evaluation.

2.2 2D Talking Heads

Another related area of research focuses on the animation of 2D Talking Heads. This area has seen also advancement recently, especially with the use of GANs. Generally, the focus of such methods is lip synchronization due to the nature of the generative models used. Even if there are differences in this field with the 3D facial animations, studying 2D Talking Heads could be proved to play a helpful part in understanding the way evaluation is done in that similar field. It has also influenced some of the metrics used for audio-driven facial animation and could help in the exploration of more novel implementations of similar existing metrics for 2D talking heads.

On that topic, the paper from Vougioukas et al. [46] achieved the creation of an end-to-end system to generate videos of talking heads, using an image of a person and an audio clip. Their method uses 3 discriminators for detailed frames, audio-visual synchronization and realistic expressions. The paper also goes in-depth to evaluate the method as well as each component. Another paper with an interesting approach was done by Li et al. [21]. In this paper, the method contains a Facial Animation generation model based on an Adversarial Network (FAAN) model which maps features of the face image and the natural speech to a public space during the encoding process and then generates a frame sequence of a talking face according to the temporal coherence features contained in the speech fragments. The approach from Lu et al. [24] aims for a more realistic result. The method could be separated into 3 stages, firstly, a deep neural network that extracts deep audio features along with a manifold projection to project the features. After that facial dynamics and motions from the projected audio features are learned. In the final stage, conditional feature maps from previous predictions are generated and sent with a candidate image set to an image-to-image translation network to synthesize photorealistic renderings. The work by Wu et al. [53] contains a methodology based on GANs. Mainly they use a representation based on a 3D geometric flow, termed facial flow, to represent the natural motion of the human face at any pose. Their synthesis framework combines the multi-scale appearance features from images and motion features from flows in a hierarchical manner, to achieve the best possible results. A more recent work by Wu et al. [51] proposes a transformer-based probabilistic mapping network that can model the variational facial animation distribution conditioned upon the input audio and autoregressively convert the audio signals into a facial animation sequence.

A survey on the topic was conducted by Chen et al. [3] with a specific focus on the evaluation and benchmarks of talking head generation. They propose a benchmark for the evaluation with standardized dataset strategies. They also do a comprehensive study and discussion of the evaluation needed to identify the appropriate metrics for each scenario and aim. Moving on to an additional survey on the research field conducted by Toshpulatov et al. [43], which puts to the test all the recently developed Deep Learning methods on the topic. By analyzing and categorizing the different techniques that exist such as the Generative Adversarial Networks, Convolutional Neural Networks and Neural Rendering Fields they identify the benefits and drawbacks of each. Additionally, during that process, they provide an in-depth look at the evaluation metrics and their utility.

Another recent addition to the research field has been emotions, as it was shown in 3D facial animations. On this spectrum, Ji et al. [19] have proposed the Emotion-Aware Motion Model (EAMM) for one-shot emotional talking faces. They also propose to render the talking face from audio-driven unsupervised motion. Another addition is the Implicit Emotion Displacement Learner which represents emotion-related facial dynamics as linearly additive displacements to

the previously acquired motion representations. On the same note, the paper from Zhang et al. [58] with SadTalker model which generates 3D motion coefficients, for the pose and expression, of the 3DMM from audio and implicitly modulates a novel 3D-aware face render for talking head generation.

To achieve our aim of analysing the differences and possible advantages that might exist in 2D face datasets we also explored the existing state-of-the-art datasets of this research field more in depth. The first one is the GRID dataset from Cooke et al. [5] which is part of a comprehensive audio-visual corpus, consisting of 34,000 sentences spoken by 34 different talkers. Each sentence follows a structured six-word format, including commands, colours, prepositions, letters, digits, and adverbs. Another useful dataset is the one presented by Cao et al. [2] in their comprehensive research providing the CREMA-D dataset. That dataset is designed for studying multimodal emotion expression and perception, featuring 7,442 clips of 91 diverse actors expressing basic emotions—happy, sad, anger, fear, disgust, and neutral. CREMA-D's extensive and diverse data supports research into the interplay of facial and vocal cues in emotion perception, enabling detailed analysis of both subtle and extreme emotional expressions.

Using the knowledge gathered from all the papers for 2D Talking Heads we can expand our understanding of evaluation for our research fields. The plurality of metrics as well as the extensive qualitative evaluation done will prove helpful for our research. Similar ways to identify benchmarks and metrics could be adapted or implemented for 3D facial animations.

2.3 3D Body animation and pose estimation

Recently, there has been notable progress in another animation domain, 3D body animation and pose estimation. This research is due to the increasing demand for realistic and accurate representations of human movements, in many different and diverse industries. The significance of 3D body animation and pose estimation lies in its ability to capture and portray the intricate dynamics of human motion, facilitating a more immersive and lifelike experience. Advanced models in this field have improved architectural aspects and transformed training methodologies, opening up new avenues for advancing the precision and authenticity of animated human figures.

One of the papers that uses Adversarial Learning to achieve realistic pose estimation was done by Yang et al. [57]. Their method employs an adversarial learning framework, the model distils knowledge from fully annotated datasets to enhance pose predictions for diverse real-world scenarios. Also, a novel multi-source discriminator, incorporating a geometric descriptor capturing pairwise relative locations and distances between body joints, is introduced to enforce valid pose generation. Another paper that uses Temporal Transformers was done by Zheng et al. [60]. In their paper, they use PoseFormer which utilizes transformer structures to comprehensively model joint relations within frames and temporal correlations across frames. This innovative design achieves accurate 3D human pose predictions for the central frame. Similarly, the paper from Cheng et al. [4] utilizes multi-scale spatial features for 2D joint predictions and multi-stride temporal convolutional networks (TCNs) for 3D joint estimation, the model adapts to diverse human appearances and motions. To enhance robustness against occlusion and accuracy, a spatio-temporal discriminator assesses the validity of predicted poses and movements based on body structures and limb motions. A really useful survey of the methods for 3D Human pose estimation was done by Wang et al. [47]. This survey evaluates the strengths and

weaknesses of methods, fostering a nuanced understanding of the field. Additionally, the survey explores commonly used benchmark datasets, conducting a comprehensive study for comparison and analysis. By explaining the current state of research in 3D human pose estimation, the study not only serves as a valuable resource for researchers but also provides insights for guiding the future design of models and algorithms in this field.

A useful review paper on the topic of audio-text-music driven body animation was done by Nyatsanga et al. [27] which explores the automatic generation of co-speech gestures, emphasizing their significance in natural human communication and applications. They organize the approaches based on input modalities, exploring systems that generate gestures from audio, text, and non-linguistic input. Additionally, they identify key research challenges in gesture generation, including data availability and quality, achieving human-like motion, grounding gestures in co-occurring speech and the environment, conducting gesture evaluation, and integrating gesture synthesis into applications. Moreover, the survey from Mourot et al. [26] on Skeleton-Based Human Animation delves into motion data representations, prevalent datasets, and enhancements to basic deep models for learning spatial and temporal patterns. The paper showcases a variety of animation techniques and elucidates how recent advancements in Deep Neural Networks and Deep Reinforcement Learning have significantly impacted the field, providing insight into the most efficient and realistic animation techniques. Furthermore, meticulously explores motion data representations, prevalent datasets, and enhancements to basic deep models for the acquisition of spatial and temporal patterns in human motion.

One of the factors that we aimed to explore more deeply as we are going to use it for our analysis and comparison is the datasets that exist for the body motion research field. The first dataset that we found used in many models is the one by Plappert et al. [32] which presents a large, open, and extensible dataset that links human motion data with natural language descriptions. The dataset combines motion capture data from multiple sources into a unified representation, independent of the capture system, and includes 3911 motions. The authors developed a web-based tool for crowd-sourcing motion annotations and used gamification to motivate participants. They also introduced a perplexity-based selection method to systematically select motions for annotation, ensuring a diverse and accurate dataset. This dataset aims to facilitate transparent and comparable research in human-robot interaction. Another useful paper is the one presenting the BABEL dataset from Punakkal et al. [34], which combines motion capture data with detailed action labels in English. The dataset contains about 43 hours of mocap sequences from the AMASS dataset and includes over 28,000 sequence labels and 63,000 frame labels, covering more than 250 unique action categories. The dataset is designed to support tasks like action recognition, temporal action localization, and motion synthesis. The authors also introduce a benchmark for 3D action recognition, highlighting the dataset's challenges due to its realistic and diverse action distribution. The last dataset that we aimed to use for our analysis is the IDEA 400 which is a motion-captured subset of the Motion-X dataset from Lin et al. [22]. The IDEA 400 dataset includes 12.5K motion sequences covering 2.6M frames. It features 400 diverse actions performed by 36 actors, capturing a wide range of human self-contact motions and expressive whole-body movements, including facial expressions and hand gestures. This dataset enhances the diversity and expressiveness of motion data by providing high-quality whole-body motion annotations.

3 Background Knowledge

This section will be used to explain the needed knowledge for the models from the papers that were refereed but it will also focus on the knowledge for the evaluation. All the important parts of the evaluation will be discussed, concerning objective and subjective evaluation. A more in-depth examination of the crucial factors influencing audio-driven 3D facial animations and the parameters impacting the models. Also, the terms deterministic models and non-deterministic, which were referred to earlier, will be explained more in-depth along with an explanation of the differences it makes in terms of evaluation.

3.1 Deterministic and Non-Deterministic Models

In this subsection we we will deep dive into the recent advanced methods for state-of-the-art models to explain the important factors that will affect our research. Focusing on audio-driven 3D facial animations and the components that also affect the evaluation differently. As stated before an important parameter of audio-driven 3D facial animations, is the distinction between deterministic and non-deterministic methods which play a crucial role in shaping the character and responsiveness of the generated facial expressions. Deterministic methods refer to approaches that produce consistent and predictable results based on specific input audio signals. These methods typically rely on predefined rules, algorithms, or deterministic mappings between audio features and facial movements. On the other hand, non-deterministic methods, or probabilistic as they are sometimes referred to, introduce an element of variability and diversity in facial animations, aiming to capture more natural and nuanced expressions. These approaches often leverage techniques where the model learns from diverse datasets to generate dynamic and contextually appropriate facial animations in response to audio cues. Basic methods falling under the deterministic category might include rule-based systems that map certain phonetic features to corresponding facial gestures, while non-deterministic methods could learn intricate patterns and dependencies between audio and facial expressions, allowing for more fluid and realistic animations. Each of the two categories has its own goals and reasons to implement.

As shown there has been diversity in terms of the methods adapting deterministic or non-deterministic approaches in state-of-the-art audio-driven 3D facial animations. It has been a discussion that will be examined further in this topic and some prior knowledge is needed to understand the approach for each of the subcategories. Traditional keyframe-based animation techniques, where animators manually specify key poses and transitions, are inherently deterministic. The animator's decisions directly determine the sequence of facial expressions, and the animation follows a predefined script. However, some of the recent methods often incorporate non-deterministic elements with techniques that introduce stochasticity into the animation process. For instance, models using random sampling during training or employing diffusion-based approaches inherently generate diverse and non-deterministic facial expressions in response to audio cues. The randomness in these methods contributes to the creation of more natural and expressive animations.

This also leads to another important parameter closely correlated with the research topic of this project: the differences in the evaluation needed for deterministic and non-deterministic methods. Deterministic methods are typically assessed using established criteria like visual fidelity, smoothness of transitions, and accuracy of specific expressions. Metrics may include measuring the accuracy of facial feature movements and the ability to faithfully reproduce pre-

defined sequences as done in [11]. In addition to those, evaluating non-deterministic methods introduces additional complexities. Metrics assessing the diversity and naturalness of generated expressions become crucial. Such metrics were recently used in [38] [59] [40] with additional variability in [56]. On the other hand, evaluation measures involve perceptual studies where human observers assess the realism and expressiveness of the animated faces. The specific identification of those metrics is an important factor in distinguishing the need for different evaluations for these two types of methods.

Additionally, the usage of discrete motion prior, such as the codebook used in CodeTalker [55] ensures that probabilistic nature in this case. The model achieves this by adopting a transformer-based VQ-VAE, to encode the facial motions into a temporal feature space and subsequently to quantize these features into the discrete codebook. This discrete representation introduces an element of variability and richness, capturing nuanced facial expressions and promoting a more natural synthesis of 3D facial animations. The codebook's finite cardinality reduces mapping ambiguity while preserving expressiveness in a context-rich latent space. On the other hand, the model from FaceFormer [11] can be categorized as a deterministic sequence-to-sequence model for speech-driven 3D facial animation. The model exhibits a clear cause-and-effect relationship between its inputs and outputs. The architecture follows a pattern, where the encoder processes raw audio into speech representations and the decoder autonomously predicts facial movements based on audio context, past facial motions, and speaker styles. The deterministic nature is evident in the autoregressive scheme during training and inference, emphasizing precision in predicting 3D facial animations. While effective in generating realistic facial expressions, the deterministic approach may limit the model's adaptability to handle uncertainties or diverse styles. Additional exploration into the presented sub-categories can be undertaken by examining Meshtalk's model [36], known for its non-deterministic nature. The model addresses the challenge of uncanny or static upper-face animation in audio-driven methods by disentangling expressive information. Specifically, MeshTalk learns a categorical latent space for facial expressions, aiming for expressiveness, categorical nature, and semantic disentanglement. The non-deterministic aspect is evident in the autoregressive sampling during inference, ensuring plausible animation of face parts uncorrelated to speech. Most of the diffusion-based models such as [38] [40] [28] have naturally resulted in non-deterministic models. Diffusion processes involve random steps or stochastic elements, making the outcomes inherently uncertain. In diffusion models, randomness plays a crucial role in the evolution of the system or the generation of samples.

3.1.1 Deterministic Methods

For the deterministic 3D facial animation models, the VOCA model by Cudeiro et al. [7] stands out as a benchmark, particularly for its distinctive techniques. This model operates as an encoder-decoder network, employing a series of techniques that contribute to its deterministic nature. Firstly, the model utilizes DeepSpeech [15] for speech feature extraction, ensuring a consistent representation of audio input. The resultant unnormalized log probabilities are then resampled using linear interpolation, forming a three-dimensional array. The deterministic encoding process involves a combination of convolutional and fully connected layers, incorporating subject labels to learn subject-specific styles. This technique, involving the encoding of subjects as one-hot vectors concatenated with speech features, ensures a deterministic style representation. Furthermore, the decoder, with a fully connected layer and linear activation, generates vertex

displacements based on weights initialized using PCA components, contributing to the overall deterministic output. Notably, the alteration of an eight-dimensional one-hot vector during inference allows for changes in the output speaking style, adding a deterministic element to the model.

Another deterministic model, FaceXHuBERT by Haque et al. [16], introduces a set of techniques that align with the deterministic paradigm. This model leverages a pre-trained HuBERT speech model as the audio encoder, ensuring deterministic audio feature extraction. The encoder, consisting of a CNN encoder, feature projection layer, positional convolution embedding layer, and 12 transformer layers, contributes to the deterministic representation of audio information. The GRU-based decoder with frozen pre-trained weights further emphasizes determinism in the generation of vertex displacements. The incorporation of subject and emotion labels as one-hot vectors linearly embedded into the network adds a deterministic aspect to the decoding process. Notably, the model's deterministic nature is validated through extensive analysis and comparison, showcasing consistent and reproducible results, particularly when utilizing the BIWI dataset.

FaceFormer, as proposed by Fan et al. [11], introduces techniques that align with the deterministic approach to 3D facial animation. This sequence-to-sequence learning model employs an encoder-decoder framework, where the encoder utilizes a self-supervised pre-trained speech model, wav2vec 2.0, for the deterministic transformation of raw audio into speech representations. The decoder, featuring periodic positional encoding (PPE) and a biased causal multi-head self-attention mechanism, adds determinism by incorporating temporal order information and capturing dependencies within the past facial motion sequence. The biased cross-modal multi-head attention mechanism further aligns audio and motion modalities in a deterministic manner. During training, the autoregressive scheme and Mean Squared Error (MSE) loss contribute to the deterministic nature of FaceFormer, ensuring consistent and realistic 3D face mesh predictions.

Imitator by Thambiraja et al. [41] is designed for person-specific 3D facial animations, introducing a set of techniques contributing to determinism. The audio encoder, utilizing the Wav2Vec 2.0 model, ensures deterministic encoding of audio inputs. The auto-regressive viseme decoder, employing a transformer architecture, consistently produces identity-agnostic viseme features, contributing to deterministic generation. The motion decoder, incorporating a style embedding layer and a motion synthesis block, adds determinism by mapping viseme features to the motion space defined by a linear deformation basis. Training objectives, including reconstruction loss, velocity loss, and lip contact loss, further enhance the deterministic nature of the Imitator. The use of the VOCASET dataset ensures a deterministic learning process, resulting in a model capable of generating person-specific 3D facial animations in a consistent and reproducible manner.

The explained deterministic models employ techniques such as consistent speech feature extraction, subject-specific encoding, frozen pre-trained weights, and structured attention mechanisms. The deterministic nature is crucial for reproducibility and reliability in 3D facial animation applications, as well as an accurate representation of the ground truth.

3.1.2 Non-Deterministic Methods

FaceDiffuser by Stan et al. [38] introduces non-determinism through a diffusion process during training. This involves iteratively adding noise sampled from a normal distribution to input frames. The model is specifically trained to predict animation data rather than noise levels, allow-

ing for diverse and stochastic predictions during inference. Additionally, the introduction of randomly sampled noise during the inference process further contributes to the non-deterministic nature of FaceDiffuser, resulting in varied and unique animation sequences.

Thambiraja et al. [40] propose 3DiFACE, a non-deterministic model that leverages a diffusion-based architecture for facial animation synthesis. The diffusion process, which predicts noise-free displacements given noisy input sequences, introduces stochastic outputs, making the generated facial animations diverse and unpredictable. Furthermore, 3DiFACE deviates from transformer-based methods by employing a 1D-convolutional network for temporal processing, introducing variability in the model's architecture compared to deterministic approaches.

CodeTalker by Xing et al. [55] adopts a non-deterministic approach through the introduction of a discrete codebook prior and quantization function. This discrete representation adds uncertainty and variability to the model, as each motion is represented as a combination of allocated items from the codebook. The training involves a straight-through gradient estimator to handle the non-differentiable nature of the quantization function, introducing stochasticity. The inclusion of a style vector during inference provides a versatile representation and adds a non-deterministic element, allowing for the control of talking styles and enhancing the model's adaptability.

ProbTalk3D by Wu et al. [52] is a newly released state-of-the-art model that was proposed by our research team in another research aiming to explore the use of non-deterministic models with the 3DMEAD dataset using the FLAME parameters of the data to train, instead of the vertice information. The paper proposes the use of VQ-VAE with adaptions on the quantizer in a two-stage approach providing evidence of outperforming recent non-deterministic models.

Lastly, the probabilistic nature of the model from the paper from Yang et al. [56] whose non-determinism relies on the residual vector-quantized codebook which is in a coarse-to-fine manner. In their process, the fixed-size codebook is employed, and residual vector quantization is applied to recursively project a vector to the nearest code in the codebook, generating a sequence of indices for the codes. The residual vector-quantized is utilized within the latent space of a 3D facial motion autoencoder, where a temporal convolutional encoder maps the target sequence to a latent embedding of the motion. The non-deterministic aspect is further emphasized by the sampling strategies proposed, aiming to balance diversity and fidelity during inference. These strategies include KNN-based sampling, code averaging, and SyncNet-based sampling, each influencing the stochasticity of the generated facial motions.

Summarizing, there are many different newly proposed techniques for non-deterministic models as a topic that is still expanding. The main contribution of those methods is the variability in the way they can incorporate non-determinism. Some examples employ diffusion processes, random noise during inference, 1D-convolutional networks, and discrete codebook priors with quantization functions. The fundamental aim of those techniques is to introduce variability, unpredictability, and adaptability into the synthesis of 3D facial animations in response to audio signals.

3.2 Evaluation

In this section, we will focus on the main topic of the project which is the evaluation of the models that were thoroughly discussed in the previous sections. Evaluation is a key component in understanding all the factors concerning 3D facial animations, 2D talking Heads and Body

Animation. As of the nature of this research, we will mainly focus on the way the evaluation is done on audio-driven 3D facial animations but similarly with the models, there are some helpful points to be gained by analyzing the other types of animations. One of the first main things that should be declared is the 4 different sub-categories of evaluation:

- **Quantitative Evaluation - Objective Metrics:** Contains objective metrics that rely on concrete, observable data.
- **Quantitative Evaluation - Ablation Study:** Systematically removing specific components within a model to assess their impact on performance.
- **Qualitative Evaluation - Subjective Metrics:** Visual comparison done by the researcher with different sequences and datasets using subjective metrics.
- **Qualitative Evaluation - Perception Study:** Conducted with user studies using subjective evaluation metrics.

The focus towards quantitative evaluation and the use of objective metrics has been imminent since the topic of facial animations is a niche topic that requires in-depth understanding and most of the time correctly analysed using numerical data. Quantitative evaluation provides a systematic and measurable approach to assess the performance of the algorithms, offering a set of metrics to measure accuracy, realism, and synchronization with audio input. Objective metrics, often rooted in mathematical formulations, yield consistent and unbiased results, reducing the potential for interpretational variations that may arise in subjective evaluations. Additionally, the use of quantitative measures facilitates the establishment of benchmarks and comparisons between different methodologies, enabling researchers to identify strengths, weaknesses, and areas for improvement with a higher degree of objectivity. With the use of quantitative evaluation and objective metrics, a robust and reliable foundation for advancing the state-of-the-art in audio-driven 3D facial animation technologies is ensured. A necessary clarification is that in certain instances, the qualitative evaluation utilizing subjective metrics is performed similarly to a perception study but conducted by different individuals. The use of subjective metrics and essentially ranking the results of the method based on them can be considered a similar approach and for some researchers perception study and their user study falls under the qualitative evaluation spectrum. While quantitative evaluation and objective metrics play a crucial role in assessing the technical aspects of audio-driven 3D facial animations, qualitative evaluation and subjective metrics are equally indispensable for capturing the nuanced and perceptual aspects of the user experience. Subjective metrics, derived from human assessments, offer insights into the emotional expressiveness, naturalness, and overall aesthetic appeal of facial animations, factors that may be challenging to quantify objectively. Qualitative evaluation allows for a more holistic understanding of how users perceive and interact with animated characters, taking into account subtle details that contribute to the overall believability of the facial expressions. Moreover, subjective metrics are valuable in situations where the ultimate goal is to enhance user engagement and immersion. By incorporating qualitative evaluation methods, researchers can gather user feedback, preferences, and emotional responses, enriching the evaluation process with a user-centric perspective and ensuring that the technology not only meets technical benchmarks but also aligns with human expectations and preferences. Moreover, it is important to establish which is the way of evaluation as a way to test the independent components of the models, and

that is Ablation Studies. Ablation studies in the context of 3D facial animation involve systematically assessing the impact of individual components within a given model or system.

3.2.1 Quantitative Evaluation and objective metrics for 3D facial animation

In this section, we will deep dive into the specifics of the evaluation metrics used for 3D facial animations to conduct the quantitative evaluation. We will look for the most commonly used metrics, some metrics that showcase potential in specific topics as well as discussing the gap that exists. There has been recently an advance in the area with new novel metrics being proposed in papers [56] [40] [55]. That advancement does not mean that research is not needed in the creation of a better understanding of what is adaptable and where. Additionally, the specific advantages of each metric are not obvious as there is a lack of clear understanding of the whole evaluation process and its benefits.

The most commonly used metric for audio-driven 3D facial animations has been the calculation of the *Lip Vertex Error (LVE)*. This metrics has been widely used in papers such as [55] [36] [30] [45] [11] [16] [38] [56] and [28]. The calculation of this metric focuses on the lip displacement comparing the generated sequence with the ground truth. Essentially, the lip error of each frame is the maximal L2 error of all lip vertices. L2 Error is used for finding the difference between two vectors in a Euclidean space with the following calculation: $MaximalL2error = max_i|u_i - v_i|$. LVE is the average of all the maximal errors of all frames in the dataset. The calculation that can be derived for the whole Lip Vertex error is the following, with x being the ground truth and $\hat{x}$ the synthesized mesh:

$$l_{vertex}(x, \hat{x}) := \max_{t,i \in lip} ||x_{ti} - \hat{x_{ti}}||_2$$

.

On a similar note to Lip Vertex Error, *Mean Vertex Error (MVE)* is essentially the same calculation but for the whole face, that metric has been used in [16] and [38]. This metric measures the deviation of all the face vertices by again the computation of the maximal L2 error for each frame, averaging all the frames to compute the Mean Vertex Error. Taking into account all the vertices of the face it is a useful addition and an interesting metric, especially for the methods that do not only focus on correct lip-syncing but also the upper-face expressions correlated to speech. The calculation can be expressed as follows:

$$f_{vertex}(x, \hat{x}) := \max_{t,i \in face} ||x_{ti} - \hat{x_{ti}}||_2$$

Another metric that proved to be used a lot in the recent methods on the field is the *Facial Dynamics Deviation (FDD)*. This metric has proved to be a useful addition to the error metrics that are commonly used with it being used in [38] and [55]. FDD measures the variation of facial dynamics for a motion sequence in comparison with that of the ground truth. Indicates how close the standard deviation of a generated sequence is compared to the observed variation in ground truth. For the calculation, it computes the difference in the standard deviation of element-wise L2 norm along the temporal axis between the predicted and ground truth non-lip vertices. The equation that can be derived is the following with $\mathrm{M}^u_{1:T}$ denoting the motions of the v-th vertex, and Su is the index set of upper-face vertices and $dyn(\cdot)$ denotes the standard deviation of the element-wise L2 norm along the temporal axis.

$$FDD(M_{1:T}, \hat{M}_{1:T}) = \frac{\sum_{u \in Su} (dyn(\mathrm{M}^u_{1:T})) - dyn(\hat{\mathrm{M}}^u_{1:T}))}{|Su|}$$

Another interesting metric that also has the factor that is mainly used for non-deterministic methods is the metric of *Diversity*. This metric could be applied to different iterations of the inference algorithm, but for the metric to also be compatible with deterministic models, it is defined across different subjects. This helps to ensure that different subjects have different neutral facial physiognomy resulting in the expected diversity. This metric is a novel metric introduced in FaceDiffuser [38] but has been influenced by another metric for human body animation diffusion models as described in [39]. The equation of this metric is presented below with S denoting the list of the training subjects and $\hat{x}_0^i$ is the predicted animation sequence conditioned on the ith subject from S.

$$Diversity = \frac{\sum_{i=1}^{|S|-1} \sum_{j=i+1}^{|S|} ||\hat{x}_0^i - \hat{x}_0^j||}{\frac{(|S|-1)\cdot|S|}{2}}$$

As expressed before there has been an attempt by recent papers to tackle the problem of the gap of the existing methods that we are discussing, alongside their implementation. One of those papers is the one from Yang et al. [56] and has been noted earlier but now we are going to take a more comprehensive look at the evaluation metrics it suggests and implements, one by one. As with almost any other state-of-the-art they implement Lip Vertex Error which they also explain is not enough. Interestingly they propose two other metrics concerning the Lip error spectrum that are closely correlated. The first one is *Coverage Error* which essentially tests the closeness of the sampling distribution of a non-deterministic model to the ground truth. The calculation is done by generating a set of samples S and computing the closest distance to the ground truth from that sample pool.

$$l_{cover} := \min_{x \in S} l_{vertex}(x, \hat{x})$$

This metric is combined with another metric proposed for the lip error which is the *Mean Estimate Error*. Mean Estimate Error assesses how close the mean of the sampling distribution is to the ground truth by computing the lip vertex error over the mean of S. The equation of Mean Estimate Error is shown below:

$$l_{mean} := l_{vertex}(x, \mathbb{E}_{\hat{S}} \hat{x})$$

The combination of those two metrics provides important context in understanding whether a non-deterministic model is capable of generating the ground truth lip sequence and proves to be better than computing error from one random sample.

Additionally, the paper proposes four key metrics to provide comprehensive insight into different aspects of the generated content. The *SyncNet Score* introduces a speech-mesh synchronization network, assessing the alignment between a 3D face mesh sequence and an audio signal. Utilizing two distinct networks, it employs a multi-modal fusion network and computes scores based on merged embeddings, enhancing the evaluation of both temporal and semantic alignment. *SyncNet Frechet Distance (SyncNet-FD)* extends this evaluation by measuring the realism and diversity of speech-related facial motions, calculating the Frechet distance [13] between embeddings of real and generated mesh sequences. Moving beyond synchronization, *Style Cosine Similarity and Rank* assess the replication of diverse speaking styles within datasets. Using the style recognition model called ArcFace [10], these metrics quantify the similarity between the speaking styles of reference and generated sequences, considering both cosine similarity

and rank relative to other speakers. Lastly, *Style Frechet Distance (Style-FD)* evaluates the diversity and distribution of speaking styles by computing the Frechet Distance between recognition model embeddings of real and generated mesh sequences. Together, these metrics provide a comprehensive framework for evaluating the fidelity, synchronization, and diversity of audio-driven 3D facial animations.

Similarly to the previously focused paper, there has been another recent paper from Thambiraja et al. [40] that implements a diffusion-based model and quantitatively evaluates it with multiple metrics. Firstly, the two metrics referred to as L_2^{lip} and L_2^{face} correspond to the previously discussed Lip Vertex Error and Mean Estimate Error accordingly. They also use a Diversity metric that is similar to the one used in [38] but has minor differences and is derived from a paper from Ren et al. [35]. We will use the abbreviation Div^E to recognize it differently from the previously showcased metric. This metric can be formulated as follows: With a set of generated 3D motions with T text inputs. For t-th motion, they randomly sample two subsets with the same size S, and we use m to represent the generated motion.

$$Div^E = \frac{1}{T \times S} \sum_{t=1}^{T} \sum_{i=1}^{S} ||m_{t,i} - \acute{m}_{t,i}||_2$$

Additionally, the paper also uses the Lip-Sync metric that is also used in [41] and is using Dynamic Time Warping to compute the temporal similarity and is influenced by [36] which refers to it as Lip-DTW.

Lastly, there have been other metrics independently used in some of the papers that could be helpful to be examined as well. An example of that is the use of *Emotional Vertex Error (EVE)*, as proposed in citepeng2023emotalk. EVE uses the vertex indexes in the eye and forehead regions and measures the maximum l2 error of the vertex coordinate displacement in the interested region. It can be described as a similar metric to MVE from [38]. Still, it only focuses on the specific regions of the eye and forehead as it is interesting to see the effect of emotions the facial motions on those regions.

Ablation Studies: An additional way to quantitatively evaluate the models and their components is the use of Ablation studies. In this process the aim is to isolate and understand the contribution of specific elements by selectively removing or altering them, helping the crucial factors influencing the overall performance of the models. Ablation studies can be instrumental in refining and optimizing 3D facial animation models, guiding improvements based on a nuanced understanding of the individual elements that constitute their functionality. For most Ablation studies there is use of the same objective metrics as the ones referred above with the difference being that they test different versions of the same model. An example of that is the use of Lip Vertex Error and L2 Error in CodeTalker [55] to evaluate the representation space with a shape-entangled codebook in their way. The result of that process is presented in Figure 1.

Variants	VOCA-Test	BIWI-Test-A	
	Rec. Error ($\times 10^{-5}$ mm)	Rec. Error ($\times 10^{-5}$ mm)	Lip Vertex Error ($\times 10^{-4}$ mm)
Shape-ent. codebook	2.75	4.07	6.41
Motion codebook (Ours)	**0.08**	**2.83**	**4.79**

Figure 1: Ablation study on the representation space of codebook for the CodeTalker [55] paper

Ablation experiments can be done on more influential components of the methodologies as

well and an example of that is conducted in FaceDiffuser [38]. In that paper there are ablation studies that put to the test both the diffusion process as well as the audio encoder, by removing the diffusion completely for the first case and testing with both Wav2Vec2 [1] and HuBERT [18] for the second case. A complete overview of those results is showcased in Figure 2

Ablation on Diffusion Process				
Model	MVE ↓ $x10^{-3}$mm	LVE ↓ $x10^{-4}$mm	FDD ↓ $x10^{-5}$mm	Training Time (m)
w/o Diffusion	6.8833	4.5870	4.6690	≈ 67
FaceDiffuser	6.8088	4.2985	3.9100	≈ 67
Ablation on Audio Encoder				
Model	MVE ↓ $x10^{-3}$mm	LVE ↓ $x10^{-4}$mm	FDD ↓ $x10^{-5}$mm	Training Time (m)
Wav2Vec2	7.4593	5.1590	4.1950	≈ 67
HuBERT	**6.8088**	**4.2985**	**3.9100**	≈ 67

Figure 2: Ablation study on the diffusion process and the different audio encoders for FaceDiffuser [38]

3.2.2 Qualitative Evaluation, subjective metrics and Perception Study

Sometimes using numbers and data is not enough especially when we are talking about a research topic which mainly focuses on human perception and connection. Qualitative evaluation and subjective metrics help us understand how realistic and emotionally authentic the facial expressions appear. Subjective metrics include user surveys and expert evaluations, capturing people's feelings and preferences. On the other hand, objective metrics, like facial landmark accuracy and lip synchronization, provide more quantitative measures. Using both types of metrics ensures a well-rounded evaluation, addressing the subjective nature of facial expression interpretation. This approach helps researchers and developers improve audio-driven 3D facial animations by understanding their strengths and limitations more comprehensively. Qualitative evaluation involves the in-depth analysis of the visual aspects, such as the naturalness and believability of facial expressions. Subjective metrics, on the other hand, delve into the subjective experiences and preferences of users or experts, capturing nuanced emotional responses. Perception studies, which often employ techniques like eye tracking, assess how viewers process and interpret facial animations. While qualitative evaluation focuses on visual quality, subjective metrics tap into human preferences, and perception studies delve into cognitive processes, these approaches can be synergistically combined. By merging qualitative insights with subjective metrics and perception study outcomes, researchers gain a more comprehensive understanding of the holistic user experience. This integrated approach not only refines the visual realism of animations but also ensures that user perceptions align with the intended emotional impact, leading to more effective and emotionally resonant audio-driven 3D facial animations.

In this section, we will take a broader look into what is usually used to conduct a qualitative evaluation by analysing state-of-the-art papers. The first paper we will start with is FaceDiffuser [38], which is subjected to extensive analysis, comparing generated animation sequences to both ground truth and existing methods. Notably, the approach excels in accurately reproducing lip shapes resembling ground truth motions, while demonstrating diverse upper-face movements, even in challenging scenarios like multiple speakers, noisy audio, and different languages. The diversity metric, introduced, is demonstrated by sampling animation sequences with the same

audio but conditioned on different training subjects. The results are visually represented through mean and standard deviation motion plots with heatmap visualizations, which are showcased in Figure 3 below. Additionally, animation graphs showcase the model's diversity by sampling multiple times with the same audio input, revealing variations, especially in eye blinks.

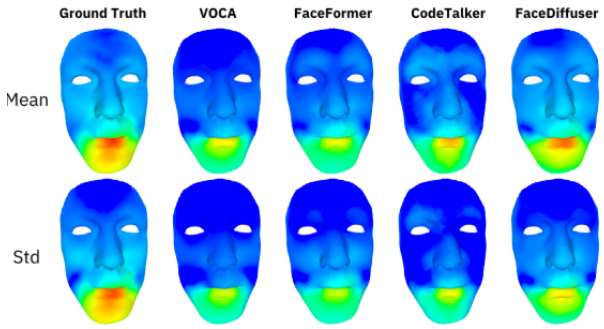


Figure 3: Animation sequences sampled from BIWI-Test-B conditioned on different training subjects, all showcased for GT and 4 models

Additionally, another paper that uses a similar qualitative evaluation was the one from Thambiraja et al. [40] for the Diffusion model. In that paper the reproduction of multiple methods is done using VOCASET [7] and they identify expressive facial animations that match the speaking style of the target subject. As a broad example with multiple methods, it is useful to have in mind the diverse output from some of the methods due to their non-deterministic approach and the opposite of deterministic. That is the case with the Imitator [41] model on that comparison as well, which achieves synthesizing convincing animations, its outputs are not diverse.

The Perception Study expands the evaluation to user studies, employing an A/B testing strategy. Participants judge realism, lip-sync, and appropriateness by comparing randomly paired videos rendered by different methods. The study encompasses three datasets: BIWI [12], VOCASET [7], and UUDaMM, with 83 survey responses. The results are provided together in the form of a percentage of approval in the different metrics that the users were asked about.

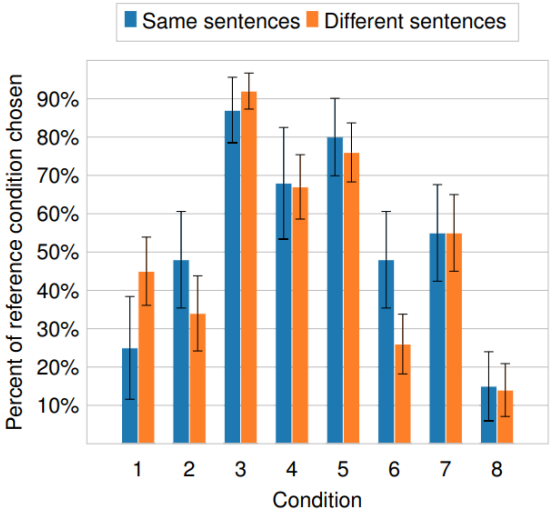
On a similar note for that part of the evaluation is the work on CodeTalker [55] which firstly is visually compared with other competitors, namely VOCA [7], FaceFormer [11], and MeshTalk [36]. To ensure a fair comparison, the same talking style is assigned to all methods as conditional input, randomly sampled for each. Lip synchronization performance is examined through the illustration of three typical frames in the synthesized facial animations, depicting specific syllables. Notably, CodeTalker exhibits more accurately articulated lip movements in alignment with speech signals, outperforming competitors in terms of both accuracy and consistency. The evaluation extends to facial expressions, where CodeTalker showcases stronger facial movements and a broader range of dynamics, as evidenced by the calculated temporal statistics of adjacent-frame facial motions. The discrete facial motion space contributes to robustness against cross-modal uncertainty, promoting superior performance. Readers are encouraged to refer to the Supplemental Video for further animation comparisons. In the User Study, a comprehensive assessment is conducted to evaluate the quality of animated faces in perceptual lip synchronization and realism. A series of A/B tests are employed for each comparison, including CodeTalker versus competitors (VOCA, MeshTalk, FaceFormer), as well as the ground truth. The study involves 31 participants, each participating in all eight kinds of comparisons to ensure broad exposure and cover the diversity of favorability. The results of the User Study are presented as an example in

Figure 4 below.

Competitors	BIWI-Test-B		VOCA-Test	
	Lip Sync	Realism	Lip Sync	Realism
Ours *vs.* VOCA	92.47	89.25	86.02	84.95
Ours *vs.* MeshTalk	80.65	82.80	95.70	92.47
Ours *vs.* FaceFormer	53.76	56.99	70.97	69.89
Ours *vs.* GT	43.01	49.46	43.01	43.01

Figure 4: User Study - A/B testing on CodeTalker: The percentage of answers where A is preferred over B

Lastly, we will look into how the setting stone of this research that is usually used to be compared with other methods is evaluated in the paper, VOCA [7]. Firstly the paper showcases the perceptual evaluation as a series of blind user studies on Amazon Mechanical Turk (AMT) are conducted to assess various aspects of VOCA's performance. The first study involves a binary comparison between held-out test sequences and VOCA conditioned on all training subjects, aiming to evaluate the naturalness of facial movements. Another study investigates the correlation between style, content, and identity. All experiments are carried out on sequences and subjects completely disjoint from the training and validation sets. Binary comparisons require participants to choose the talking head that moves more naturally and follows the audio from a pair of videos with the same animated subject and audio clip. Style comparisons evaluate learned speaking styles, asking Turkers to determine which of two predictions is more similar to a reference video. The investigation into speech-driven facial motion independently from identity-dependent face shape reveals varied results across conditions, emphasizing the challenges of disentangling style, identity, and content in perception. The study of styles is showcased in Figure 5a for reference as it is an interesting way of analyzing the results. The bars show the percentage of choosing the reference condition when the same sentence was being shown for reference and prediction, and with different sentences.



(a) Study of styles: Percentage choosing the reference condition when the same sentence was being shown for reference and prediction

(b) Generalization across face shapes with the template on the left and three randomly selected animation frames

Figure 5: a and b: Study of Styles and Generalization

In the qualitative evaluation, VOCA's generalization capabilities are showcased across subjects, languages, and speaker styles. The model demonstrates the ability to animate a wide range of adult faces with large shape variations. Generalization to non-English sentences is illustrated in Figure 5b below, and different speaking styles are achieved by conditioning on different subjects during inference. The paper highlights the linearity of the decoder, allowing for the generation of new intermediate speaking styles through convex combinations. VOCA proves robust to noise, maintaining plausible facial animations even in the presence of high noise levels. Comparison to [20] demonstrates VOCA's ability to produce similar facial animation without using subject-specific training data. Animation control is showcased by changing identity-dependent shapes and head poses during animation while maintaining realistic facial motion.

Finally, we will make a more clear overview of the subjective metrics that are used in qualitative evaluation on the topic of audio-driven 3D facial animations.

- **Visual comparison:** Showcasing generated sequences of multiple models and comparing them with GT.
- **Diversity Metric:** Animation sequences with the same audio but conditioned on different training subjects.
- **Realism:** Metric used for A/B testing for Perception Studies to represent the naturalness of the animation.
- **Lip-Sync:** Metric used for A/B testing for Perception Studies to represent the lip synchronization accuracy of the animation
- **Styles:** Showcasing the preference towards the Stylistic differences which include variation in lip articulation.
- **Generalization:** Showcase the generalization capabilities across a wide range of adult faces.

3.2.3 Metrics used for Non-deterministic and Deterministic models

After the thorough examination of most of the metrics used to evaluate state-of-the-art models both qualitatively and quantitatively, it is useful to identify the difference between the ones used in Deterministic and Non-deterministic models. As the topic of the difference between the two categories will be a major part of our research the examination of how the evaluation is done differently, or at least needs to be done differently, is of high importance.

Deterministic Models: Deterministic models adhere to a fixed set of parameters and inputs, resulting in consistent and reproducible outputs for a given set of conditions. In the context of 3D facial animation synthesis, deterministic models generate facial animations with minimal variability across multiple executions, given the same input data. For deterministic models, the evaluation predominantly centres around metrics that assess accuracy, precision, and adherence to ground truth. Metrics such as Lip Vertex Error (LVE), Mean Vertex Error (MVE), and Facial Dynamics Deviation (FDD) provide valuable insights into the fidelity of the generated sequences compared to the expected outcomes. These metrics are well-suited for deterministic models where consistency in output is a key criterion.

Non-deterministic Models: In contrast, non-deterministic models introduce an element of variability in their outputs, even when presented with identical inputs. This variability could

stem from the use of probabilistic or stochastic processes within the model architecture, resulting in diverse facial animations for the same input conditions. Evaluation for non-deterministic models requires a more nuanced approach, considering not only the accuracy of individual samples but also the diversity and realism of the generated sequences. Metrics calculating the Diversity, both the one referred to as Diversity and DivE, and the proposed metrics from Yang et al. [56], including Coverage Error and Mean Estimate Error, play a crucial role in assessing the dispersion and quality of outputs from non-deterministic models. Also, the Diversity metric as proposed in [38] but more accurately implemented in [40] is one of the most useful ones to evaluate the probabilistic nature of those models.

Adapting Evaluation for Both: Given that our research aims to bridge the understanding between deterministic and non-deterministic methods, the evaluation framework must be flexible. It should encompass metrics capable of gauging precision and consistency for deterministic models, while also capturing the variability and diversity intrinsic to non-deterministic counterparts. Some of the already existing metrics can be proved useful for both categories, especially those that are mainly used for accuracy. Most of the metrics, especially the ones referred to as metrics used for deterministic models are equally useful for non-deterministic models, as they are mostly not directly testing the determinism of models but the general accuracy of the representation. Most of the models that use a non-deterministic method are also testing it with metrics such as LVE, MVE, FDD and more, to also evaluate the accuracy of certain parameters of the results, such examples are the models of [38] and [40]. The main usefulness of those metrics for non-deterministic models can be identified in the in-depth work of Yang et al. [56]. In that paper, there is a clear implementation of two new metrics that are adaptations of the Lip Vertex Error to use different samples of results to test the possible similarity of some of them towards the Ground truth. Using the basic calculation used for LVE the new metrics, Coverage Error and Mean Estimate Error, are testing at the same time the variability of the results as well as the similarity of the samples with the ground truth sequence.

3.3 Datasets

In this section we will explain the details of the dataset that will be used for both the training of the models as well as the dataset analysis, focusing on the motivation behind those datasets. Additionally, we will showcase some preliminary information for the task of capturing our dataset.

Multiface [54]: This dataset stands out as a valuable resource, encompassing high-quality recordings of 13 individuals captured across two versions (v1 and v2), each offering a diverse set of features. Notably, the dataset presents an average of 12,200 to 23,000 frames per subject, recorded at 30 frames per second, with variations in the number of camera views and illumination between versions. Each frame captures a wealth of information, including images from multiple camera views at a resolution of 2048 × 1334 pixels, tracked meshes with head poses, unwrapped textures at 1024 × 1024 pixels, metadata with intrinsic and extrinsic camera calibrations, and audio. The Mugsy capture studio, characterized by a dome structure with cameras arranged on a sphere's surface, facilitates synchronized multi-view videos. The capture script intentionally spans a range of facial expressions, gaze directions, and phonetically balanced sentences. The dataset's richness in high-resolution data makes it a compelling choice for training models in our research. When evaluating existing datasets, Multiface emerges as particularly useful due to its comprehensive nature, providing a diverse and detailed representation essen-

tial for addressing the research questions.

BIWI [12]: This dataset represents a pivotal contribution to our research, offering a nuanced exploration of both neutral and emotional facial expressions. Comprising synchronized audio4D scan pairs from 14 human subjects uttering 40 phonetically balanced English sentences in two distinct scenarios—neutral and emotional—the dataset presents a comprehensive collection of 1120 sequences. Each sequence averages 4.39 seconds in duration, encompassing both the initial neutral utterances and subsequent emotionally expressive renditions. The deliberate inclusion of emotional expressions adds a unique dimension to the dataset, crucial for training models capable of capturing a broad spectrum of facial dynamics. The dataset's value is further underscored by the suggested split detailed in the FaceDiffuser paper [38]. This split results in BIWI-Test-A, comprising 24 sentences from familiar subjects, and BIWI-Test-B, containing 32 sentences from the 8 remaining unseen subjects. This stratification allows for an insightful evaluation of model generalization, essential in understanding how well the trained models perform on familiar versus unseen subjects. In the broader context of dataset selection, BIWI distinguishes itself by providing not only essential emotion labels but also identity labels, a unique combination that significantly enriches the dataset's utility for our research. The BIWI dataset, with its meticulous design and inclusion of diverse expressions, emerges as an indispensable resource for training and evaluating models in the pursuit of addressing the research questions at hand.

3DMEAD [48]: The 3DMEAD dataset stands as a significant resource in our research, particularly in its unique emphasis on emotional expressions. This 3D reconstruction, an extension of the Mead dataset, captures the dynamic interplay of emotions across 60 actors and actresses engaging in conversations. Notably, the dataset includes dialogues featuring 8 distinct emotions at 3 different intensity levels, providing a nuanced representation of emotional dynamics in facial expressions. The inclusion of emotions at varying intensity levels adds a layer of complexity crucial for training models capable of discerning subtle nuances in facial reactions. Each audio-visual clip within the dataset is captured from 7 different view angles, resulting in a comprehensive set of visual information. With a substantial collection of 40 hours of audio-visual clips for each person and view, 3DMEAD ensures a rich and diverse dataset for training and evaluation. The deliberate focus on emotions in this dataset aligns with the core objectives of our research, emphasizing the importance of understanding and accurately representing emotional states in 3D facial animations. Considering the broader landscape of available datasets, 3DMEAD distinguishes itself by providing an extensive collection of emotional expressions, offering a unique resource for training models that can effectively capture the nuances of human emotion. The inclusion of various emotions at different intensity levels enriches the dataset's utility, positioning it as a valuable asset in our pursuit of addressing the research questions and advancing the understanding of 3D facial animation within the realm of emotions.

VOCASET [7]: The dataset captured to be used on the VOCA model proved to be a useful benchmarking dataset for many models to come. This dataset captured from 6 female and 6 male subjects, comprises 40 sequences of sentences spoken in English by each individual, lasting three to five seconds. The selection of sentences draws from diverse linguistic sources which ensures a broad phonetic diversity crucial for training and evaluating models effectively. The dataset's organizational strategy ensures a balance between shared sentences across subjects, sentences spoken by a subset of individuals, and those uniquely spoken by one or two subjects, resulting in a total of 250 unique sentences. The capture setup employs a sophisticated multi-camera active stereo system, capturing high-quality 3D head scans and synchronized audio at 60fps and 22 kHz,

respectively. This process results in unposed meshes that faithfully represent the raw data, and additional post-processing focuses on fixing neck boundaries and ears while preserving subtle facial motions.

BEAT [23]: The BEAT dataset represents a substantial advancement in the domain of multimodal data for conversational gesture synthesis. This dataset contains a comprehensive spectrum of semantic and emotional annotations, derived from 30 speakers over 76 hours of motion capture data. It includes 2508 sequences, meticulously segmented into conversational and self-talk sessions. The conversation sessions span 10 minutes each, involving interactions on 20 predefined topics, while the self-talk sessions consist of 1-minute segments designed to evoke eight distinct emotions. With more than 3 million frames annotated for emotion, BEAT facilitates a comprehensive analysis of the interplay between facial expressions and emotions. The dataset's diversity is further highlighted by its inclusion of data in four languages—English, Chinese, Spanish and Japanese.

4 Nature of Facial Animations

After understanding the previous knowledge of the metrics, the models and the datasets we are going to separate the processes that will be followed to conduct our study. Due to the nature of the study and the variety of topics that will be covered, we aim to separate them into different sections for better understanding. In this section, we will explain the analysis and the processes that were conducted to conclude on the question about the nature of Facial Animations. That process was mainly done by analysing the ground truth of the existing datasets in various ways, also identifying the differences between the latest and most broadly used datasets for 3D facial animations, as well as other areas of research. To answer the question about the nature of facial animations we have to divide the specific part of non-determinism that will be explored. This will be done by first evaluating the difference each subject makes across the same sequence as well as specific parameters such as emotions or intensity. The two latter parameters due to the nature of the datasets that will be explored will mainly be identified for the 3DMEAD dataset [48] which specifies those parameters for each motion.

The main variable that we aim to explore is the effect of a specific subject/actor on the way a motion is uttered, which will help identify the reality behind determinism and non-determinism in the ground truth. At this point, we have to identify a limitation that would make the process more straightforward in different aspects. As previously explained, all the current datasets contain subjects uttering multiple different sequences at once. One way that would definitely be helpful for this specific research question is to have the subjects perform the same sequence multiple times. That would undoubtedly be a more complex dataset to explore, while at the same time providing an easier way to answer the question about the nature of facial animations. Since this type of dataset is not yet available, we aim to use the existing state-of-the-art datasets to analyse in-depth and provide an answer to this question.

3DMEAD: Starting with the explanation of the process we will focus on the variables that were tested for the 3DMEAD dataset as it is the one with the broader points to evaluate. Firstly, to have a clearer view of how emotions affect the way a sequence is uttered across all subjects, we will use the FLAME parameters of the dataset. While using the mean jaw movement of all the subjects across different emotions for a specific sequence it becomes clear that some patterns are followed but are not the rule every time as seen in Figure 6.

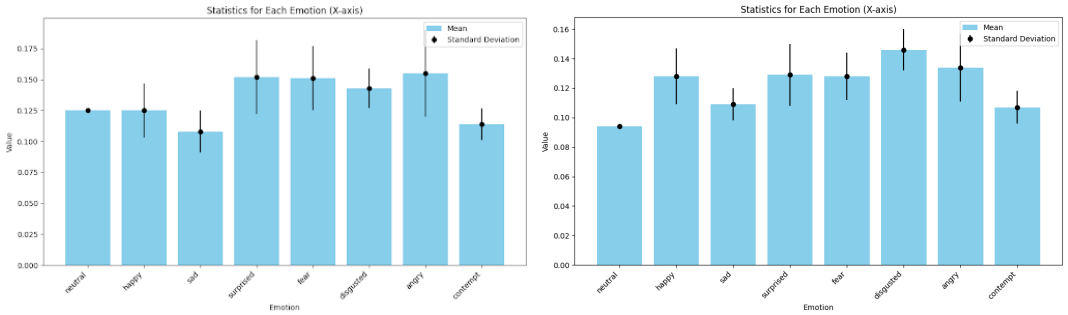


Figure 6: Mean Jaw Movement of each emotion across all subjects for sequence 1 and 17

The mean of all the subjects across each of the three axes (X, Y and Z) was calculated with focus on the emotions for each motion. The standard deviation showcases the difference across the intensity level that each emotion was performed in. On a similar note, the same process was done but this time the test was focused on the subjects. Meaning that we explored the difference the mean jaw movement has when the same sequence is uttered by different subjects. The test included all emotion and intensity levels, and from the outcome in Figure 7 it becomes clear that there are some major differences across some specific subjects.

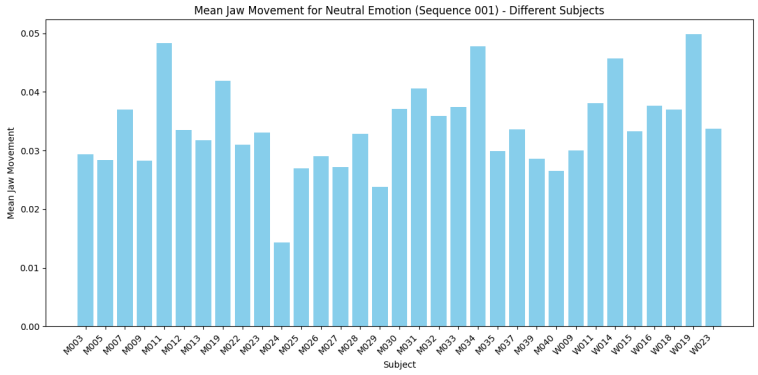


Figure 7: Mean Jaw Movement of each subject uttering sequence 1

After this first step which provided an understanding of how the datasets work, the focus was concentrated on other kinds of analysis tools which would be more applicable to what we are trying to test. Such tools are the T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [44] which is a dimensionality reduction technique that aims to visualize the structure of huge datasets by reducing the hundreds of different parameters into 2D space for easier understanding. That technique is one of the main ones used to explore the datasets more in-depth to capture all the parameters that matter for our study. We use the standard t-SNE implementation from the open source sklearn Python library [29] with perplexity set to 15 and learning rate to 200, with the mean used as the merging strategy for all experiments.

After identifying the process that needed to be done and the holistic view that a dimensionality reduction tool would provide to us, we aimed to create a subject-based view to evaluate the effect each subject has on the sequences of the dataset. To explore the effect that the subjects have when uttering a sentence we have to dig into the whole dataset to explore the possibility that each subject has its own style that is represented after the use of dimensionality reduction. For that purpose we conducted two different experiments, the first one being the complete iso-

lation of the data of a single sentence and the second one containing the whole dataset. For the first approach, we conducted t-SNE using only the sequences of sentence '001' for all subjects, emotions and intensities, which would give a zoomed-in approach to the differences between the subjects uttering the sentence "She had your dark suit in greasy wash water all year". For the second approach, we will use the whole dataset using the same subject-based colour mapping, with more colours to understand the different subjects clearly. The main purpose of using all the sequences is to evaluate if the same route is followed across all sentences. The outcome of these approaches can be identified in Figure 8.

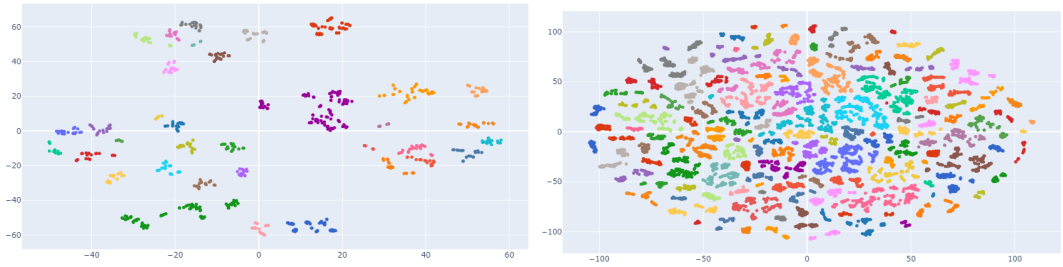


Figure 8: t-SNE conducted across sequence '001' on the left and the whole dataset on the right with subject-based colour mapping

Following the same idea, using the whole dataset and understanding the usefulness of dimensionality reduction techniques, we conducted Principal Component Analysis [42][25] on the dataset focusing on the subjects. For every experiment using PCA, we will use the standard implementation of sklearn Python library [29]. The main purpose of using a second dimensionality reduction technique was to create a more holistic and clearer view of the result while being able to interpret it by two outcomes. Additionally, due to the multiple tests that were conducted with multiple different parameters taking place, we ought to focus on the two main aspects the subjects and the emotions, with two different techniques. With that in mind, we conducted the same experiment using all the sequences in the dataset with subject-based colour mapping to identify the effect of the individual subjects on the sequences, which can be seen in Figure 9

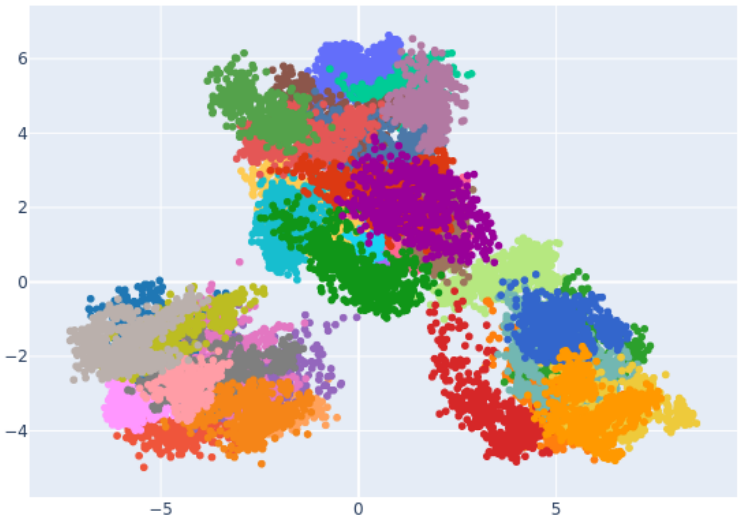


Figure 9: PCA on all subjects in the 3DMEAD dataset with colouring across the subjects

As we explore the non-determinism and the effect of the subject towards the sequences, another factor that needs to be explored is the effect of emotions towards the sequences and how much they affect the way the sentences are uttered. Even if we can understand the possible existence of an effect following the mean jaw movement calculation, we need to understand if this can also be identified when we use these kinds of techniques. First off, we explored using the algorithm to conduct t-SNE for a specific subject to identify the differences across the emotions. Then, similarly to before we explored the effect of emotions towards the whole dataset, by conducting t-SNE with all the sequences performed by all subjects with emotion-based colour mapping. Both outcomes can be seen in Figure 10.

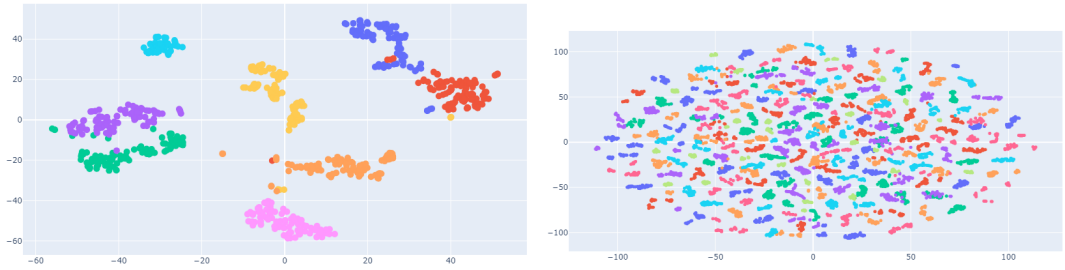


Figure 10: t-SNE on subject 'M003' on the top and the whole dataset on the bottom, with colouring across the emotions

After experimenting with t-SNE with a focus on the emotions we found it logical to also test the hypothesis and the outcome using PCA. For that reason, we conducted a similar experiment with the whole dataset using this dimensionality reduction technique. The result of this process is presented in Figure 11.

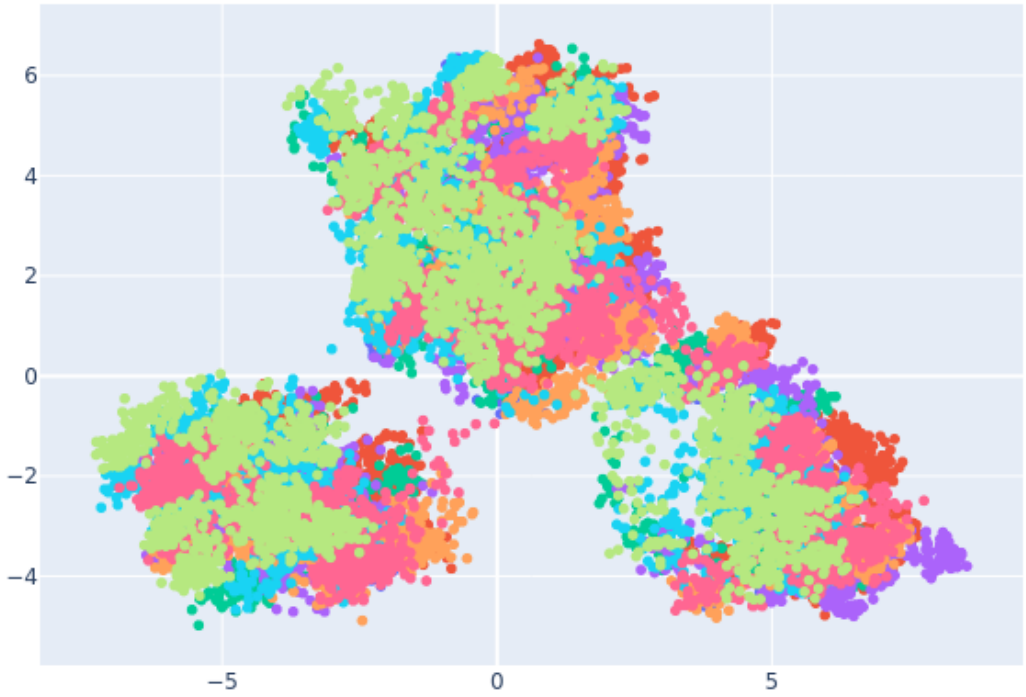


Figure 11: PCA conducted on the whole dataset with colouring across the emotions

BIWI and Multiface: After experimenting and exploring multiple aspects of non-determinism for the 3DMEAD dataset we aim to adapt similar methods and tests for other datasets as well. One important aspect that should be made clear is that due to the lack of emotional categorization of the sequences in those two datasets, we are not able to test the difference in emotion. BIWI contains two categories, emotional and neutral, but that does not enable us to make tests similar to 3DMEAD for emotions. For the reasons explained above and the need to identify the possible non-determinism in terms of subjects we will follow a similar path in the experiments using dimensionality reduction techniques in the same way. First off, for the testing of any possible existence of a strong correlation between the sequences with the same subject uttering them we conducted t-SNE with the same parameters as before and the same colour mapping based on the subjects. In Figure 12 we can see the results for both datasets for a clearer comparison.

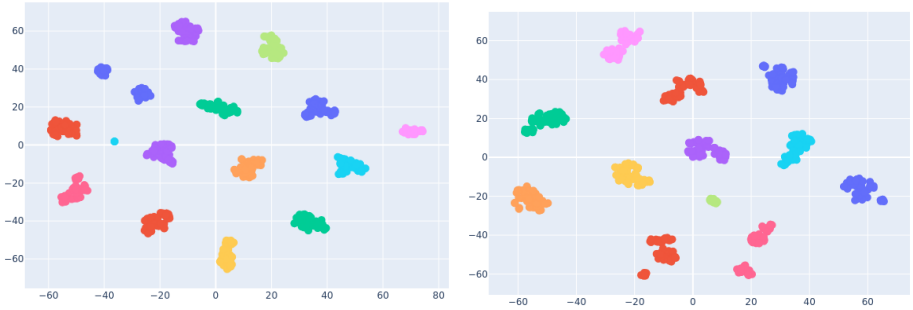


Figure 12: t-SNE conducted on the BIWI dataset on the left and the Multiface dataset on the right, with colouring across the subjects

For the next step, we followed the same route using the PCA dimensionality reduction technique as a cross-reference for the t-SNE technique. With the same parameters, we conducted the same experiment to evaluate the possible clustering that might also appear when using PCA for the two datasets. The outcome of those tests can be seen in Figure 13

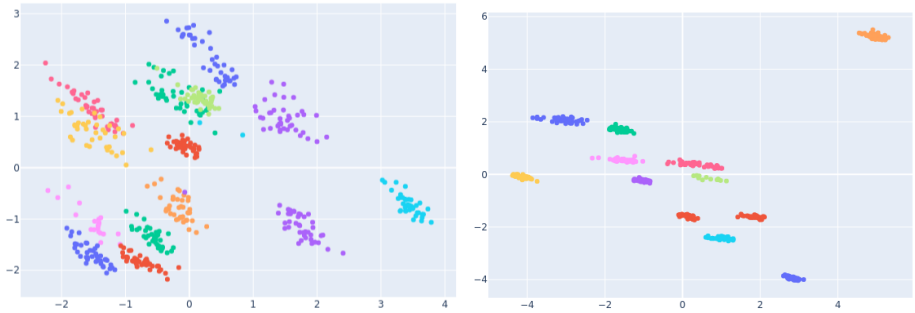


Figure 13: PCA conducted on the BIWI dataset on the left and the Multiface dataset on the right, with colouring across the subjects

BEAT: The final dataset that was explored based on the non-deterministic nature of its parameters is the BEAT dataset. This dataset contains emotions categories and the subjects perform the sequences based on those emotions, which means that we can also explore this variable of the dataset. At first, we will experiment with the t-SNE dimensionality reduction technique with both subject-based and emotion-based colour mapping. A similar procedure, as the one for 3DMEAD with the same parameters, was done and the output is presented in Figure 14.

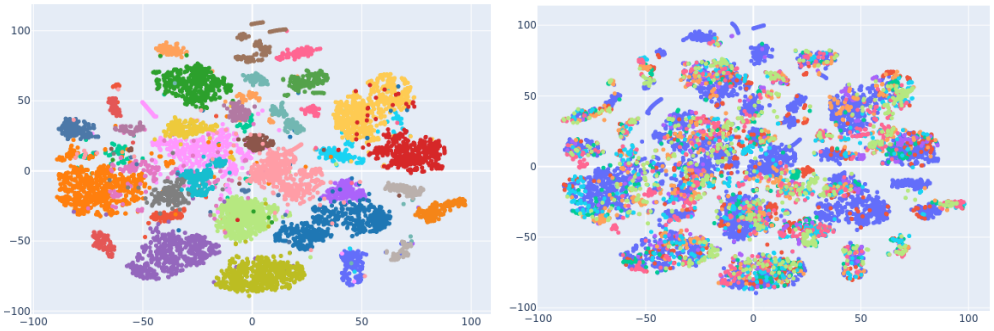


Figure 14: t-SNE conducted on the BEAT dataset: On the left with subject-based colour mapping and on the right, with emotion-based colour mapping

Furthermore, to explore the possible similarities and differences that using a second dimensionality reduction technique could provide, we also experimented with using PCA on this dataset, similar to the previous datasets. The experiment with the PCA used the whole dataset and focused on the possible correlations between the sequences uttered by the same subject. The outcome of the PCA experiment can be seen in Figure 15

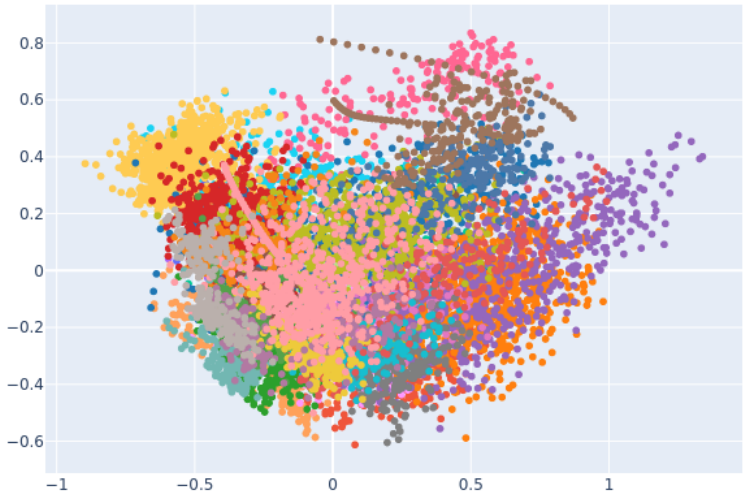


Figure 15: PCA conducted on the BEAT dataset with subject-based colour mapping

Datasets Comparison: After, the completion of the exploration conducted mainly using dimensionality reduction techniques for the datasets we aimed to continue our focus on the core parameters that exist in these datasets. To identify possible pros and cons that the existing facial animation datasets have, the need to broaden the spectrum and compare them with datasets used for other research fields occurred. That would come as an additional research point to identify if there is, in general, a huge room for improvement in the existing and state-of-the-art datasets for 3D facial animations with one main focus being the diversity they provide. We aimed to analyse the differences both internally, in 3D facial animation datasets, and externally, with other datasets, to understand where and how some datasets fall behind. That straightforward process would also be insightful for the previous work that focused on aspects such as the subjects and emotions contained in the existing datasets.

Moreover, we gathered all the needed information for the datasets that will be examined, with categories such as:

- Subjects
- Emotions
- Sentences uttered by each subject
- Total number of sequences
- Intensities of the performed sequence
- Total number of hours that the dataset contains
- Total number of frames that the dataset contains

The gathered information for the datasets examined in our study can be seen in Table 1.

Dataset	Subjects	Sequences	Total Sequences	Emotions	Intensities	Hours	Frames
BIWI	14 (8M + 6F)	40	536	1	1	0.71	64 K
Multiface	13	50	562	1	1	0.67	65 K
BEAT	30 (15M + 15F)	118	11399	8	1	31.66	3.4 M
3DMEAD	35 (27M + 8F)	40	21844	8	3	26.04	2.3 M

Table 1: Core Information of 3D facial Animation datasets

Moving on, to produce a simple comparison with datasets in other research fields, we first selected the most obvious choice which is 2D datasets mainly used for 2D talking heads models. There can be a direct comparison between those two classifications as the information is similar in terms of categories which can only be proved useful. For that process, we selected to use the GRID dataset [5], the CREMA-D dataset [2] and the TCD dataset [17], with the same information categorisation. The information of these datasets can be identified in Table 2

Dataset	Subjects	Sequences	Total Sequences	Emotions	Intensities	Hours	Frames
GRID	34 (18M + 16F)	1000	34000	-	-	40.51	3.7 M
CREMA-D	91 (48M + 43F)	12	7442	6	4	11.08	1.2 M
TCD	62 (59 Tr + 3 Te)	100	6913	-	-	11.1	1.2 M

Table 2: Core Information of 2D face datasets

Furthermore, another insightful research field whose datasets we explored is the body motion topic. This topic could be considered probably the most advanced as it has been a major research field for many years. Having those ideas in mind we aimed to use the datasets that exist for this topic intending to understand the possible showcasing of that advancement as well as identify the specific key points that 3D facial animation datasets need to evolve. Some noteworthy differences must be considered before identifying any direct correlations between the datasets, with the main one being the difference in aim. 3D facial datasets have evolved in a way that most of the advancements focus on the emotional or even intense performances of the subjects, which is not the case for body motion datasets. Additionally, the whole pipeline that is followed to create or use a body motion dataset, especially the actions, has nothing to do with the common 3D facial animation datasets. For this part of our research, we used the datasets KIT-ML [32], BABEL [34] and IDEA 400 [22]. Having those in mind we have categorised the information from the datasets and we present the findings in Table 3.

Dataset	Subjects	Total Sequences	Hours	Actions	Frames
KIT-ML	111	3911	11.2	-	1.2 M
BABEL	300	13220	43.5	250	4.6 M
IDEA 400	36	12500	24.4	400	2.6 M

Table 3: Core Information of Body-Motion Datasets

5 Evaluation metrics inventory

In this section, we aim to present the process that was followed to create and evaluate the evaluation metrics inventory we needed to create. The inventory's purpose was to have a complete and holistic view of all the metrics across multiple different models and datasets. By creating that inventory ourselves we can identify the differences that exist in the metrics with all of them tested in the same way, with the same system and same dataset splits. That process would also provide us with useful insight in terms of the variety of metrics and their adaptiveness to the models, especially in terms of determinism and non-determinism. This process requires a precise selection of metrics and models, as well as the datasets that would be used to create that inventory with one of the most important parameters being the variety in all aspects so that most fields are covered.

5.1 Experiment Selections

Models selection: Even if our main goal for creating the inventory is not the comparison between the models themselves, making an appropriate selection of the models that will be displayed in our inventory is of high importance. The reasoning behind that is the attempt to have models of a quite different structure in multiple aspects, as well as, models that have claimed in their recent research to be outperforming others based on the evaluation metrics. One of the main parameters that were taken into account was the use of both deterministic and non-deterministic models in our inventory. In that way, we aim to have a holistic outcome that will provide clear results in terms of how the metrics evaluate differently between the models that are deterministic and non-deterministic. Following that decision we will use FaceDiffuser[38] as our main non-deterministic model and FaceXHuBERT [16] as our deterministic model, both being state-of-the-art in their respective areas. Additionally, to test the metrics using a model with a motion prior as the main aspect of the architecture we will use CodeTalker [55] for our experiments. The experiments on CodeTalker will be done in two ways, firstly using the released model that produces deterministic results and afterwards making minor changes to the quantization to achieve non-deterministic results, providing us with CodeTalker-ND. Following that, we will use FaceFormer [11] with its transformer-based architecture and the two modules in its decoder will provide insightful results. The last model we will use in the experiments will be a VQ-VAE model proposed by a fellow student in our research team, ProbTalk3D [52], which previously proved to be outperforming other state-of-the-art models. ProbTalk3D will only be tested on 3DMEAD for both subjective and objective metrics due to the model being trained on the FLAME parameters of the 3DMEAD dataset and not the vertices, as all the other models will be trained on. We consider our selection of models to be the one that covers the most variety possible and in most aspects possible, taking into account the limitations. One of those limitations is the availability of the models for training as some of the most recent ones have not published their code.

Dataset selection: Similarly to the selection of the models, we have to precisely select the datasets that we are going to use for our experiment with a focus on multiple aspects. One of the aspects that creates a path towards the selection is how many of the models have already been tested on those datasets and proven to perform. Likewise, one useful parameter would be that the original implementation of the models was done on some of the datasets to have a backbone to rely on how they are expected to perform. With those parameters in mind, one of the most obvious selections of our experiments was using BIWI [12]. That selection is derived from the fact that all of our models and other proposed models use BIWI as the main dataset to experiment and evaluate their results. Similarly, another dataset that has recently been used in two of our models is Multiface [54] which we will also include in our experiments. Those two datasets have some advantages and disadvantages that were also taken into account. Except for the obvious advantage of our experiment that has already been discussed, following the research on the parameters of each dataset, the consistency of BIWI and Multiface is obvious. This means that both datasets have a specific number of sentences performed by all the subjects with clear numbering and categorization, something that is missing from other datasets such as BEAT [23]. As done from exploring the datasets in terms of determinism and non-determinism of their ground truth, we also needed to include in our experiments a dataset that covered the spectrum of emotions. That dataset was selected to be 3DMEAD [48], which includes both emotion categorization and intensity levels for each sequence. That dataset will help us understand how differently the results can be represented on the metrics with the categorization of emotions and with the 3DMEAD being a larger dataset compared to the others that are going to be used, we will identify possible differences. One problematic factor of the 3DMEAD dataset is the fact that none of the models' architecture was built around emotion categorization or at least previously tested on that dataset. The handling of such a new dataset that was only used for one similar model, EMOTE [9], will be an additional challenge but the results provided by using that dataset will be insightful for our study.

Metrics Selection: The final selection that we are required to make before starting our experiments is to decide which metrics we are going to calculate for the models. That selection comes with some prerequisites required to cover the most aspects to have both metrics that have been already established in the research field as well as additional ones that we find insightful. After the completion of our literature study, it has become clear that two metrics that are used by most if not all prior experiments are Lip Vertex Error (LVE) and Facial Dynamics Deviation (FDD). Those metrics provide coverage of both the lip area and the upper face area in terms of comparison of the predicted sequences with the corresponding ground truth. Likewise, the Mean Vertex Error is an additional useful metric for us as it takes the calculation process of LVE and uses it for the whole face. That metric becomes extremely interesting as we ought to identify the effect 3DMEAD has on the whole face due to this dataset's emotion categorization is considered to be present on the whole face. The three metrics stated above are the metrics that have already been widely tested and we have results to compare directly with, especially for the selected models of our experiment. In addition to those metrics, we decided to add more variation in terms of metrics that are mainly meaningful for non-deterministic models. That examination needed to be wisely done and the selection needed to be precise, in both useful previously used metrics as well as metrics that have been suggested recently that we seem fit for our experiment. The first selection of a metric that comes to mind for non-deterministic evaluation is the Diversity metric from FaceDiffuser[38], but after careful consideration, we came across the fact that this metric is

not suitable for actual testing of the variety of different results for the same motion. The reason being that this specific Diversity metric focuses on calculating the difference between motions with different subject conditioning and not using the same motion with multiple samples. We could argue that this metric is more suitable for testing the effect the subject conditioning has on a motion and not the variety that multiple generated samples of the same motion have between them. For those reasons and also need to use a metric that tests the diversity across generated samples for each motion we decided to use the previously referred DivE metric that was first used by Ren et al. [35] and adapted for 3D facial animations by Thambiraja et al. in 3DiFace [40]. For that metric, which from now on we will refer to as Diversity, we needed to implement our understanding of the calculation due to the lack of publication of 3DiFace's code. The decision to use this metric as our main Diversity metric comes from the fact that it aims to calculate the direct difference between multiple different samples for the exact motion, directly testing the non-determinism of the results. The two final metrics that were selected to be used in our inventory are the ones derived from the paper from Yang et al. [56] and those are Coverage Error and Mean Estimate Error. Coverage Error was decided to be used as a non-deterministic metric to evaluate the possible existence of motions generated samples that achieve a closer representation of the Ground Truth. At the same time, Mean Estimate Error has been introduced to check the mean of the generated samples and its difference with the GT, testing the performance of the non-determinism more universally.

5.2 Experiment Settings

One of the main purposes of training and testing all the models ourselves was to make sure that all experiment settings were the same. The settings that are the same include both machine variables as well as the dataset split used for each dataset. The machine that all the training was conducted is a Dell PowerEdge R7525 server which is equipped with 2 x AMD 7313 processors, offering a clock speed of 3.0GHz, 16 cores, and 32 threads, with a 128MB cache. Memory is abundant with 1024GB of 3200MT/s RDIMM. Additionally, the server is equipped with 8 x 3.84TB SSDs running at 6Gbps in a RAID6 configuration, providing a total of 23TB of storage for data. The operating system installed is AlmaLinux 8 with the configuration including an NVIDIA Ampere A16 GPU with 64GB of memory.

For the data split, we decided to use the sentence split used in most papers for BIWI and Multiface, and follow the same logic with the 3DMEAD dataset. The splits for each dataset are presented below in Table 4.

Each of the models was trained at 100 epochs using the exact same split for each dataset with the models being the same as the ones published by their authors. Minor changes were done for 3DMEAD's purposes in terms of adapting the hyperparameters with the vertice dim being 15069 as the dataset suggests. For all the metrics tested, we used the same implementation as the one presented in Section 3, with LVE, MVE and FDD being directly implemented as they are in the code of FaceDiffuser[38]. For the metrics for which no code is available such as CE and MEE we followed Yang et al. [56] in the description given in the paper in our best understanding. For the Diversity metric, we performed our adaptation of the code from Ren et al. [35] for 3D facial animations since 3DiFace's [40] code was not available at the start of the experiment process.

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 = 15080 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 = 1869 sequences	9 seen subjects 5 sequences per subject Total = 45 sequences
Test Set A	6 seen subjects 4 sequences per subject Total = 24 sequences	-	9 seen subjects 5 sequences per subject Total = 45 sequences
Test Set B	8 unseen subjects 4 sequences per subject 6 conditions per sequence Total = 192 sequences	32 seen subjects 3 sequences per subject (emotional) 4 sequences per subject (neutral) Total = 1646 sequences	4 unseen subjects 5 sequences per subject 9 conditions per sequence Total = 180 sequences

Table 4: Datasets Specifications

5.3 Metrics Inventory Results

In this subsection, we will present the results of the metrics provided from training the models on the datasets. We will present each dataset's results for each model for all metrics we tested to get a first idea of the outcome of the experiment which we will analyse in further sections.

Firstly we will present the results for the BIWI dataset [12] which is the dataset that all the authors of the respecting models have previously tested. For that dataset, the process was completely straightforward since all of them have been similarly implemented for that dataset as we intended to do. The results of that part of the experiment are visible in Table 5

Model	LVE (↓) $x10^{-4}$	FDD (↓) $x10^{-5}$	MVE (↓) $x10^{-3}$	Diversity (↑) $x10^{-3}$	MEE (↓) $x10^{-4}$	CE (↓) $x10^{-4}$
FaceDiffuser [38]	4.9462	4.4298	6.8088	0.0024564	**4.9515**	**4.99313**
CodeTalker-ND [55]	6.333	5.1863	7.357	**0.3234**	6.670	6.180
CodeTalker [55]	4.7915	4.1172	**6.0130**	-	-	-
FaceFormer [11]	4.9282	4.6275	7.1352	-	-	-
FaceXHuBERT [16]	**4.7293**	**3.9006**	6.2955	-	-	-

Table 5: Objective Evaluation Metrics for the Models on the BIWI Dataset

After completing our tests for the BIWI dataset we moved on to test the results of the metrics after training all models on the Multiface dataset [54]. Due to Multiface being a similar dataset, even though the exact implementation or instructions were not given the process was similarly straightforward. The dataset has a similar structure and the results from some of the papers came out useful for us to cross-reference some of our findings. The outcome of our experiment on the Multiface dataset is presented in Table 6.

Model	LVE (↓) $x10^{-4}$	FDD (↓) $x10^{-5}$	MVE (↓) $x10^{-3}$	Diversity (↑) $x10^{-3}$	MEE (↓) $x10^{-4}$	CE (↓) $x10^{-4}$
FaceDiffuser [38]	**5.7662**	5.1600	**6.7387**	0.00079924	**5.7667**	**5.7656**
CodeTalker-ND [55]	20.262	9.3500	14.288	**0.011661**	21.6225	20.059
CodeTalker [55]	16.94	5.4309	10.945	-	-	-
FaceFormer [11]	13.405	6.9343	8.4464	-	-	-
FaceXHuBERT [16]	18.001	**5.0023**	9.1676	-	-	-

Table 6: Objective Evaluation Metrics for the Models on the Multiface Dataset

Moving on to the final dataset that we tested the models on, which is 3DMEAD [48], it seems to have the most interesting results for the metrics. As the dataset is so much different than the others and it also has never been tested on those models it is hard to make strong judgements towards the models' performance on that dataset. Our main aim was to use the models exactly as intended by the authors without changes such as emotion control which could be useful for the dataset since we did not want to interfere with the models themselves. The results of the metrics on the models being trained on 3DMEAD are presented in Table 7.

Model	LVE (↓) $x10^{-4}$	FDD (↓) $x10^{-5}$	MVE (↓) $x10^{-3}$	Diversity (↑) $x10^{-3}$	MEE (↓) $x10^{-4}$	CE (↓) $x10^{-4}$
FaceDiffuser [38]	0.89389	0.090677	1.3242	0.044977	0.88488	0.87845
CodeTalker-ND [55]	3.1325	0.30893	3.4783	0.1696	3.1593	3.0205
ProbTalk3D [52]	**0.6040**	**0.04515**	**0.7243**	**0.3274**	**0.5549**	**0.5227**
CodeTalker [55]	1.5978	0.20364	1.7121	-	-	-
FaceFormer [11]	2.0266	0.065979	2.8548	-	-	-
FaceXHuBERT [16]	2.969	0.085135	3.1428	-	-	-

Table 7: Objective Evaluation Metrics for the Models on the 3DMEAD Dataset

5.4 Subjective Metrics

After the completion of the experiments focused on training the models on the different datasets and calculating the objective metrics, we aimed to complete a perception study to gather subjective metrics from users rating the results. The two aspects of the rating were going to be the realism and lip-syncing accuracy of the results, rated individually as standout motions. For our user study, we decided to only use the 3DMEAD dataset as most of the models have already conducted their own perception studies containing most of the models on the BIWI and the Multiface datasets. The standard practice in perception studies for 3D facial animations is the use of A/B testing which aims to directly compare the proposed model with other state-of-the-art models and the ground truth. Since our aim is not to prove the performance of a specific model,

we decided to not use this methodology and go for an individual rating process in which the user does not compare two motions at the same time. After generating four random predictions for each motion for all the models trained on the 3DMEAD dataset we rendered the results for all the models and the Ground truth, for the user to rate. One random example from the four rendered results for each motion was randomly shown to the user adding also intermixing between the emotions and the models, with the aim of eliminating the chance of favorability. Each of those motions is going to be rated on a scale of 1 to 7 in terms of realism and lip-syncing to the audio performed, with 7 signifying best. An example of how the screen for each individual video looks is shown in Figure 16.

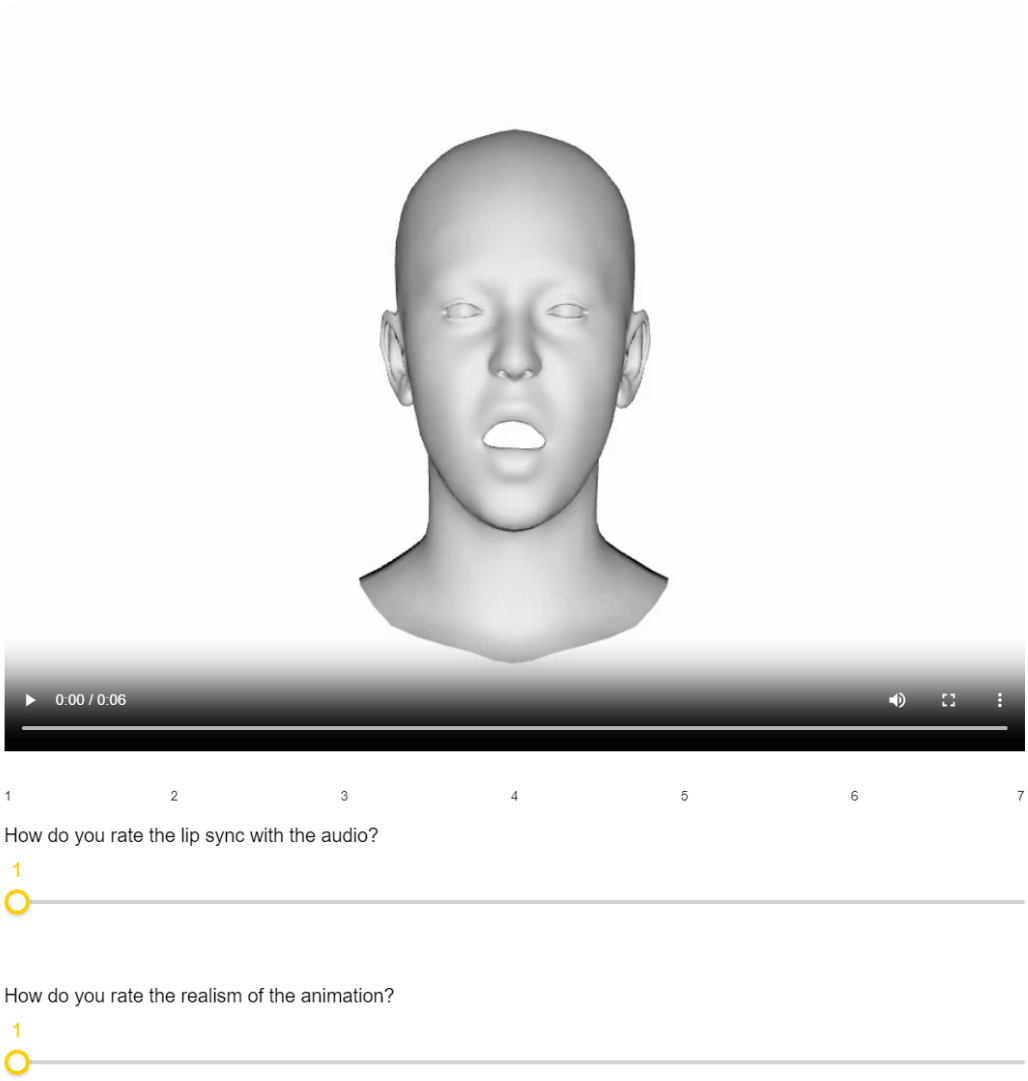


Figure 16: User Study screen for each individual motion showcased to the user

Due to the large number of individual motions, each user would have to rate, 6 models + GT for every emotion resulting in 56 motions, we decided to randomly split the motions into two parts. The way the split was executed was by randomly placing each user in a group where the motions of 4 random emotions for all models were going to be shown to them. That way we avoid the case that the study takes too long and the users do not pay attention to how they rate

the motions. The complete contents about the process followed while conducting the User Study are presented in Section 8.

To gather participants for our User Study, we used a website that allows researchers to pay the desired amount of users to complete our study. Using that website we gathered 45 users and also used the responses of 16 more volunteers that helped us in our study. We also made sure that there was balance for both the time it took to complete the study and the results themselves, between paid users and volunteers to ensure unbiased and accurate results. Using the answers from all users for each individual motion they watched, we would use those results to calculate the ratings mean for each model and the GT. The two aspects that the users are required to rate, differ from each other for both the user and the way we are going to utilise them. The user has to rate lip-sync based on how accurate the lip movement is in terms of the audio that is performed, while realism is a broader and intuitive rating of how accurate the facial movement is overall. For the analysis, we will use the results of lip-sync to directly compare with the results of LVE for the models during our objective metrics study. On the opposite side, we will use the rating for realism as a complete interpretation of the achievement of the models to create good results while comparing them with all the results for the objective metrics. After conducting our Perception study with 61 recruited users we calculated the mean result for each of the two metrics across the models, the results are presented in Table 8.

Model	**Lip-Sync (↑)**	**Realism (↑)**
FaceDiffuser [38]	3.861	3.194
CodeTalker-ND [55]	3.753	3.030
ProbTalk3D [52]	3.671	3.118
CodeTalker [55]	3.574	2.992
FaceFormer [11]	3.858	3.335
FaceXHuBERT [16]	4.020	3.345
Ground Truth	**4.035**	**3.375**

Table 8: Subjective Evaluation Metrics for the Models on the 3DMEAD Dataset

6 Results Analysis

In this section of our research, we will focus on analyzing the results previously gathered from our methodology to answer our research question and conclude on different factors concerning 3D facial animations. As most of our results have been previously showcased as part of our methodology we will use those results for our analysis expanding the understanding of certain topics.

6.1 Nature of Facial Animations

Firstly, the main goal of the process followed and presented in previous sections was to identify the possible classification of the nature of Facial Animation as deterministic or non-deterministic. Using the resources available we aimed to answer that question in numerous different ways all connected with each other and focusing on multiple aspects. As explained earlier, that process would have been more straightforward if a dataset containing multiple samples of the same motion existed but due to the lack of this kind of dataset we aimed to separate that question in different levels. Using the results gathered from our processing, we aim to first identify the different factors that can suggest the existence of non-determinism for Facial Animations. The first obvious factor that we can identify is the distinction across different subjects uttering the same sentence.

Subject Non-Determinism: To determine the possible existence of that variety we first need to understand its meaning. We aimed to identify and provide evidence of the fact that even if multiple people perform the same sentence in the same context, sometimes even with emotion and intensity control, not a single sequence will be the same as another. The next step is to determine how far apart those performances have to be from one another to suggest that they are completely uncorrelated between them and the only aspect that affects them is the subject performing them. For that exact reason, we aimed to use as much data as possible to have a more holistic view and not try to select specific points which might end up more confusing than helpful. Using the 3DMEAD [48] dataset we calculated the mean jaw movement of each subject performing sequence '001', across all emotions and intensities, as seen in Figure 7. The results of this test show a huge difference across the subject's mean jaw movement, which essentially controls the opening and closing of the mouth, even for one specific sequence. Subjects such as 'M024' and 'M029' show a really small movement across the jaw in all axes when in contrast with subjects such as 'M011', 'M034' and 'W019' which show immense movement in the jaw area. That overall variety across the subjects is a result of the different speaking styles that each subject has and can be seen even by examining the ground truth. Moving on, to have a more detailed and overall overview of how the data are de facto separated by the subject performing the sequence we will analyse the various dimensionality reductions we conducted on that aspect. On a path to the mean jaw movement test, we conducted t-SNE on sequence '001' with colour mapping based on the subject. That result which can be seen in Figure 8, produces a clear distinction between the sequences performed by the same subject. Clusters seem to be created with the sequences not affected by the emotions or intensities but by the subjects, showcasing a high correlation between the ones performed by the same subject. Even if some of them showcase some intermix with some neighbouring clusters the distinction between them is too obvious to be affected by such small outliers. In some cases, the clusters contain some mini-clusters of them which, after examination, are sequences from the same emotion group which will be examined closer

later. In the result using the whole dataset instead of a sequence it becomes obvious that some clusters still exist but not to the obvious extent that they can be identified when using a single sequence. The usage of PCA for the 3DMEAD dataset as a whole provided us with a more useful result in terms of understanding the clustering between the same subject sequence. The result of that, which is seen in Figure 9, showcases an interesting variation of clustering with most of the clusters being closely connected and intermixing with others while at the same time, their shape is outlined clearly. Due to the nature of that dimensionality reduction technique, it is harder to showcase a more extended distance between the huge chunks of data points from one another, even if the clustering is obvious. Another interesting factor is the three bigger clusters that seem to be created with the one at the top being mainly female performances and the two at the bottom being male performances. This adds to the layers of style that are responsible for the non-determinism as it shows the gender's significant difference in the performance of the sentences. Moving on from the subject-oriented analysis of 3DMEAD, we aim to analyse the results of the similar process we conducted for the BIWI [12] and Multiface [54]. The results of this process were clearer due to the consistency that exists among both datasets and also due to the lack of emotion categorisation which might disturb the clear view of the dimensionality reduction. The t-SNE results for both datasets, as seen in Figure 12, there is a clear distinction between the clusters of each subject with also more notable compactness between the data points of each cluster. Interestingly both datasets which hold a similar number of subjects and sequences per subject also have a really low number of outliers with almost no data point exiting the limits of a very compact cluster. The results themselves prove an extremely high correlation between the sequences performed by each subject at a higher level than expected even if the consistency among those datasets was previously stated. Similarly when PCA was conducted on those datasets, Figure 13, the data points tended to follow that pattern. A minor difference that can be stated is the rare intermixing between the clusters of certain subjects in the BIWI dataset and the data points being more spread overall. Still, both datasets and dimensionality reduction techniques prove an extreme amount of correlation of the same subject sequences. Moving on to the last dataset we tested using this methodology and focusing on the subject non-determinism, the BEAT dataset [23], we followed the same process as before to generate similar results for a precise analysis. The inconsistency of the BEAT dataset which does not follow the structure of the rest datasets, with not all subjects performing the exact sequences is an interesting factor to identify in that process. The results after conducting the t-SNE, Figure 14, showcase a more complex situation on the data points than the rest datasets. Even if most clusters are clearly visible, in that specific case a lot of outliers and intermixing of the clusters takes place. Some of the subjects have data points that spread all around the axes while others are more compact with data points closely mapped. That variety of subjects having compact or loose clusters comes as a result of the inconsistency of the dataset and the structural difference with others, meaning that some of the data points that become outliers from their clusters are sequences not performed by all subjects. The extent of this problem is mainly visible when conducting PCA on the BEAT dataset, Figure 15, which as proven before is a lot less forgiving towards the non-compact clusters. As a result, the clusters of the subjects when using PCA are a lot harder to distinguish between one another while still following a pattern which can be identified after closer examination. As a final evaluation for both dimensionality reduction techniques on the BEAT dataset, it is clear in both cases that this dataset performs a lot worse in distinguishing the same-subject data points due to the major inconsistencies of the data.

As an overall conclusion from the subject-focused non-determinism of the datasets, we seem to have used appropriate tools to identify and prove the existing assumption. As seen in all datasets there can be a clear distinction between the same subject data points, meaning that the speaking style that every person has, influences those sequences the most. The apparent evidence of that was mainly dimensionality reduction techniques to use the raw data provided from the datasets, with a resulting clustering of the data points from the same subject. The length that this clustering exists is dependent on the consistency of the dataset and other factors that may influence it, such as emotion categorization, which will be the next factor we will examine.

Emotion Non-Determinism: The second factor of the potential existence of a probabilistic nature is the emotion categories of the performances for the sequences by the subjects. As the inclusion of emotion categories has been a more recent addition in 3D facial animation datasets, we only find those categories in two datasets, 3DMEAD and BEAT. By similarly examining those two datasets as before we aim to identify if the sequences from the same emotional category are more tend to have a correlation between them proving the existence of a non-determinism inside the same subjects. As the subjects are required to perform the same sentences using different emotions, and sometimes intensities, the potential correlation between those emotions and not the sentences will help us understand the nature of facial animation more deeply. For the first part of our analysis of the datasets and the results we used the same t-SNE experiment as before but this time we used it on one specific subject. Using the 'M003' subject from the 3DMEAD dataset and all its sequences with emotion category colouring, provided us with results that prove the existence of a clear distinction between the same-emotion data points, as in Figure 10. It becomes apparent that even if the sentence is the same the most important factor that affects the final sequence is the emotion that the subject performs on it. The same individual test was done for most of the subjects and all of them suggest that correlation between the same emotion category data points. In those specific cases, we can identify after closer examination that the only data points that tend to become outliers and expand further than their emotion clusters are the data points with intensity level 1, the lowest one, proving that only if the subject aims to subtly perform the emotion it might end up being too subtle and hard to distinguish with other emotions. When using the whole dataset to conduct the emotion-colouring t-SNE we come across a different form of result. The result when using the whole dataset showcases smaller chunks of clusters that contain data points from the same subject and emotion category. The data points in this case show the connection of both factors and the data points being affected by both. The useful insight that is provided in terms of the correlation between the same emotion data points, is that even in the bigger picture the sequences of the same emotion and subject are still closely correlated. Interestingly, when testing the same idea with the whole dataset for the 3DMEAD using the PCA dimensionality technique we come across a different result. In that case, all the emotions are spread around the area equally without any obvious signs of clusters being made. That results in the acknowledgement of the PCA technique which makes the distances between data points smaller even in prior tests that showcased clusters. With really close inspection we can see the existence of some really broad clusters that span across bigger parts of the data but nothing that can lead us to the idea that emotional correlation exists between the data points. That additional result and by of course taking into account the previous ones we can easily conclude that even for the 3DMEAD dataset which is focused on emotional categorisation, there is little to no connection between the data points of the same category. That conclusion can identify the potential lack of obvious existence for non-determinism

based on the emotion category. In order to conclude even further, we needed to also adapt these ideas to the BEAT dataset, which is the only other dataset with emotion categories, using the same techniques as before. The results of implementing a t-SNE test on the BEAT dataset with an emotion category colour mapping are visualized in Figure 14. Those results give a similar idea to the results of 3DMEAD which provided us with a mixed-up outcome. We can see most of the emotions spread across the axes with no clear distinction between the categories. The inconsistency of the dataset might also be a leading factor to that issue, but it becomes evident that even the t-SNE is unable to create any form of minor clustering there is no clear correlation between the emotion categories across subjects. The BEAT dataset also provided some imperfect results in terms of clear identification of the classifications even with the subjects, however, the results about emotion categories suggest the same thing we came across with 3DMEAD. In alignment with the previously mentioned, we conducted our final test, PCA on BEAT with emotion colour mapping 15, with the results confirming the initial assumption. The outcome produced makes it impossible to identify any form of clustering, with the data points all over the axes. Once again, especially for the BEAT dataset, it becomes evident that the dataset is not consistent with a clear correlation between sequences from the same emotion category across different subjects.

Overall Results: After analysing the results of our methodology to identify the possible existence of non-determinism in facial animations, we will summarize our interpretation of the results and their analysis. The process that was followed allowed us to be able to make direct comparisons between the datasets and their results mainly when using the dimensionality techniques. With our existing impression of what a non-deterministic result was supposed to be, we were able to evaluate the results. The hypothesis used that data points of the same subject or emotion of multiple sequences were supposed to showcase a strong correlation, visualized with the clustering, which would prove the high connection between them. Having four datasets, with the variety they had between them, proved to be of much help since the results were what we expected after carefully analysing the datasets before the process. As understood from the analysis the more consistent a dataset is the more the non-determinism can be identified using those methods, while the two datasets that had two or more kinds of categorizations of the sequences were the ones with less clear results. In addition, even if the 3DMEAD dataset also contained intensity levels, it proved to be a more compatible dataset in the representation of subject-based non-determinism in most cases. As previously explained, the nature of the existing datasets does not allow a concrete and straightforward process to identify the same parameter sequences as different, since they do not exist. The need to subvert the process helped us understand other interesting parameters of facial animations in terms of their non-determinism. When directly comparing the differences in the results of the subject-based and emotion-based tests on the two datasets with emotion categorisation, proves that the most important parameter that affects the data points is the subjects uttering the sentence. Since the difference in the visualization of the correlation between the data points is apparent, we can easily suggest the existence of more non-determinism between the subjects than the emotions. Conclusively, even if two subjects utter the same sentence with the same emotion, the data points are not correlated as much with one subject uttering different sentences with different emotions.

These findings suggest that individual differences play a major role in how unique facial animations are. This has important implications for creating more realistic and personalized animation systems. To make animated characters look more natural, it's crucial to consider how different each person's facial movements can be. Additionally, the weaker correlation found

with emotions means that while emotions do change facial expressions, they don't overshadow the unique traits of each person. This insight can help future research focus more on capturing individual facial details rather than just relying on emotions. By using these understandings, researchers and developers can create better animation systems that closely simulate real human facial behaviour, leading to more engaging and believable digital experiences.

Datasets parameters: An additional component we aimed to specify was the possible differences between 3D facial animation datasets with datasets from other similar research fields, in terms of ground truth data parameters. We gathered data from the main datasets we used for our study and compared them with the data gathered from datasets from the body motion and 2D talking heads research fields. Exploring the data from the state-of-the-art datasets from this research field could provide us with insight into the depth of variables and categorization they contain as well as quantity-wise advantages from them. Firstly, we have to acknowledge the differences that the datasets in these research fields hold, which makes it hard to have a direct comparison with the 3D facial animation datasets. However, there are some factors, such as the number of subjects or the total amount of content gathered, which could be an indicator for comparison. One factor that we also wanted to compare is the possible categorization of the sequences such as the emotions and intensity levels that also exist in 2D talking head datasets. That category is not directly comparable with the categories of body motion datasets even if it has some representation as the action labels given to the sequences. With those in mind, we will try to analyse the differences in the factor suitable to do so but also keep in mind the rest data from the datasets of all the fields. Between the datasets of 3D facial animation and the 2D talking head, we can recognise the existing pattern of almost all 2D datasets containing more subjects than all the 3D ones. Only the GRID dataset contains one less subject than 3DMEAD which we already recognised as the most recent state-of-the-art dataset for 3D facial animations. With huge differences between the rest datasets such as the BIWI and Multiface after comparing them with CREMA-D and TCD datasets, which have more than five times the number of subjects performing the sequences. That clear comparison can also be of additional value in terms of the variety that those datasets contain affecting the non-deterministic nature of the animations. With expanding the number of subjects performing the sequences in the ground truth, the variety of speaking styles covered is broader which helps the generation of more diverse results. Additionally, another factor that seems to showcase a difference between the datasets is the amount of sequences performed. The sequences performed by the subjects are a harder topic of comparison due to the influence of other factors on it. As the BEAT dataset contains the most individual sequences performed, it lacks consistency meaning that not all subjects perform them in all emotional categories. On the other hand, we can identify that the CREMA-D dataset selected just 12 sequences to perform in all 6 emotions and 4 intensities, similar to 3DMEAD. One of the datasets that stands out for their individual sequences is the GRID dataset which uses 1000 sequences performed by the subjects covering all phonetic possibilities. That also affects the total length of the dataset in hours which is another indicator for us, with the GRID dataset containing 40.51 hours worth of data, with BEAT and 3DMEAD having 31 and 26 respectively. For the total amount of hours, the BIWI and Multiface fall behind a lot with only 0.7 hours as they contain way fewer subjects and no repetitiveness of the sequences with the emotion categorisation. The CREMA-D and TCD datasets both contain around 11 hours of total data with the TCD dataset following a typical pattern with many subjects and many sequences but no emotional categorisation. For a better understanding of the differences between the datasets, we

showcase a visualization comparing the number of subjects and the number of sequences per subject in the 3D and 2D datasets in Figure 17.

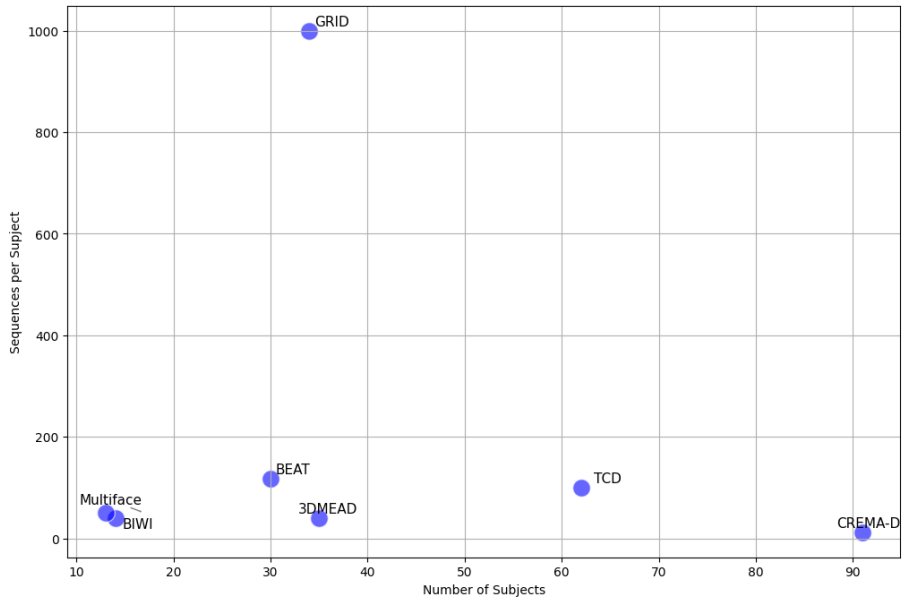


Figure 17: Visualization of the number of Subjects on the X-axis and number of Sequences per subject on the Y-axis for all datasets

It is evident that one of the standout factors is the aim of the 2D dataset to cover as much as possible for either subjects or sequences from the ground truth and that in the end, the datasets are more fulfilling for the models. The effect on some of them not including the emotional categorisation does not affect the amount of content for those datasets, something that undoubtedly exists for 3D datasets. The second research field that we aimed to compare directly with the 3D datasets is the datasets used in various body motion methodologies. That area is harder to correlate with for the 3D datasets as it contains various techniques in both the capturing and usage of the datasets, but we aim to identify the usable factors for comparison. One clear comparison again can be made with the number of subjects that exist within the datasets, and in that part, the body motion datasets stand out a lot. With all of the datasets, that we selected as the useful state-of-the-art for our study, having more subjects than the 3D face datasets. Interestingly, the only dataset that is close to the 3D face datasets in terms of subjects, is the IDEA 400 which is a MoCap subset used to generate the Motion-X dataset through a deep learning pipeline. In terms of the total amount of hours in the datasets, we can also identify a similar pattern as before with the datasets ranging from 11 to 43 hours. Once again the two of the main datasets mainly used from almost all of the recent state-of-the-art 3D facial animation models fall behind a lot in terms of the total amount of data. The body motion datasets also contain a specific number of actions that correspond to the movement that is conducted by the subject, something that does not correlate directly with anything concerning the 3D face datasets, but still showcases the depths of different data that are gathered from these datasets. A direct visualization of the number of hours contained in the datasets and total sequences is presented in Figure 18.

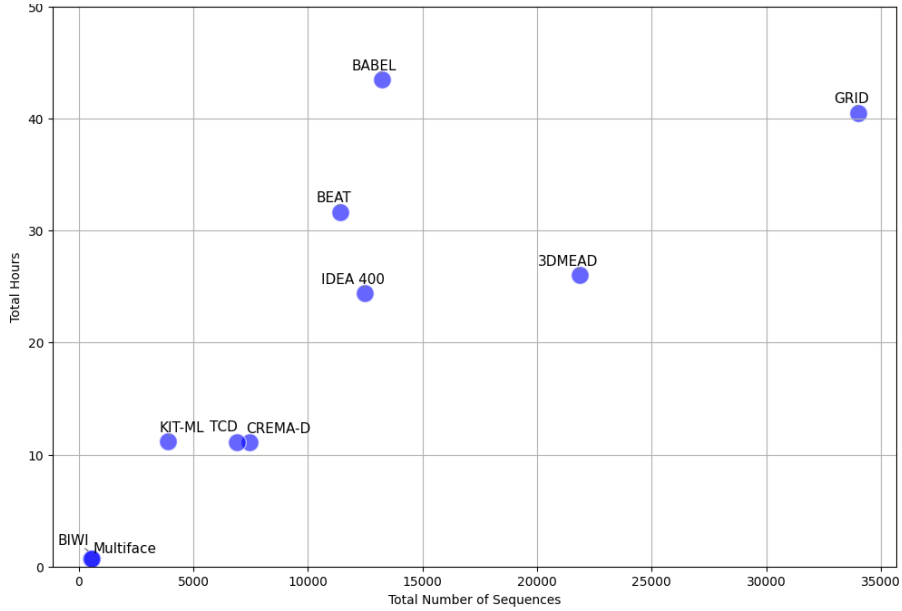


Figure 18: Visualization of the total amount of hours on the Y-axis and the total number of Sequences on the X-axis for all datasets

The comparison between 3D facial animation datasets and those from related fields like 2D talking heads and body motion highlights key differences in dataset parameters. Notably, 2D talking head datasets contain a significantly larger number of subjects, providing a broader variety of speaking styles and enhancing the diversity of results. This suggests that expanding the number of subjects in 3D facial animation datasets could similarly improve their robustness. In terms of sequence quantity and total dataset duration, datasets like BEAT, despite having many sequences, lack consistency, especially between emotion categories. In contrast, datasets such as CREMA-D and TCD, with structured emotional and intensity variations, offer a more comprehensive approach. Future 3D facial animation datasets should aim for this level of comprehensiveness to generate realistic animations adaptable to various emotional contexts. Compared with body motion datasets, which have more subjects and extensive data, underscores the need for larger and more diverse 3D facial animation datasets. Adopting similar data collection and categorization methodologies from body motion research can enhance the robustness of 3D facial animation datasets. Expanding and diversifying these datasets will better capture the variability of facial animations, leading to more realistic and adaptable models.

6.2 Objective Metrics Results

In this subsection, we will analyse the results that we gathered from training, testing and calculating the metrics of the selected models on three different datasets. That experiment aimed to provide a concrete idea with all the rest of the variables the same about the evaluation metrics and not the models themselves. We do not aim to identify the model that performs best on those datasets in terms of the metrics but rather try to identify which of the metrics are good in guiding us to that path. Keeping in mind the different factors that the models were selected based on, we aim to answer the question about the different insights that the objective metrics provide us. Using the advantages of each metric as well as the knowledge we have of the models' differences, we aim to achieve a holistic exposition of the inventory especially after analysing

the metrics results in-depth.

BIWI dataset: The first results that we are going to analyse are the ones from the BIWI dataset [12] with training FaceDiffuser[38], FaceXHuBERT[16], CodeTalker[55], both deterministic and non-deterministic, and FaceFormer[11] as shown in Table 5. Firstly, the LVE metric showcases a small distribution among the models with all of them performing similarly well with not many differences. The models being implemented in a way that they focus on the achievement of good lip-sync as well as the fact that this is one, if not the main, metric that those models tested their results in their ablation studies are the two main reasons behind that. As this metric is a distance metric between the ground truth and the generated data, it proves the capability of the models to generate results that are near the ground truth motion around the lip region. A parallel case exists for the MVE metric where we have most models performing similarly and CodeTalker showcasing a better result, in contrast with LVE where FaceXHuBERT performs slightly better. As this metric is also a distance metric comparing the ground truth and the generated results, we can identify again the capability of the models to achieve results close to the ground truth, this time for the whole face region. The difference in the absolute values of MVE and LVE shows that the models are more capable of generating correctly lip-synced results than the whole face, something that we expected. The reason behind that is both the LVE being used more often for testing reasons and the models being approached in a way that they focus on achieving the best lip-sync possible. One of the models that seems to be a slight outlier is our implementation of CodeTalker-ND, even if the model does not perform badly, it has the highest value in those two metrics. To show a more transparent view of how these two metrics show their results we present a graph with both metrics as the axes and all the models placed according to their performance in Figure 19.

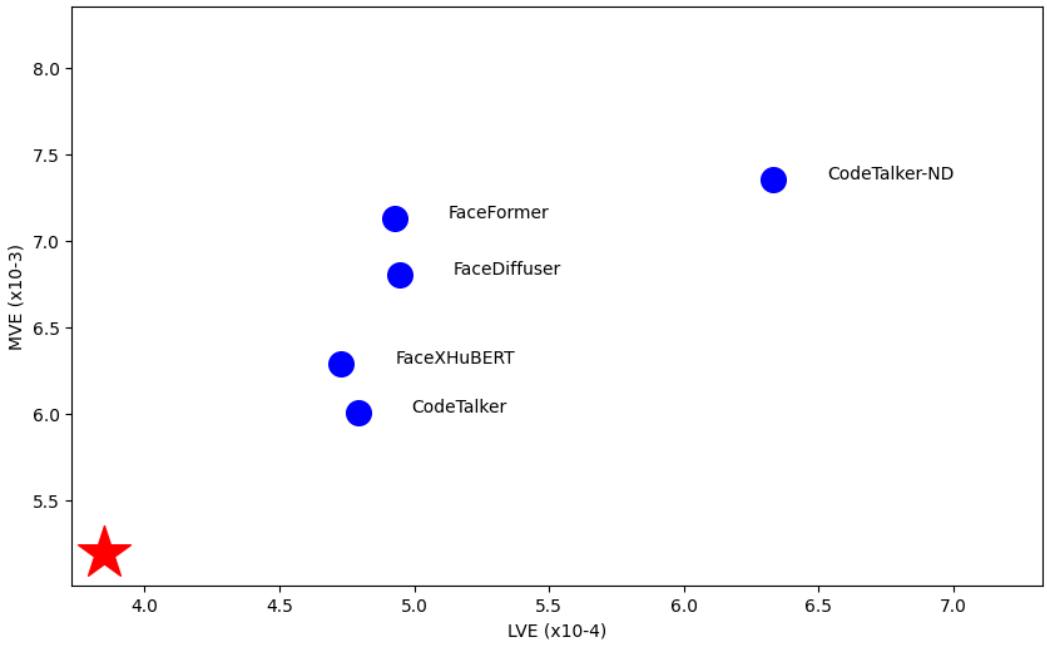


Figure 19: MVE and LVE results on the BIWI dataset visualized

In both metrics, the visualized result seems to indicate an advantage to FaceXHuBERT and CodeTalker while still, all the models performing regularly well and with results close to each other. A similar case can be interpreted for the FDD metric which tests the movement differ-

ence of the upper face of the GT and the generated motions. All the results are in a small range, showing the small difference that exists between the movement of the GT and the models' results. As this part of the face is usually not the one that is moved a lot, especially in non-emotional datasets, a clear interpretation is the fact that those models do not reshape a lot the upper face when generating the results. The metrics that test the non-determinism of a model were only calculated on the different samples that were generated for the same motion on the FaceDiffuser and CodeTalker-ND. Due to the non-determinism of the models, they can generate different samples given the same parameters for the same motion sequence. For this process, 10 samples were generated for each motion and used to calculate the different metrics that require different samples, Diversity, Mean Estimate Error and Coverage Error. Using the samples to calculate the Diversity metric we separate the samples into two subsets to calculate the difference between each subset position and calculate the mean for each motion. The Diversity results showcased should not be confused with the Diversity stated in the paper of FaceDiffuser as it contains a different calculation using the multiple samples generated, previously referred to as DivE. The result of this metric showcases a relatively small Diversity between the samples even if FaceDiffuser's main component is diffusion, but that should not be taken as negative. The fact that the model performs equally well with the deterministic models in terms of the metrics that compare the generated results with the GT proves that it covers both a realistic and lip-sync accurate result and the capability of generating different samples. That effect is highly highlighted when we compare the Diversity results of CodeTalker-ND, which has a higher Diversity value but provides worse results for the metrics calculating the difference with the GT. That proves that CodeTalker-ND is able to generate different samples with enough distance between them but not always accurate. In terms of the two other metrics that use the calculations of the LVE on the different samples, MEE and CE, we can identify results that are parallel to what we expected for FaceDiffuser. Due to the low Diversity that we identified, we expected the samples to be similar between them and those two metrics prove exactly that. The coverage error proves the fact that there exists a sample that is slightly closer to the GT, whereas the Mean Estimate Error shows the generated samples are not far from each other when compared with the LVE. For CodeTalker-ND, we can identify the effect of the higher Diversity as the CE and MEE are further apart from the LVE as the results differentiate between them on a larger scale. Recognising that pattern that happens the opposite way on those two non-deterministic models helps us understand the way the models perform when generating different samples. As far as the model achieves to generate results close to the GT the difference between the results will be minor, and vice versa. Even if the main aim of our study is not to evaluate the models themselves, but rather use that evaluation to understand the metrics, we aim to gather the ranking based on the models' performance for each metric. That ranking will help us understand the final results of the metrics and explain the differences between our results and the claims from the corresponding authors. The ranking of the models for each metric is presented in Table 9.

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	-	-	-

Table 9: Ranking results of the models on the BIWI dataset for each evaluation metric

The models perform similarly on the deterministic metrics with CodeTalker and FaceXHuBERT standing out as the best for those metrics. The fact that all of those models perform differently, even with slight changes to the number, is a result of the experiment settings being different than the ones that the authors of those papers had when doing so. While every author has different local experiment settings slightly which might even lead to ranking changes as the numbers are that close. Nevertheless, one result that stands out for the BIWI dataset is the FaceDiffuser's low ranking, contrary to the author's experiment results. The main reason behind that is the different approach we took when generating the results, with multiple samples generated for one motion instead of the different conditioning from subjects. For the calculation of those metrics we randomly selected one of those samples that are generated from the motions to calculate LVE, MVE and FDD, which is the same process done for CodeTalker-ND. The fact that there was no different conditioning and no use of those results from the conditioning to calculate the difference with GT, led to the different results and the lower ranking. Nonetheless, the ranking on those metrics should not disregard FaceDiffuser's capability of both the variety in results and the accurate representation of GT.

The analysis of the BIWI dataset reveals that the models FaceDiffuser, FaceXHuBERT, CodeTalker, CodeTalker-ND and FaceFormer perform similarly in achieving good lip-sync, as evidenced by the LVE metric. This metric, which measures the distance between ground truth and generated lip movements, indicates that all models generate lip-sync results close to the ground truth. FaceXHuBERT shows a slight advantage in LVE, while CodeTalker excels in the MVE metric, which assesses the whole face region. The consistency across models in these metrics suggests a strong focus on lip-sync accuracy. However, the MVE values being higher than LVE indicates a greater challenge in accurately generating full-face movements. The FDD metric, which evaluates upper-face movement differences, also shows minor variations, reflecting the models' conservative approach to upper-face motion, typical for non-emotional datasets. For the non-deterministic FaceDiffuser model, additional metrics like Diversity, Mean Estimate Error and Coverage Error were calculated using multiple samples. Despite a low Diversity, indicating similar outputs across samples, FaceDiffuser maintains realistic and accurate lip-sync results, relative to deterministic models. The MEE and CE metrics confirm the generated samples' consistency and proximity to the ground truth, reinforcing the model's effectiveness in producing diverse yet accurate results.

Multiface dataset: After understanding the results of the BIWI dataset we attempt to make a similar approach to understand how the metrics correspond to the results of Multiface. After conducting our experiment with Multiface we calculated the same metrics, using the same calculations, which are visualized in Table 6. Starting with LVE, we can already identify a completely different scenario happening with the Multiface dataset, with clear and broad differences between the models. The model that stands out with its performance on the Lip Vertex Error is the FaceDiffuser with less than half of the error than the second best. In that context, the only two models that specify their calculation results for the Multiface dataset are FaceDiffuser and FaceXHuBERT. The results concerning the LVE provide us with a clear distinction between the models' capabilities in terms of generating results close to the GT in the lip region. The FaceDiffuser model can more accurately represent the GT movement focused on the lip region with the rest of the models having higher absolute values. To be more precise with our understanding we can also use the results of the MVE to make a comparison since those two metrics are using the same calculation. Even if FaceDiffuser performs better than the other models also in MVE,

it does not have that huge gap with the rest of the models, which exist for LVE. The only model that systematically stands out as performing the furthest from the GT is the CodeTalker-ND which due to its high non-determinism does not achieve an accurate representation of the GT. Additionally, if we make a clear comparison of the absolute distance between the results of BIWI and Multiface for LVE and MVE, we can identify an improvement concerning the MVE of Multiface. By understanding the fact that Multiface contains even less movement around the upper face which results in a more accurate representation of the results as the metrics suggest. That proves the need to understand the content and the advantages of each dataset and how those affect the result on the metrics before reaching any conclusions. The effect the data has on the results is obvious in this case as there exists a bigger focus around the lip area resulting in the generated results not being able to identically align with the ground truth. The effect of those two metrics and the holistic results of the models are visualized in Figure 20.

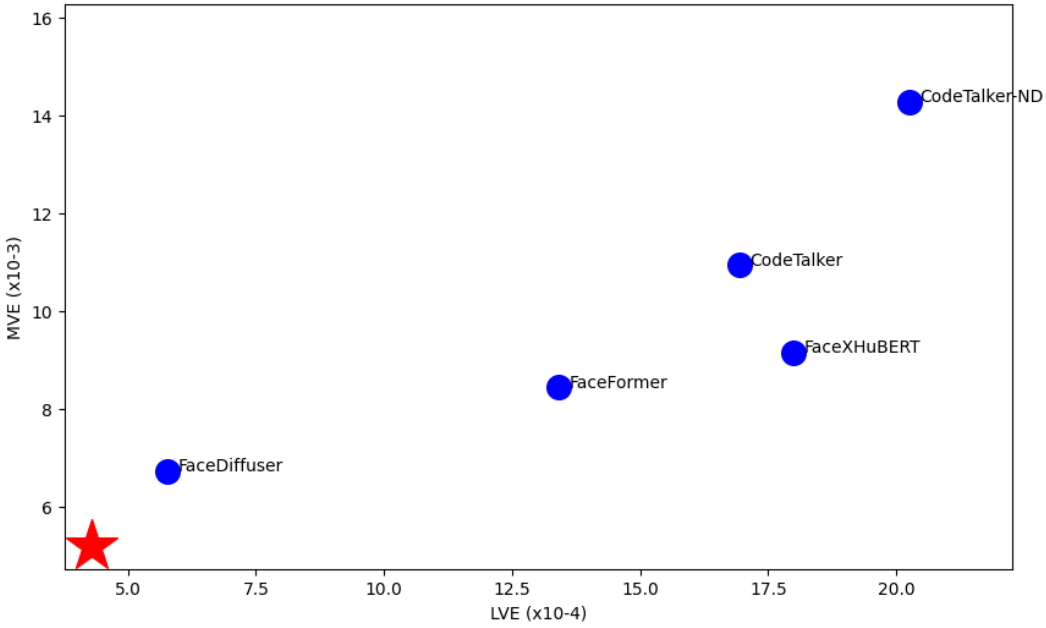


Figure 20: MVE and LVE results on the Multiface dataset visuialized

Using the information gained from understanding the LVE and MVE we can also identify a similar pattern existing for the third metric that calculates the difference between the GT and the generated results, FDD. Even if FDD calculates the standard deviation of the movement of the upper face we can see again all the models providing closer results than for the lip region as expected. FaceXHuBERT is the best one concerning the absolute value of those differences with FaceDiffuser being a close second. As FaceXHuBERT provided the best results for both datasets on that metric and after understanding its architecture we can also understand that it was one of the first models to focus also in an accurate representation of the whole face, including the upper face and not only the lip region. That also can be derived from the fact that it was the first model to include the L2 error (MVE) for the whole face and not only the lip region, which seems to have also been used in tests in ablation studies. The insight we gain from that analysis is that one big advantage of the objective metrics is that they can be valuable to the authors of the models while running tests to improve specific factors of their models to achieve specific capabilities, and FaceXHuBERT is a clear example of that. To better understand the results of

the FDD on those two similar datasets we present a visualization of them concerning the models and the axes being the FDD for the two datasets, that visualization is presented in Figure 21.

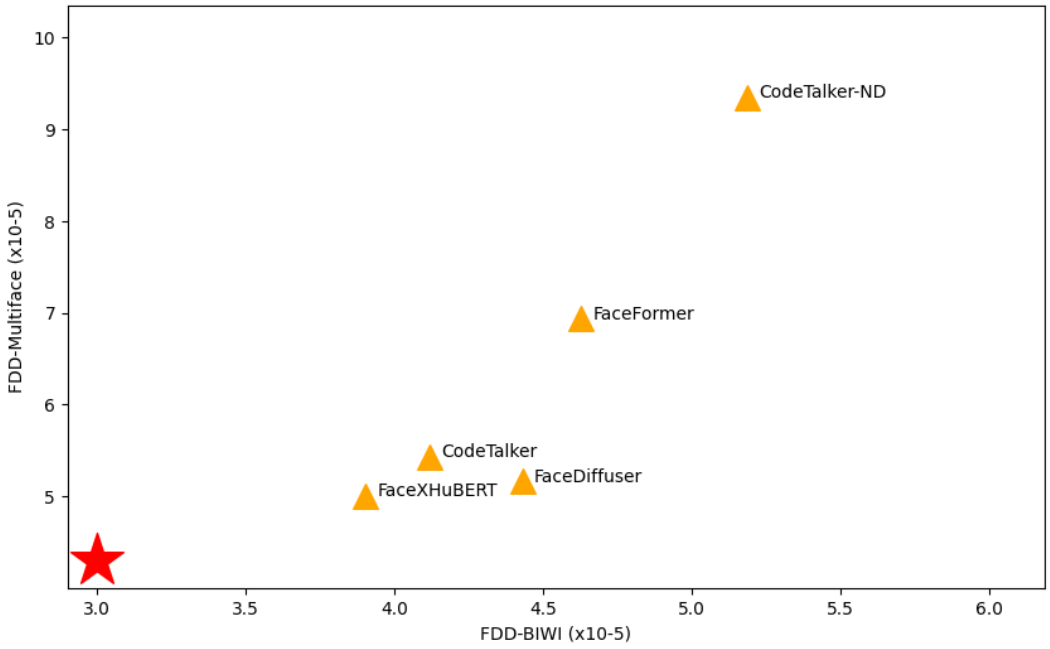


Figure 21: FDD results representation for BIWI and Multiface datasets

Moving on to the last three metrics that we calculated only for FaceDiffuser and CodeTalker-ND due to only them being able to generate different samples for the same motion, we can identify interesting results. Firstly, for FaceDiffuser the Diversity metric showcases a lack of capability to produce highly diverse results and they only contain minor differences. This does not come as a surprise after we identify FaceDiffuser's performance of the 3 metrics concerning similarity with the GT. As the model is capable of generating results that are this close, at least closer than the rest, to the GT it is not able to also generate too diverse results. In most cases that is the option that needs to be made for non-deterministic models, when the model is tested on both deterministic or non-deterministic evaluation metrics. The model is not possible to be able to both generate an accurate representation of the GT and diversify between the ten samples that are generated for each motion. As for the non-deterministic metrics we randomly select one of the samples of each motion to use for the calculation of LVE, MVE and FDD, so that we can make an honest comparison with the deterministic models' results. The Diversity metric proves the fact that those samples are not far off to each other which might not be a bad thing if we mainly need to use the model to generate results close to the GT. That exact same insight is also provided by the MEE and CE, the Coverage Error of FaceDiffuser seems to be just a fraction smaller than the LVE which is already achieving good results. While the Mean Estimate Error, again by being close to the LVE, provides us with more perspicuity concerning the non-determinism of those results. Interestingly, the results for those two models and their comparison with the LVE follow the same pattern for the two datasets, BIWI and Multiface. Following the same path we can identify, similarly to BIWI results, the opposite existing for CodeTalker-ND. As FaceDiffuser can generate accurate but not diverse results, we can see the capability of diverse results from CodeTalker-ND damaging the accuracy of the representation of the GT. With higher Diversity

from FaceDiffuser, the model seems to be able to cover a broader spectrum of motions with the disadvantage of some of them not being accurate. That can also be seen from the MEE and CE results of CodeTalker-ND which have a bigger absolute distance from the LVE, especially the Mean Estimate Error showing the extent of the bad representation among the samples.

Additionally, for the purpose of both analysis of the results as well as the concrete evidence of how the models perform we used the same methodology as in BIWI to showcase the ranking of the models. The ranking of the models for each metric is presented in Table 10.

Rank	**LVE**	**FDD**	**MVE**	**Diversity**	**MEE**	**CE**
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 10: Ranking results of the models on the Multiface dataset for each evaluation metric

The results concerning the rank that the models achieve for each metric show again some of the advantages and disadvantages that exist in each model. The use of Multiface from FaceDiffuser's authors results in really good results overall with it being mostly first and second in ranking. The effect of using MVE during testing is showcased again by FaceXHuBERT which produces the best results on FDD and even third on MVE, while being really close in numbers. Overall the models performed as expected since the two datasets look alike concerning the way the data are handled during training.

The analysis of the Multiface dataset shows more variation among models compared to the BIWI dataset. FaceDiffuser excels in Lip Vertex Error, performing much better than the others, demonstrating its strong capability in generating accurate lip movements. While it also leads in Mean Vertex Error, the difference is smaller, suggesting all models are fairly good at modelling the entire face. The Multiface dataset's lower upper face movement likely contributes to this improved MVE. FaceXHuBERT performs best in the FDD metric, which measures upper-face movement, highlighting its focus on full-face accuracy. This model's architecture and metrics have influenced in making better its performance. For the non-deterministic FaceDiffuser, metrics like Diversity, Mean Estimate Error and Coverage Error show limited variability in generated samples but strong alignment with the ground truth. This trade-off between accuracy and diversity is typical. Overall, the knowledge about the usability and insight provided from the metrics follows the same pattern as the previously analyzed BIWI dataset results.

3DMEAD dataset: The final dataset we decided to test the models on and calculate the metrics is the 3DMEAD dataset which could be described as unknown territory. To the best of our knowledge, no other training of those models has been reported by other researchers, mainly due to the recency of the publishing of this dataset. For that reason, and because of how different the dataset is from the others, we struggled to identify the correct process we should have followed for the training. We came to the conclusion that due to the main aspect of the comparison we want to make not being the models but the metrics, we are going to use the models exactly the same way as before. That means that, even if it would affect the results for some models, we decided to keep the published versions of them exactly the same doing only minimal changes. One example is that we could have added emotion control in the models to also affect the result based on the emotional category, but we decided that changing any of the models

would lead them to be different from the ones published. In that sense, it is important to point out that the only model that was implemented with emotion control is ProbTalk3D, which also shows the best results in all aspects. Since the model was not available at the start of our study, that step was done by the authors since it focused on training with 3DMEAD contrary to the rest of the models we used in our study. As the line between having a new adaptation of the model and improving it for the sake of that dataset is so thin, we decided to keep the rest of the models as plain as possible and complete only changes that do not affect the architecture or approach of the models. With that in mind, we trained the models the exact same way, gathered the results and calculated the metrics, which are visible in Table 7. The first observation that we can make is the existence of smaller absolute values for the metrics, compared with the previous two datasets. That is caused by the fact that 3DMEAD contains a lot more data as it provides the models with multiple different emotional categories and intensity levels for each motion. Firstly, the LVE metric seems to provide low values, especially for the FaceDiffuser and ProbTalk3D which score the best of all the models with similar results. We can already establish a pattern from the three datasets' results at this point, that the models showcase their advantages through the metrics. The clear view of the capabilities of the models that the metrics provide can be interpreted especially on the location they are focused, especially LVE, FDD and MVE. As we have already seen FaceDiffuser and FaceXHuBERT have been the best in the lip region and for the upper face FaceXHuBERT has had the best performance for the previous two datasets. In the case of 3DMEAD, FaceXHuBERT performs well in terms of FDD but it is surpassed by FaceFormer and ProbTalk3D. We can identify a similar case for the MVE, with all the models performing well and showcasing close absolute numbers as results. One important aspect that can be identified for all the deterministic metrics is the excellence of ProbTalk3D even if it is a non-deterministic model that does not only aim to achieve results close to the GT but also diverse results. Due to the implementation of emotion control and the two stages of training, the model is able to handle better the huge amount of data during training and improve the results for each emotion without any intermixing. Another insight that we get from this dataset's results is in terms of the metrics, as we understand the importance of using the metrics during the testing of the architecture selections. Using specific metrics to evaluate the model during the process of creating the architecture and the selection of specific parts would increase the capabilities of the models. That is mostly understood from the ablation studies that are presented from the models and with the insights and patterns that we identify from the results from datasets that were not explored by the original authors. Additionally, another interesting factor is the difference between the performance of CodeTalker in Multiface and 3DMEAD. In the two distance metrics, LVE and MVE, we can identify CodeTalker performing better for 3DMEAD in comparison with the other models, with the main reason that was identified during training, being the use of the motion prior that helped handle that immense amount of data 3DMEAD provides. That was not the case for Multiface as it performed worst in those metrics in comparison with the rest of the models, giving an additional insight towards the variety that exists even for the same model performance for different datasets. Overall, metric values like LVE, FDD, and MVE were generally lower compared to other datasets due to the extensive emotional categories in 3DMEAD. ProbTalk3D, which includes emotion control, showed the best results across all these metrics. FaceDiffuser and FaceXHuBERT performed well in specific regions, with FaceXHuBERT which excelled in FDD for the other two datasets being surpassed by FaceFormer and ProbTalk3D for this dataset. CodeTalker performed better on 3DMEAD, highlighting performance variations

across datasets. These findings underscore the importance of metrics in evaluating model performance. For a better understanding of how the models perform for the 3DMEAD dataset, we showcase a visualization with MVE and LVE on the Y and X axes respectively, in Figure 22.

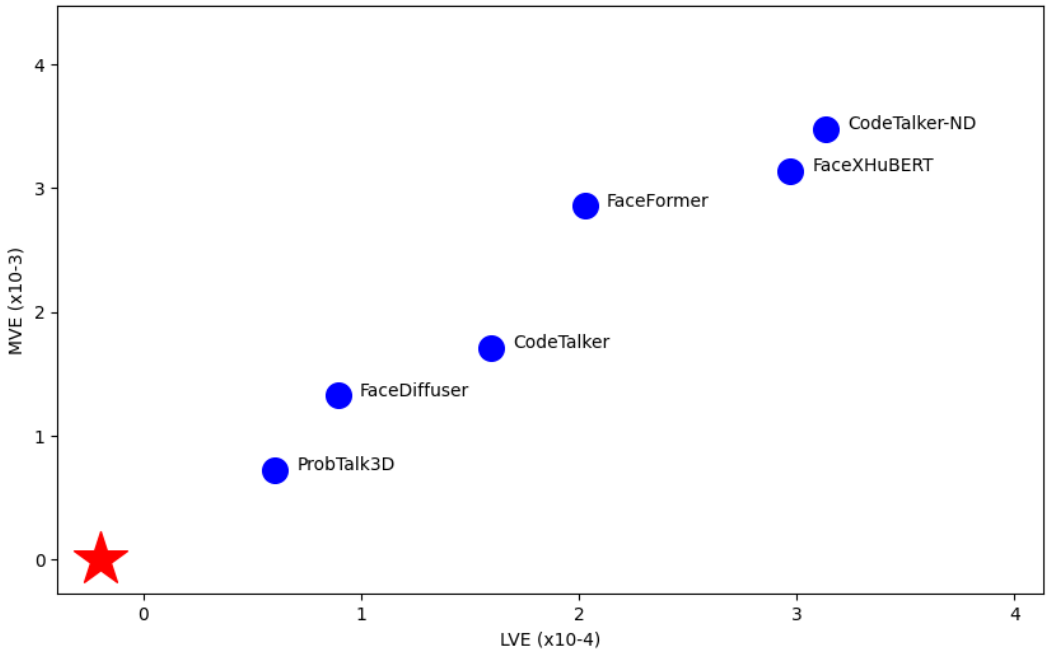


Figure 22: MVE and LVE results on the 3DMEAD dataset visuialized

In the case of non-deterministic metrics, ProbTalk3D excels in Diversity showing capabilities of different and accurate results. That advantage exists for this model due to the different way it handles the data with emotion control and the reasons priorly explained about its architecture. For the rest of the models for which we can make a direct comparison with the results that were gathered in the previous datasets, we identify the pattern of low Diversity for FaceDiffuser in our results for 3DMEAD and slightly higher for CodeTalker-ND. That is an expected result, especially for this dataset, which contains a huge variety of data for each motion for the models to train on, the expectation was that it would lead to low variety in results for the Diversity metric on the generated samples. Notably, as FaceDiffuser performs well in almost all the other metrics calculating the distance from the ground truth, we can provide the same argument about the deterministic and non-deterministic metrics not being able to mutually align in most cases. Similarly, CodeTalker-ND performs the worst in almost all of the deterministic metrics struggling to generate samples close to the GT but diverse at least as seen from the Diversity metric. Additionally, the two metrics that we previously used alongside the LVE to provide us with more information towards the non-determinism of the model present a matching tone for the expectations of the three non-deterministic models. The Mean Estimate Error of FaceDiffuser for 3DMEAD is again close in terms of absolute numbers to the LVE, which shows the low distance between the multiple generated samples of the model. At the same time, the Coverage Error, which provides the smallest distance that exists between the samples and the GT, shows again that it has a short distance with the randomly sampled LVE. For ProbTalker3D we get a smaller distance of MEE and CE from the LVE value, but that happens due to the low absolute values of those metrics. CodeTalker-ND shows the same behavior as before with the MEE and

CE being a little bit further to the LVE value in terms of absolute distance. Even if all the absolute values for all the metrics are low we can identify the same pattern we experienced in the previous datasets, which guides us to the same conclusions. Due to the model being able to generate an accurate representation of the GT for all the samples, it is not able to create a huge variance between them. To showcase the difference that exists for the Mean Estimate Error and the Coverage Error results, we gathered the data for FaceDiffuser and CodeTalker-ND among the three datasets, with two separate visualizations in Figure 23.

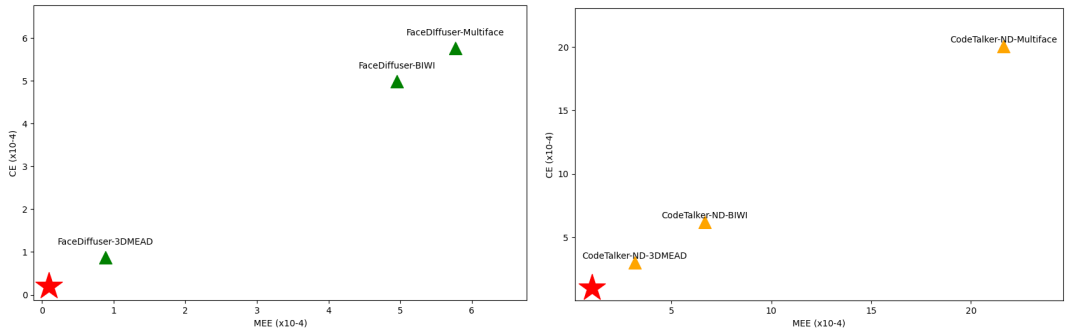


Figure 23: Mean Estimate Error and Coverage Error results for FaceDiffuser and CodeTalker-ND across all datasets, BIWI, Multiface and 3DMEAD

With both the visualization and after examining the numbers of the results, it becomes obvious that those two metrics perform a lot better for both models when training with 3DMEAD. As the obvious reason is the amount of data provided to the models, we can also add the assumption that FaceDiffuser performs well in synchronizing the lip movement of the generated results with the GT. Those metrics are just an addition to our insights about the general advantages that 3DMEAD as a dataset contains, which were identified in earlier stages as well. The amount of data for each motion as well as the general addition of more sequences and subjects provides the models with better tools to generate results closer to the ground truth they aim for. That can be seen across all models when the values of the distance errors are compared with the rest of the results for the rest of the metrics. An intriguing factor after evaluating all the results of the datasets, as well as how FaceDiffuser and CodeTalker-ND perform in terms of MEE and CE, is the huge gap that exists between the metrics absolute numbers for the three distance metrics for the GT. Even if we expected a difference the gap for all the models in some metrics is beyond our expectations and adds to the argumentation towards more diversity and an increase in the amount of subjects and sequences in the 3D face datasets. Summarizing the results for the non-deterministic metrics, ProbTalk3D excels in Diversity due to its unique handling of data with emotion control. For the other models, FaceDiffuser shows low Diversity, while CodeTalker-ND has slightly higher Diversity, reflecting the dataset's vast variety. FaceDiffuser performs well in deterministic metrics but less so in non-deterministic ones, illustrating the misalignment between these metric types. CodeTalker-ND struggles with deterministic metrics but shows diverse results. The Mean Estimate Error and Coverage Error for FaceDiffuser on 3DMEAD align closely with LVE, indicating low variance in generated samples. ProbTalk3D's MEE and CE are also close to LVE but with generally lower absolute values. CodeTalker-ND shows greater distance in MEE and CE from LVE, similar to its performance in other datasets. These patterns, consistent across datasets, highlight how models generating accurate representations of the Ground Truth have lower sample variance. Visualizations of MEE and CE reveal that

FaceDiffuser performs significantly better with 3DMEAD, likely due to the dataset's extensive data. This suggests that 3DMEAD's abundance of sequences and subjects helps models generate results closer to the GT. Comparing FDD across datasets further underscores these insights. Using a similar approach as before we visualized the differences existing for the models in terms of FDD for all the datasets in Figure 24. Finally, the same process will be followed presenting the

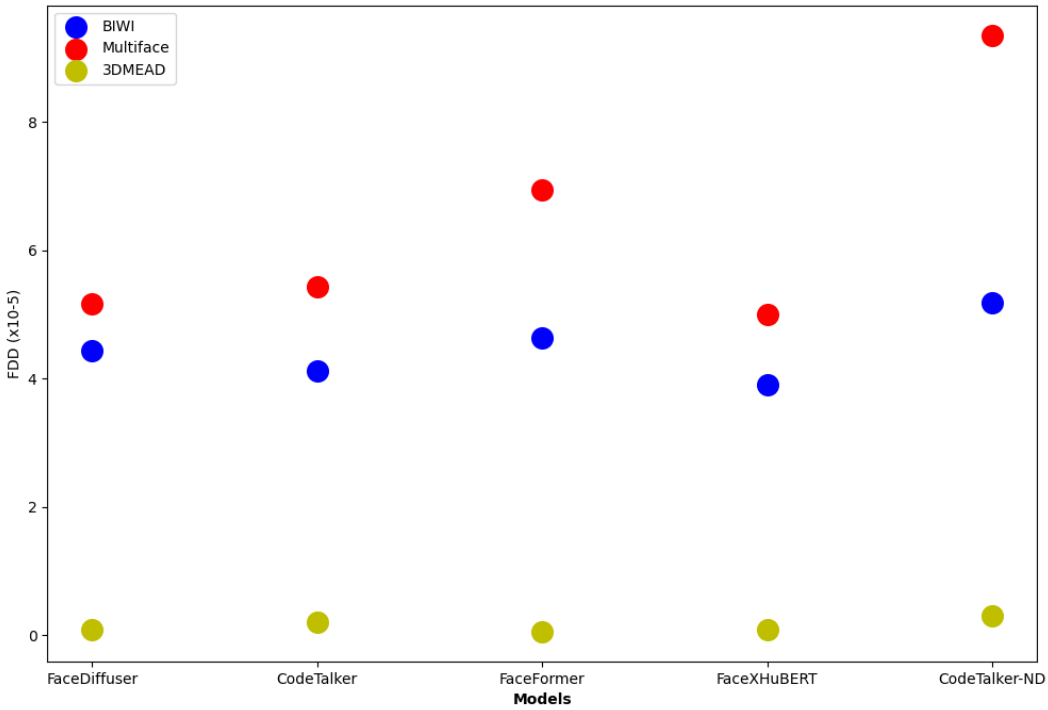


Figure 24: Facial Dynamics Deviation results for all the models across all datasets

ranking of the models on the 3DMEAD dataset with a difference this time. The ranking based only on the objective metrics is presented in Table 11, while we will later use those rankings to combine them with the results from the subjective evaluation that will be conducted.

Rank	**LVE**	**FDD**	**MVE**	**Diversity**	**MEE**	**CE**
1	ProbTalk3D	ProbTalk3D	ProbTalk3D	ProbTalk3D	ProbTalk3D	ProbTalk3D
2	FaceDiffuser	FaceFormer	FaceDiffuser	CodeTalker-ND	FaceDiffuser	FaceDiffuser
3	CodeTalker	FaceXHuBERT	CodeTalker	FaceDiffuser	CodeTalker-ND	CodeTalker-ND
4	FaceFormer	FaceDiffuser	FaceFormer	-	-	-
5	FaceXHuBERT	CodeTalker	FaceXHuBERT	-	-	-
6	CodeTalker-ND	CodeTalker-ND	CodeTalker-ND	-	-	-

Table 11: Ranking results of the models on the 3DMEAD dataset for each evaluation metric

The result of the ranking for this dataset provides us with a clearer picture of what has already been discussed and expected. ProbTalk3D excels in all the metrics for the reasons discussed earlier, while the rest of the models follow a route similar to the previous rankings. If we exclude ProbTalk3D which was trained on 3DMEAD, that dataset is the only one that none of the models were previously trained on by the authors which makes the results more honest. In comparison with the Multiface dataset, FaceDiffuser scores best in most metrics while being the only one that was trained by its authors on that dataset. 3DMEAD dataset provides clear outcomes on how the models can perform towards new data and how close they can represent the data they were

not originally trained to do. For the deterministic metrics, FaceDiffuser shows again resilience and performs second best in LVE and MVE, while we can identify the better positioning of CodeTalker on the rankings than the previous datasets, due to the two-stage handling of the large amount of data that this dataset contains. FaceXHuBeRT follows the same pattern with its best position being the one concerning the FDD metric. A similar case exists for the non-deterministic metrics while ProbTalk3D is the best in all categories and even if CodeTalker-ND performs second best for Diversity the deterministic results are not as good as FaceDiffuser's.

Overall Analysis: Analyzing the results from the BIWI, Multiface and 3DMEAD datasets reveals significant insights into the utility and behaviour of different evaluation metrics, as well as the patterns that emerge from the models' performances across these datasets. The Lip Vertex Error metric consistently shows small variations among models across all datasets, highlighting its effectiveness in assessing lip-sync accuracy. This consistency suggests that LVE is a reliable metric for evaluating how closely generated lip movements match the ground truth. However, the Mean Vertex Error, which evaluates the whole face region, generally presents higher values than LVE. This discrepancy indicates that achieving accurate full-face motion is more challenging than focusing solely on lip-sync, emphasizing the importance of including MVE in evaluations to capture the broader performance of facial models. The Facial Dynamics Deviation metric, which measures the movement difference of the upper face, also reveals interesting patterns. Models that used either FDD or MVE during their testing achieved regularly good results across all the datasets in comparison with models that did not use such techniques during the testing of the architecture. Non-deterministic metrics, including Diversity, Mean Estimate Error and Coverage Error, provide crucial insights into models' variability and their ability to generate different outputs for the same input. The Diversity metric highlights how varied the generated samples are, with low diversity indicating similar outputs across samples. This metric is particularly useful for understanding a model's potential to produce diverse results, which is essential for applications requiring variability. The MEE and CE metrics, on the other hand, offer a deeper understanding of the generated samples' consistency and proximity to the ground truth. Consistent MEE and CE values close to LVE across datasets suggest that models generate samples that are not only close to each other but also close to the ground truth, underscoring these metrics' importance in evaluating non-deterministic models. The impact of dataset characteristics on metric values is also evident. For instance, the BIWI dataset, with its focus on lip-sync, results in lower LVE values and minimal variation across models, demonstrating the dataset's alignment with lip-sync evaluation. In contrast, the Multiface dataset, with less upper-face movement, shows a broader range of LVE and MVE values, indicating that this dataset is more challenging for full-face motion generation. The 3DMEAD dataset, with its extensive emotional categories, generally produces lower metric values, highlighting the richness of the data and its ability to improve model performance. This dataset also underscores the importance of including diverse emotional expressions in training data to enhance models' generalizability and accuracy. Examining the models' performance patterns across these datasets provides additional insights. FaceDiffuser consistently stands out, when compared only with the models tested on all datasets, particularly excelling in the LVE metric, which indicates its strong capability to generate accurate lip movements. This model also shows robust performance in MVE, though its lower Diversity suggests limited variability in generated samples. FaceXHuBERT consistently delivers strong results, especially in the FDD, due to its architecture designed to accurately represent the entire face. CodeTalker-ND presents a contrasting behaviour, exhibiting higher Diversity but

struggling with deterministic metrics like LVE and MVE, indicating that its generated samples are varied but not always accurate representations of the ground truth. FaceFormer generally maintains a balanced performance without specializing in any particular metric consistently across datasets. The results that stand out are the ones of ProbTalk3D that were only gathered for the 3DMEAD dataset. The model performs the best across all metrics achieving both variety and accuracy which are the two main aims of a non-deterministic model. Due to the model's architecture which was implemented around the 3DMEAD dataset, there is a clear advantage in providing the best results from all the models.

Overall, these findings highlight the complementary nature of these metrics. While LVE and MVE provide insights into lip-sync and full-face accuracy, FDD offers additional perspectives on upper-face movements. Non-deterministic metrics such as Diversity, MEE, and CE reveal the extent of variability and consistency in generated samples.

6.3 Subjective Metrics Analysis

After completing the User Study that was conducted based on the experiment setting explained in Section 5 we gathered the results from the user's responses to the two questions. Using the results we calculated the mean across all the users for each model's motions for the two metrics that were set, as seen in Table 8. As we expected the results for both metrics are really close for the models, with not much deviation from the best to the last, while still recognizing models performing better. That can be explained based on two things, firstly the use of four motions for each emotional category, with the aim of straightening the curve of possible disadvantages or advantages a model might have for a specific motion, ended up bringing most results around the same small range. Also, the foreseen similarity between the results, which was obvious also from the objective metrics, is not something that most people can rate so easily. The subtle differences between the motions may not be identifiable from an untrained human eye to the extent that the differences are clearly showcased in the results. For the results themselves, we can identify that the Ground Truth stands first for both the Lip-Sync and the Realism, while for the rest of the models, there is intermixing in the performance. A model that performs consistently enough is the FaceXHuBERT which performs best from the models in both metrics, with quite a distance, especially regarding the Lip-Sync. Even if in some of the objective metrics for the 3DMEAD, FaceXHuBERT does not stand in the first places, it shows the best results regarding the subjective metrics, interestingly enough. Another model with a similar case is FaceFormer which has results that place it in third place, slightly behind FaceDiffuser, for Lip-Sync and second place regarding Realism, while it also does not showcase that efficiency with the objective metrics. On the opposite side, even if ProbTalk3D showcased amazing results for the objective metrics, performing best in all categories, it failed to keep that same position for the subjective metrics. That result is the one that is more interesting due to the extent of the good results of ProbTalk3D regarding the objective metrics, showing a clear inconsistency between the two types of metrics. Other models, such as FaceDiffuser follow a similar path with results around the ones of the objective metrics with slight changes in the positioning. That is also the case with both the deterministic and the non-deterministic versions of CodeTalker, which perform around the same range of positions for both the objective and the subjective metrics. Overall, from a first analysis of the results, the two models that stand out due to their results are ProbTalk3D and FaceXHuBERT. Each for completely contrasting reasons, with ProbTalk3D excelling in objective

metrics but lacking good results in the subjective metrics, while FaceXHuBERT does the complete opposite. To identify clearly the combination of results of the two metrics we visualized them in Figure 25.

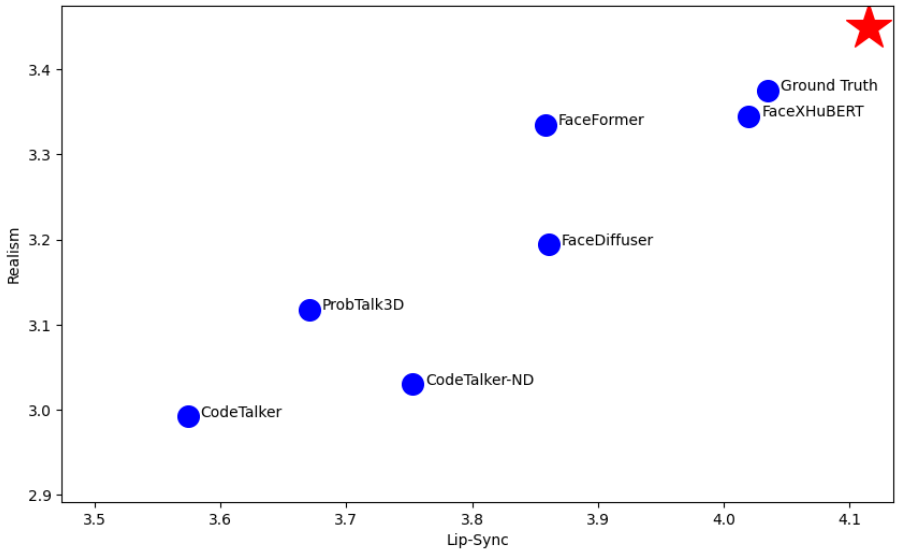


Figure 25: Subjective metrics results visualization with Realism on the Y-axis and Lip-Sync on the X-axis

The two subjective metrics are going to be used to generate a ranking for the models based on how well they performed on those metrics and then also compared with the objective metrics rankings. As for the difference between those two metrics that we have results on, we will utilize that difference to cover as much as possible for the comparison with the objective metrics. The main meaningful comparison will be the ranking on the Lip-Sync metric and the three objective metrics concerning the Lip area, LVE, CE and MEE. We will mainly use LVE as it is a direct representation of what was asked to rank but the addition of CE and MEE could prove to be helpful for our study. Additionally, the Realism metric will be used alongside MVE and FDD as they utilize the whole and upper face respectively, being more suited to showcase similar results concerning realistic movements. Concerning the Diversity metric, we will exclude it from the analysis concerning the comparison between subjective and objective metrics due to the constraint of it not being able to be rated by users. The clear ranking representation of the results for the objective metrics and the subjective metrics is presented in Table 12 and Table 13 in two separate tables according to the metrics involved, for that ranking we will exclude the Ground Truth as it is obviously not capable to be used for objective metrics.

Rank	**Lip-Sync**	**LVE**	**MEE**	**CE**
1	FaceXHuBERT	ProbTalk3D	ProbTalk3D	ProbTalk3D
2	FaceDiffuser	FaceDiffuser	FaceDiffuser	FaceDiffuser
3	FaceFormer	CodeTalker	CodeTalker-ND	CodeTalker-ND
4	CodeTalker-ND	FaceFormer	-	-
5	ProbTalk3D	FaceXHuBERT	-	-
6	CodeTalker	CodeTalker-ND	-	-

Table 12: Ranking results of the models on the 3DMEAD dataset for LVE, MEE, CE, and Lip-Sync metrics

The table showcases the full ranking for the models concerning the Lip area metrics both objective and subjective, with some interesting results, especially after the addition of the Lip-Sync subjective metric. Since the results for the MEE and CE metrics are only available for three models, we will focus more on the comparison of Lip-Sync with LVE but they would come in handy to understand how the multiple samples affect the result.

It becomes evident that most models perform differently for the LVE and the Lip-Sync metrics, and that there is no clear analogy of the rankings. Firstly, the clear example is the drop from first place of ProbTalk3D all the way to the fifth place for the Lip-Sync, suggesting a clear difference between the way users conceive the results. On the same note, users seemed to find the Lip-Sync of FaceXHuBERT the best contrasting the calculation of the distance with the Ground Truth. Especially for the lip area, we can suggest the existence of a complicated difference between the metrics. The calculation of the LVE is done by comparing the results with the Ground Truth, but that is not the case with Lip-Sync which asked the users to individually rate each motion without knowing which one is the GT. Even though Ground Truth achieved the best results, the users were not aware that this was the motion they were supposed to use as a reference to what is accurate. Even if ProbTalk3D achieved the best on the objective metrics, the users did not find it achieving what they perceived as the best, due to their lack of knowledge on what was the desired one. Even if at first those two facts seem to be mutually exclusive, they let us understand that the untrained human eye is not able to at least reproduce the same result as the calculations. The fact that the users seemed to identify the GT as the best one did not lead them to approach all the rest of the motions looking for similarity with it.

Furthermore, a similar case exists with CodeTalker which performs relatively well for the LVE landing in third place but dropped last by the users' rating of Lip-Sync. The issue with the lack of perception from the users as well as the really close distances between the mean results led to those differences that can be identified for this model as well. Another interesting observation can be made by comparing FaceDiffuser's performance. While it secures the second position across both the Lip-Sync and LVE metrics, this consistency highlights that its approach to lip movement generation is both technically accurate and perceptually convincing. This suggests that FaceDiffuser might be employing an effective strategy in capturing lip movements that align well with human visual expectations as well as objective measurements. However, the overall trend in these rankings indicates that a high objective metric score does not always translate to a high subjective score, emphasizing the complexity of human perception in evaluating lip-sync accuracy. Moreover, the rankings also show the impact of subjective perception on models like FaceFormer and CodeTalker-ND, which have differing positions across the metrics. FaceFormer ranks higher in Lip-Sync compared to its LVE score, suggesting that users might perceive its output as more natural despite the objective measurements indicating otherwise. Similarly, CodeTalker-ND shows a significant drop in Lip-Sync ranking compared to its relatively better performance in LVE, pointing towards possible deficiencies in capturing the nuances that users deem important for realistic lip-syncing.

These distinctions between subjective and objective evaluations highlight the inherent challenges in designing models that can satisfy both technical and perceptual benchmarks. They underscore the necessity of incorporating diverse evaluation metrics to obtain a holistic understanding of a model's performance and to identify areas that need improvement from both a technical and a user-experience perspective.

Rank	**Realism**	**MVE**	**FDD**
1	FaceXHuBERT	ProbTalk3D	ProbTalk3D
2	FaceFormer	FaceDiffuser	FaceFormer
3	FaceDiffuser	CodeTalker	FaceXHuBERT
4	ProbTalk3D	FaceFormer	FaceDiffuser
5	CodeTalker-ND	FaceXHuBERT	CodeTalker
6	CodeTalker	CodeTalker-ND	CodeTalker-ND

Table 13: Ranking results of the models on the 3DMEAD dataset for FDD, MVE, and Realism metrics

For the comparison of the second metric that we tested, Realism, we used the results of the objective evaluation of MVE and FDD. Both objective metrics are focused on the whole or the upper face which can be correlated with how realistic a motion is perceived to be. For the ranking on the Realism metric, we once again recognize FaceXHuBERT performing best, with a similar case with the Lip area metrics taking place. FaceXHuBERT was not recognized as the best for 3DMEAD for the rest of the metrics, with the FDD ranking being third and fifth for the MVE.

An interesting aspect that was previously identified is the fact that FaceXHuBERT performed best for the FDD metric for both BIWI and Multiface due to its architecture and testing methodology. Even if for the same metric that model did not excel, it shows that it was capable of generating the best results based on the perception of the user. That is one aspect that additionally affected the rating of FaceXHuBERT. Interestingly, the last positions on the ranking table for Realism are models that are placed in relatively low positions for the objective metrics as well, with CodeTalker-ND and CodeTalker which even if it achieves third place for MVE, its FDD results place it fifth. Due to the difference between those three metrics, no model keeps a specific place in all rankings, and that is also influenced by the small differences between the results. That means that even if a model is low on rankings the distance from the models above is really small absolute numbers since all the models achieved results in the range of acceptable ratings. That is also the case with FaceDiffuser as well, which does not rank in the same spot for all metrics but covers a range from second to fourth place.

A closer look at FaceFormer reveals its unique consistency across the metrics, securing the second position in both the Realism and MVE rankings, and fourth in FDD. This consistency suggests that FaceFormer has a balanced performance, producing realistic motions that align well with both user perceptions and objective measurements. The model's architecture likely contributes to its robustness, allowing it to adapt effectively to different datasets and evaluation metrics. Its strong performance in the Realism metric indicates that users perceive its outputs as highly lifelike, corroborating its objective metric rankings. FaceDiffuser's performance is notable for its variability across the metrics, ranking second in Realism, third in MVE, and fourth in FDD. This variability highlights the model's strengths and weaknesses in different aspects of motion generation. The high Realism score suggests that the model excels in producing visually convincing motions, which might be attributed to its nuanced approach to facial dynamics. However, its lower ranking in FDD implies that it may not capture certain subtle facial deformations as effectively as some other models. This discrepancy underscores the importance of using multiple metrics to gain a comprehensive understanding of a model's performance. The case of ProbTalk3D, on the other hand, shows a significant contrast between its subjective and objective evaluations. While it ranks fourth in Realism, it achieves the top spot in both MVE and FDD. This indicates that despite its technical accuracy in modelling facial movements, it may not fully

resonate with user perceptions of realism. This could be due to factors such as the naturalness of transitions or the expressiveness of the generated motions, which might not be fully captured by the objective metrics.

Overall, these comparisons highlight the complex relationship between subjective user perceptions and objective metrics. The varied performance of the models across Realism, MVE and FDD metrics demonstrates that no single model excels universally. FaceXHuBERT stands out for its user-perceived realism despite not leading in all objective metrics, while FaceFormer and FaceDiffuser show balanced yet variable performances and ProbTalk3D excels technically but lags in subjective evaluation. This underscores the necessity of a multi-faceted evaluation approach to fully capture the efficacy of these models in generating realistic facial motions.

Overall Analysis: The results, summarized in Table 8 and compared with Objective Metrics rankings in Tables 12 and 13, show that the subjective evaluations of Lip-Sync and Realism are close across different models, with Ground Truth consistently performing best. FaceXHuBERT also stands out, showing strong results in subjective metrics despite not always excelling in objective metrics. Contrariwise, ProbTalk3D, which excels in objective metrics, did not perform as well subjectively.

The close similarity in results can be attributed to the use of multiple motions per emotional category, which balanced out individual model advantages. The differences in user perception, especially regarding Lip-Sync, highlight that untrained users may not easily identify subtle differences that objective metrics can measure. This discrepancy is evident with models like FaceFormer and CodeTalker, which show significant variations between objective and subjective rankings. The findings indicate a notable inconsistency between objective and subjective metrics, particularly for models like ProbTalk3D and FaceXHuBERT. While ProbTalk3D excels objectively, it ranks lower subjectively, and FaceXHuBERT shows the opposite trend. These results underline the complexity of human perception and the challenges in aligning subjective impressions with objective measures. Overall, this analysis reveals that user perceptions can significantly differ from calculated metrics, emphasizing the importance of considering both perspectives in model evaluations.

7 Discussion, Limitations and Future Work

7.1 Disussion

Our research aimed to examine the usage of evaluation metrics in 3D facial animation datasets and understand the nature of facial animations. This exploration involved several key stages, including analysing datasets, training models and calculating multiple different evaluation metrics, both objective and subjective. The goal was to identify the evolution and specific needs of 3D facial animation.

Creating a robust inventory of evaluation metrics was a critical part of our research. This inventory aimed to provide a comprehensive view of all metrics across different models and datasets, allowing for a consistent and holistic evaluation. The primary metrics used included Lip Vertex Error (LVE), Mean Vertex Error (MVE), Facial Dynamics Deviation (FDD), Diversity, Coverage Error (CE), and Mean Estimate Error (MEE). Each of these metrics was chosen to capture various aspects of the models' performance, from deterministic to non-deterministic behaviours.

Selecting appropriate models for our experiments was crucial to ensure a diverse range of architectural characteristics and performance claims. We included both deterministic and non-deterministic models to provide a broad spectrum of performance insights. The selected models were FaceDiffuser, FaceXHuBERT, CodeTalker, CodeTalker-ND, FaceFormer and ProbTalk3D. Each model was chosen for its state-of-the-art capabilities in either deterministic or non-deterministic facial animation. To maintain consistency in our experiments, all models were trained using the same settings and data splits. This included using the exact same server for training and following specific data splits for each dataset. The training process was carefully controlled to ensure that the results were comparable across different models and datasets.

The BIWI dataset provided a straightforward baseline since all models had been previously tested on this dataset. The results highlighted the performance of each model across the primary metrics. FaceDiffuser and FaceXHuBERT performed exceptionally well in LVE and FDD, indicating their strong capability to generate accurate lip movements and upper face dynamics. The Multiface dataset presented a more varied challenge due to its different structure. FaceDiffuser again showed strong performance, particularly in LVE, highlighting its ability to generate precise lip movements. However, CodeTalker-ND demonstrated a higher diversity in generated samples, although with less accuracy. The 3DMEAD dataset, being relatively new and complex, provided unique insights. The results showed that ProbTalk3D excelled in all metrics due to its architecture being specifically designed around emotion control. This model demonstrated both high accuracy and diversity, which are crucial for non-deterministic facial animation.

Across all datasets, deterministic metrics such as LVE, MVE, and FDD provided consistent measures of the models' performance. FaceDiffuser consistently excelled in LVE, indicating its strong capability to generate accurate lip movements. FaceXHuBERT showed robust performance in FDD, reflecting its focus on full-face accuracy. Non-deterministic metrics, including Diversity, MEE, and CE, revealed the variability and consistency of the generated samples. ProbTalk3D, with its emotion control mechanism, showed the highest diversity, demonstrating its capability to generate varied and accurate results. FaceDiffuser, while performing well in deterministic metrics, showed lower diversity, highlighting the trade-off between accuracy and variability in non-deterministic models.

The impact of dataset characteristics on metric values was evident. The BIWI dataset, with its focus on lip-sync, resulted in lower LVE values and minimal variation across models, demonstrating its alignment with lip-sync evaluation. In contrast, the Multiface dataset, with less upper-face movement, showed a broader range of LVE and MVE values, indicating that this dataset is more challenging for full-face motion generation. The 3DMEAD dataset, with its extensive emotional categories, generally produced lower metric values, highlighting the richness of the data and its ability to improve model performance. This dataset also underscored the importance of including diverse emotional expressions in training data to enhance models' generalizability and accuracy.

Examining the models' performance patterns across these datasets provided additional insights. FaceDiffuser consistently stood out, particularly excelling in the LVE metric, which indicated its strong capability to generate accurate lip movements. This model also showed robust performance in MVE, though its lower Diversity suggested limited variability in generated samples. FaceXHuBERT consistently delivered strong results, especially in FDD, due to its architecture designed to accurately represent the entire face. CodeTalker-ND presented a contrasting behaviour, exhibiting higher Diversity but struggling with deterministic metrics like LVE and

MVE, indicating that its generated samples are varied but not always accurate representations of the ground truth. FaceFormer generally maintained a balanced performance without consistently specializing in any particular metric across datasets. The results that stand out are those of ProbTalk3D that were only gathered for the 3DMEAD dataset. The model performed the best across all metrics, achieving both variety and accuracy, which are the two main aims of a non-deterministic model. Due to the model's architecture, which was implemented around the 3DMEAD dataset, there is a clear advantage in providing the best results from all the models.

We utilized several datasets for our experiments concerning the nature of 3D facial animations, each with unique characteristics that provided a comprehensive understanding of the field. For the 3D facial animation datasets, we focused on BIWI, Multiface, BEAT and 3DMEAD, which we tested using methods such as PCA and t-SNE. The results provided us with information about the existence of non-determinism concerning subjects varying for each dataset. The possible existence of emotionally non-determinism was also analysed focusing on the BEAT and 3DMEAD datasets. Additionally, for the comparison with other research fields such as the 2D talking head, we compared with datasets including GRID, CREMA-D and TCD, while for the body motion datasets, we selected KIT-ML, BABEL and IDEA 400 for body motion data. These datasets varied significantly in terms of subjects, sequences, hours of data, and frames, which provided a broad spectrum of data for analysis. These datasets were selected based on their extensive coverage of actions, sequences, and frames, making them ideal for examining the detailed nuances of facial animations and body motions.

To extend our understanding of the evaluation of the models we used the trained models to generate results that were rated across realism and lip-sync during a User Study we conducted. The results of this User study provided us with contradictory information than the objective metrics with most of the models not performing the same for subjective and objective evaluation. After examining the results and the reasons behind that inconsistency we understood the differences those two methods aim to achieve. The contrasting results stated the fact that perception studies and subjective evaluation from inexperienced users do not cover the spectrum of accuracy as well as multiple objective metrics, even though their addition is useful. The addition of subjective metrics provides us with what users seem to identify as the best motion and not the most accurate representation of GT, which is what the objective metrics usually do.

These findings highlight the complementary nature of these metrics. While LVE and MVE provide insights into lip-sync and full-face accuracy, FDD offers additional perspectives on upper-face movements. Non-deterministic metrics such as Diversity, MEE, and CE reveal the extent of variability and consistency in generated samples. This comprehensive analysis underscores the importance of a diverse and robust set of evaluation metrics in assessing the performance of facial animation models across different datasets.

7.2 Limitations

During this study, we came across multiple limitations, both some that exist in the research field and some that were due to the progress of our study. Starting with limitations that exist in the research field, firstly we have to acknowledge the fact that for our first Research Question, our evidence would be more concrete if we had a dataset that contains the same motion performed multiple times. To the best of our knowledge, no dataset that contains the same subject performing the same sentence, with the same emotion category exists. That led us to use a replacement

of this type of dataset, using the recent state-of-the-art datasets in different ways. An addition of this kind could be proven really useful to a more straightforward approach in answering the question regarding the nature of facial animations. Additionally, as seen from the results concerning the objective metrics and the comparison of the datasets with the ones from other fields, it would be proven a good addition if a similar dataset such as 3DMEAD existed. The main effect of this would be another big dataset with diversity across the speakers and the motions, as this helps the model train better for more accurate results. On that note, a similar dataset in terms of emotion categorization that follows a strict consistent pattern towards filling each emotion with the same amount of sentences would help us add to our search about the existence of emotion non-determinism on the ground truth. BEAT was tested but proved to be an inconsistent dataset, and its results are not meaningful towards our claim because of that reason.

Furthermore, as the study started one of the problems that we intercepted was the lack of published non-deterministic models for 3D facial animations. Many models have recently been proposed, but during our study, either their codebase was not provided or they were not accepted, while those tests would prove to be helpful in our case. We tackled this problem by manipulating CodeTalker-ND to work as a non-deterministic model even though it was not tested in this way before. A direct comparison of FaceDiffuser with recent state-of-the-art models, in the way that we did in 3DMEAD, would help us point out the differences between these types of models and how they affect the metrics. Due to the existing models, we focus our approach around the deterministic models, especially for the tests on BIWI and Multiface, which covers this spectrum but adding on top of that some more non-deterministic models would also provide useful insights. Nonetheless, we also have to acknowledge that due to the time restriction, we were not able to compile an even bigger inventory of metrics using more models or more datasets. The training of most of the models was a timely process, which led us to minimize them in numbers to be as effective as possible. That time restriction that affected the training of the models also affected the metrics that were calculated. We aimed to also explore the possibility of changing a metric or adapting some metrics from other research fields with the aim of creating something new for 3D facial animations. That process was only conducted around the Diversity metric used in FaceDiffuser for which we explored its disadvantages and decided to use the Diversity proposed in 3Diface which was derived from a body motion paper. Even if we conducted multiple studies around that metric, using slightly different calculations on different aspects, in the end, we decided to use the same approach that has already been proposed to avoid any conflict with the results. Our aim was to conduct a similar methodology around the other metrics as well, aiming to achieve more accurate results and expand the inventory. In the end, that expansion was not something that could provide meaningful insight if we only slightly altered the way the metrics were calculated.

In addition to the aforementioned limitations, the task of performing a perception study utilizing paid users or volunteers does not provide a holistic and professional viewpoint for the subjective metrics. The subjective nature of some evaluation metrics poses a challenge to a more accurate evaluation. While objective metrics like LVE, MVE and FDD provide quantifiable measures of model performance, subjective assessments of animation quality and emotional expression can vary significantly between evaluators. To address this, future studies could benefit from incorporating more standardized subjective evaluation protocols or utilizing professionals in this research field to rate the videos. Another limitation was the generalizability of our findings across different facial animation tasks. While our study focused on specific datasets and

models, the performance and applicability of these models in other contexts, such as real-time applications or different cultural expressions, remain uncertain. Expanding the scope of datasets to include more diverse cultural and linguistic backgrounds could provide more comprehensive insights into the global applicability of these models.

In conclusion, the limitations of this study stemmed from both the broader research field and specific constraints during our study. The absence of an ideal dataset and the shortage of published non-deterministic models were significant challenges. Additionally, time restrictions affected the number of models and metrics we could explore. Despite these challenges, we adapted our approach to maximize the effectiveness of our study and provide valuable insights into 3D facial animations. Addressing these limitations in future research could lead to more robust and comprehensive advancements in this field.

7.3 Future Work

Based on the limitations and insights gained from this study, several directions can be pursued to improve our understanding and evaluation of 3D facial animations. A primary limitation identified was the absence of datasets containing repeated performances of the same motions by the same subjects. Developing such datasets in the future would provide concrete evidence for understanding the nature of facial animations. Creating a dataset where the same sentences are performed multiple times with consistent emotion categories would significantly improve the ability to analyze both deterministic and non-deterministic aspects of facial animations.

Additionally, incorporating datasets similar to 3DMEAD, which offer extensive diversity across speakers and motions, would be beneficial. Ensuring a consistent and balanced representation of different emotions in future datasets would enhance model training, leading to more accurate and generalizable results. Addressing the inconsistency found in the BEAT dataset by designing rigorous datasets that evenly represent each emotion and sentence category would also be advantageous.

Furthermore, the lack of published non-deterministic models for 3D facial animations posed a challenge in this study. Integrating recent non-deterministic models and comparing their performance with models like FaceDiffuser and ProbTalk3D across various datasets, including 3DMEAD, would clarify differences in performance metrics and their implications on facial animation quality. Expanding the inventory of evaluation metrics remains a crucial area for future work. Developing new metrics or adapting metrics from other research fields to better capture the aspects of 3D facial animations would be beneficial. For instance, exploring the adaptation of metrics used in body motion analysis could yield valuable insights. Investigating the potential of modifying existing metrics to enhance their applicability to 3D facial animations could involve creating new variations of metrics like Diversity, tailored to the specific challenges of facial animation evaluation.

Another area identified for advancement is the usage of professionals in facial animations as participants in perception studies. Their professional viewpoint would provide results that are more accurate than those obtained from random, inexperienced participants. Utilizing professional participants would reduce the number of required responses, as they can more easily identify differences and rate videos more accurately. Also, to improve the generalizability of findings, expanding the scope of datasets to include more diverse cultural and linguistic backgrounds would provide a comprehensive understanding of the applicability of facial animation

models across different contexts. Exploring the performance of these models in real-time applications and other facial animation tasks would provide valuable insights into their practical utility.

Addressing these directions in future work can overcome the limitations identified in this study and contribute to more robust and comprehensive advancements in 3D facial animation.

8 Conclusion

Our study has explored the usage of evaluation metrics in 3D facial animation datasets, aiming to understand the nature of facial animations through comprehensive analysis and model training. By developing a robust inventory of evaluation metrics, including Lip Vertex Error (LVE), Mean Vertex Error (MVE), Facial Dynamics Deviation (FDD), Diversity, Coverage Error (CE), and Mean Estimate Error (MEE), we provided a consistent framework for assessing model performance across multiple datasets. Our analysis encompassed both deterministic and non-deterministic models, ensuring a diverse range of architectural characteristics and performance claims were examined. Concluding our study we aim to provide complete and comprehensive answers for our Research Questions set in Section 1.

Is facial animation deterministic or non-deterministic by nature? How can we measure that using existing 2D and 3D facial motion datasets? Our exploration of the nature of facial animations provided significant insights into the non-deterministic characteristics of facial animations. Through various analyses, including dimensionality reduction techniques like t-SNE and PCA, we examined the effects of different factors such as subject individuality and emotional expression on facial animation sequences. The 3DMEAD dataset, with its extensive emotional categories, revealed clear patterns in how emotions influence jaw movements, indicating that while there are common trends, individual expressions vary greatly. By analyzing the mean jaw movement across different emotions and subjects, we observed substantial variability, underscoring the impact of individual speaking styles on animation sequences. This was further supported by t-SNE and PCA visualizations, which demonstrated distinct clustering based on subjects, suggesting a strong subject-specific influence on the animations. Similar analyses on the BIWI and Multiface datasets, despite their lack of emotional categorization, supported these findings. The consistency in subject-specific clustering across these datasets highlighted the non-deterministic nature of facial animations, driven primarily by individual differences.

What are the evaluation metrics to assess the quality of generated 3D facial animations? What are the advantages and limitations of these metrics? How can we test them using state-of-the-art approaches to speech-driven 3D facial animation synthesis? The impact of dataset characteristics on metric values was evident. The BIWI dataset aligned well with lip-sync evaluation, resulting in lower LVE values and minimal variation across models. The Multiface dataset, with less upper-face movement, posed a greater challenge for full-face motion generation, as reflected in the broader range of LVE and MVE values. The 3DMEAD dataset, with its extensive emotional categories, highlighted the richness of the data and its ability to improve model performance, emphasizing the importance of diverse emotional expressions in training data. Examining the models' performance patterns across these datasets provided additional insights. FaceDiffuser consistently excelled in LVE, indicating its strong capability to generate accurate lip movements, while FaceXHuBERT delivered robust results in FDD due to its architecture designed for full-face accuracy. CodeTalker-ND showed higher Diversity but struggled

with deterministic metrics, highlighting the trade-off between accuracy and variability in non-deterministic models. ProbTalk3D, tested only on the 3DMEAD dataset, demonstrated the best performance across all metrics, achieving both variety and accuracy, the main aims of a non-deterministic model. Non-deterministic metrics such as Diversity, MEE, and CE were crucial in understanding models' variability and their ability to generate different outputs for the same input. These metrics revealed the extent of variability and consistency in generated samples, emphasizing the importance of a diverse and robust set of evaluation metrics. Additionally, the results from the BIWI, Multiface and 3DMEAD datasets revealed significant insights into the utility and behaviour of different evaluation metrics. LVE consistently showed small variations among models, underscoring its effectiveness in lip-sync accuracy. In contrast, MVE presented higher values, highlighting the challenges of achieving accurate full-face motion. FDD provided valuable insights into upper-face movements, with models that used FDD or MVE during testing achieving consistently good results.

Are the current quantitative and qualitative evaluation metrics consistent with each other? How about different metrics in the same category? What are the different metrics that are meaningful for deterministic and non-deterministic approaches? The Subjective Evaluation done using a perception study provided results that allow us to understand the importance of both the Subjective Metrics as well as the Objective metrics. The existence of some contradictions between the two types of metrics showcases the usefulness of taking into account both methodologies to evaluate a model accurately. Also, the different approaches that each of those methods requires point out further improvements in the field. The small deviation of the results for each of the models for the subjective metrics identifies the possible improvement that could be done by using professionals in the field and changing the whole process to make the differences between the models clearer. Although the subjective and objective metrics do not directly evaluate the same parameters, there could still be a useful connection between them based on the area they cover. Furthermore, the existence of a correlation between some objective metrics and the subjective metrics can be identified in which of the models used the objective metrics during training, a prime example is FaceXHuBERT's excellence in Realism, while using MVE as a metric to test the architecture.

In conclusion, this study has provided a detailed analysis of evaluation metrics and model performance in 3D facial animations. Addressing the identified limitations and pursuing the suggested directions for future work can lead to more robust and comprehensive advancements in this field, ultimately improving the accuracy, variability, and applicability of 3D facial animation models.

References

[1] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems* (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., pp. 12449–12460.

[2] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing 5*, 4 (2014), 377–390.

[3] Chen, L., Cui, G., Kou, Z., Zheng, H., and Xu, C. What comprises a good talking-head video generation?: A survey and benchmark.

[4] Cheng, Y., Yang, B., Wang, B., and Tan, R. T. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 07 (Apr. 2020), 10631–10638.

[5] Cooke, M., Barker, J., Cunningham, S., and Shao, X. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America 120*, 5 (11 2006), 2421–2424.

[6] Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*, 9 (Sept. 2023), 10850–10869.

[7] Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., and Black, M. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 10101–10111.

[8] Danecek, R., Black, M. J., and Bolkart, T. Emoca: Emotion driven monocular face capture and animation.

[9] Daněček, R., Chhatre, K., Tripathi, S., Wen, Y., Black, M., and Bolkart, T. Emotional speech-driven animation with content-emotion disentanglement. ACM.

[10] Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 10 (Oct. 2022), 5962–5979.

[11] Fan, Y., Lin, Z., Saito, J., Wang, W., and Komura, T. Faceformer: Speech-driven 3d facial animation with transformers. *CVPR 2022* (2022).

[12] Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., and Van Gool, L. A 3-d audio-visual corpus of affective communication. *Multimedia, IEEE Transactions on 12* (11 2010), 591 – 598.

[13] Fréchet, M. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940) 22* (1906), 1–72.

[14] Han, T., Gui, S., Huang, Y., Li, B., Liu, L., Zhou, B., Jiang, N., Lu, Q., Zhi, R., Liang, Y., Zhang, D., and Wan, J. Pmmtalk: Speech-driven 3d facial animation from complementary pseudo multi-modal features. *Submitted* (2023).

[15] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., and Ng, A. Y. Deep speech: Scaling up end-to-end speech recognition.

[16] Haque, K. I., and Yumak, Z. Facexhubert: Text-less speech-driven e(x)pressive 3d facial animation synthesis using self-supervised speech representation learning. *ICMI '23* (2023).

[17] Harte, N., and Gillen, E. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia 17* (2015), 603–615.

[18] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc. 29* (oct 2021), 3451–3460.

[19] Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., and Cao, X. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. *SIGGRAPH Conference Proceedings 2022* (2022).

[20] Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *SIGGRAPH 2017 36*, 4 (2017).

[21] Li, X., Zhang, J., and Liu, Y. Speech driven facial animation generation based on gan. *Displays 74* (2022), 102260.

[22] Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., and Zhang, L. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems* (2023).

[23] Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., and Zheng, B. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis.

[24] Lu, Y., Chai, J., and Cao, X. Live speech portraits: Real-time photorealistic talking-head animation. *ACM Trans. Graph. 40*, 6 (dec 2021).

[25] Minka, T. Automatic choice of dimensionality for pca. In *Advances in Neural Information Processing Systems* (2000), T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13, MIT Press.

[26] Mourot, L., Hoyet, L., Le Clerc, F., Schnitzler, F., and Hellier, P. A survey on deep learning for skeleton-based human animation. *Computer Graphics Forum 41*, 1 (Nov. 2021), 122–157.

[27] Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G. E., and Neff, M. A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum 42*, 2 (may 2023), 569–596.

[28] Park, S. J., Hong, J., Kim, M., and Ro, Y. M. Df-3dface: One-to-many speech synchronized 3d face animation with diffusion.

[29] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*, 85 (2011), 2825–2830.

[30] Peng, Z., Wu, H., Song, Z., Xu, H., Zhu, X., He, J., Liu, H., and Fan, Z. Emotalk: Speech-driven emotional disentanglement for 3d face animation. *ICCV 2023* (2023).

[31] Pham, H. X., Wang, Y., and Pavlovic, V. End-to-end learning for 3d facial animation from speech. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (New York, NY, USA, 2018), ICMI '18, Association for Computing Machinery, p. 361–365.

[32] Plappert, M., Mandery, C., and Asfour, T. The KIT motion-language dataset. *Big Data 4*, 4 (dec 2016), 236–252.

[33] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., and Jawahar, C. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia* (Oct. 2020), MM '20, ACM.

[34] Punnakkal, A. R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., and Black, M. J. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 722–731.

[35] Ren, Z., Pan, Z., Zhou, X., and Kang, L. Diffusion motion: Generate text-guided 3d human motion by diffusion model.

[36] Richard, A., Zollhoefer, M., Wen, Y., de la Torre, F., and Sheikh, Y. Meshtalk: 3d face animation from speech using cross-modality disentanglement. *ICCV 2022* (2022).

[37] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning* (Lille, France, 07–09 Jul 2015), F. Bach and D. Blei, Eds., vol. 37 of *Proceedings of Machine Learning Research*, PMLR, pp. 2256–2265.

[38] Stan, S., Haque, K. I., and Yumak, Z. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion.

[39] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., and Bermano, A. H. Human motion diffusion model. *ICLR2023* (2022).

[40] Thambiraja, B., Aliakbarian, S., Cosker, D., and Thies, J. 3diface: Diffusion-based speech-driven 3d facial animation and editing. *Submitted* (2023).

[41] Thambiraja, B., Habibie, I., Aliakbarian, S., Cosker, D., Theobalt, C., and Thies, J. Imitator: Personalized speech-driven 3d facial animation. *Submitted* (2022).

[42] Tipping, M. E., and Bishop, C. M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology 61*, 3 (01 2002), 611–622.

[43] Toshpulatov, M., Lee, W., and Lee, S. Talking human face generation: A survey. *Expert Systems with Applications 219* (2023), 119678.

[44] van der Maaten, L., and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research 9*, 86 (2008), 2579–2605.

[45] Villanueva Aylagas, M., Anadon Leon, H., Teye, M., and Tollmar, K. Voice2Face: Audio-driven Facial and Tongue Rig Animations with cVAEs. *Computer Graphics Forum* (2022).

[46] Vougioukas, K., Petridis, S., and Pantic, M. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision 128* (2019), 1398 – 1413.

[47] Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., and Shao, L. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding 210* (2021), 103225.

[48] Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV* (August 2020).

[49] Wolfert, P., Robinson, N., and Belpaeme, T. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems 52*, 3 (June 2022), 379–389.

[50] Wu, H., Jia, J., Xing, J., Xu, H., Wang, X., and Wang, J. Mmface4d: A large-scale multi-modal 4d face dataset for audio-driven 3d face animation.

[51] Wu, R., Yu, Y., Zhan, F., Zhang, J., Zhang, X., and Lu, S. Audio-driven talking face generation with diverse yet realistic facial animations.

[52] Wu, S., Haque, K. I., and Yumak, Z. Probtalk3d.

[53] Wu, X., Zhang, Q., Wu, Y., Wang, H., Li, S., Sun, L., and Li, X. F^{3}a-gan: Facial flow for face animation with generative adversarial networks. *IEEE Transactions on Image Processing 30* (2021), 8658–8670.

[54] Wuu, C.-h., Zheng, N., Ardisson, S., Bali, R., Belko, D., Brockmeyer, E., Evans, L., Godisart, T., Ha, H., Huang, X., Hypes, A., Koska, T., Krenn, S., Lombardi, S., Luo, X., McPhail, K., Millerschoen, L., Perdoch, M., Pitts, M., Richard, A., Saragih, J., Saragih, J., Shiratori, T., Simon, T., Stewart, M., Trimble, A., Weng, X., Whitewolf, D., Wu, C., Yu, S.-I., and Sheikh, Y. Multiface: A dataset for neural face rendering. In *arXiv* (2022).

[55] Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., and Wong, T.-T. Codetalker: Speech-driven 3d facial animation with discrete motion prior.

[56] Yang, K. D., Ranjan, A., Chang, J.-H. R., Vemulapalli, R., and Tuzel, O. Probabilistic speech-driven 3d facial motion synthesis: New benchmarks, methods, and applications. *Submitted* (2023).

[57] Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., and Wang, X. 3d human pose estimation in the wild by adversarial learning.

[58] Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., and Wang, F. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation.

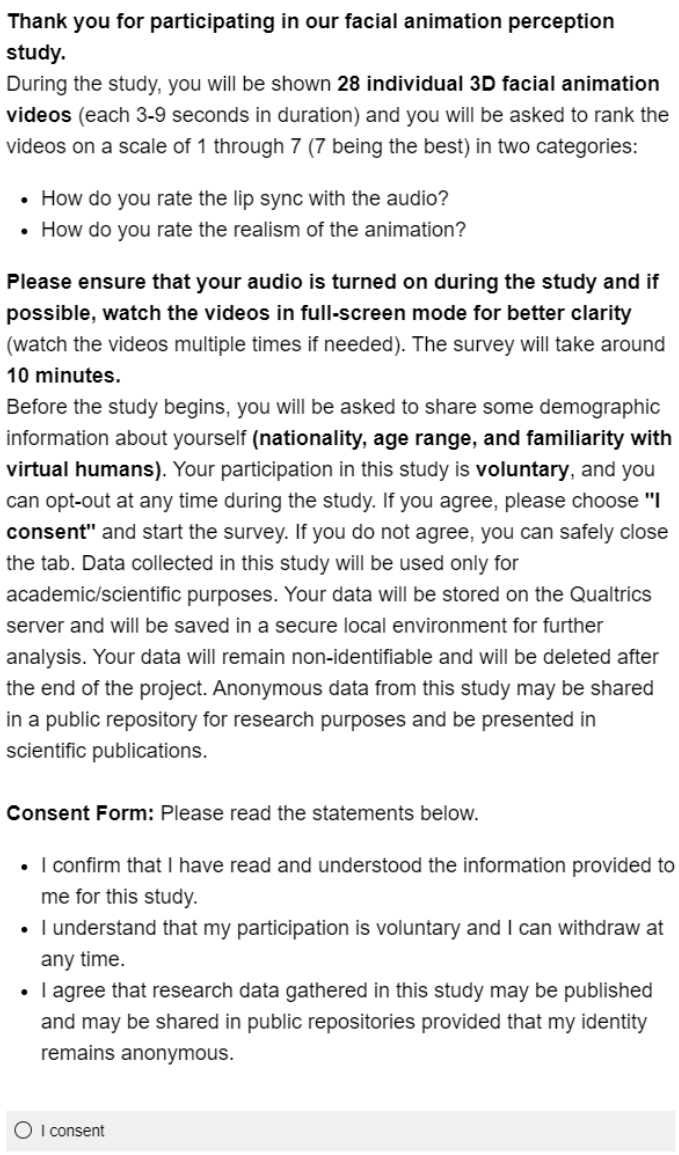
[59] Zhao, W., Wang, Y., He, T., Yin, L., Lin, J., and Jin, X. Breathing Life into Faces: Speech-driven 3D Facial Animation with Natural Head Pose and Detailed Shape. *arXiv e-prints* (Oct. 2023), arXiv:2310.20240.

[60] Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., and Ding, Z. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2021), pp. 11656–11665.

Appendix

8.1 User Study

The process of evaluating the models subjectively required the creation of a survey that contained motions across all models that would be rated by the users. The motions were rendered using the results and predictions for the models and then we started forming our study. Firstly, the instruction page contained the required information for the user to complete the study and the consent form that the user needed to accept to proceed, as shown in Image 26.



Thank you for participating in our facial animation perception study.
During the study, you will be shown **28 individual 3D facial animation videos** (each 3-9 seconds in duration) and you will be asked to rank the videos on a scale of 1 through 7 (7 being the best) in two categories:

- How do you rate the lip sync with the audio?
- How do you rate the realism of the animation?

Please ensure that your audio is turned on during the study and if possible, watch the videos in full-screen mode for better clarity (watch the videos multiple times if needed). The survey will take around **10 minutes.**
Before the study begins, you will be asked to share some demographic information about yourself **(nationality, age range, and familiarity with virtual humans)**. Your participation in this study is **voluntary**, and you can opt-out at any time during the study. If you agree, please choose **"I consent"** and start the survey. If you do not agree, you can safely close the tab. Data collected in this study will be used only for academic/scientific purposes. Your data will be stored on the Qualtrics server and will be saved in a secure local environment for further analysis. Your data will remain non-identifiable and will be deleted after the end of the project. Anonymous data from this study may be shared in a public repository for research purposes and be presented in scientific publications.

Consent Form: Please read the statements below.

- I confirm that I have read and understood the information provided to me for this study.
- I understand that my participation is voluntary and I can withdraw at any time.
- I agree that research data gathered in this study may be published and may be shared in public repositories provided that my identity remains anonymous.

○ I consent

Figure 26: User Study welcome screen and consent form

The complete overview of motion randomization and an example of how a user would randomly be assigned to watch the motions of four emotions for all the models is showcased in Figure 27.

Emotions:	FaceDiffuser	FaceXHuBERT	FaceForner	CodeTalker	CodeTalker-ND	ProbTalk3D	Ground Truth
Angry	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4
Disgust	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4
Contampt	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4
Fear	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4
Happy	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4
Sad	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4
Surprise	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4
Neutral	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4	1 in 4
Total Videos for each Model	4	4	4	4	4	4	4
Audio Source:	3DMEAD dataset (2 level 2 intensity sequence + 2 level 3 intensity sequences)						

Figure 27: User Study motions by model and emotion, with a random split of emotions for each user

The users were also required to fill in some information about themselves such as age, country of residence and their prior experience with Virtual Humans, video games and 3D Animation movies. The distribution among age categories and experience on specific factors is showcased in Image 28 and Image 29 respectively.

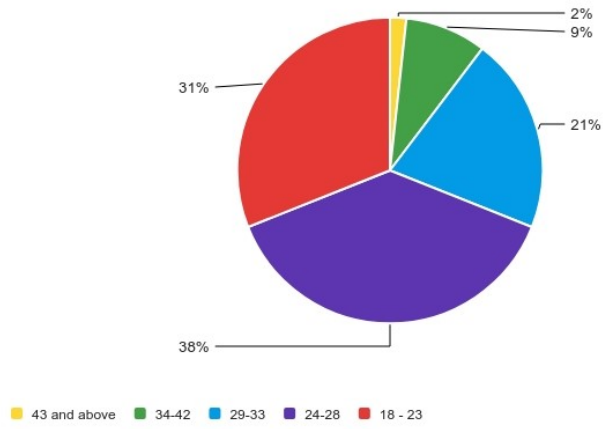


Figure 28: User Study age distribution among categories

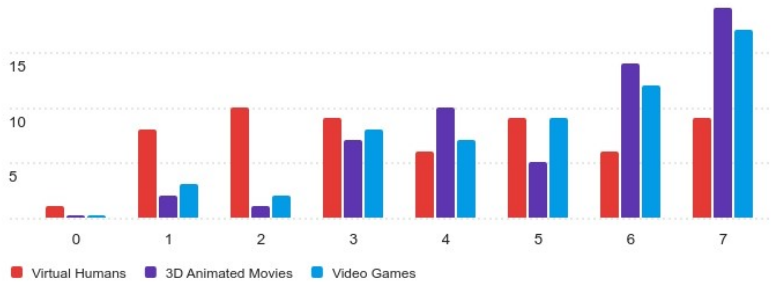


Figure 29: User Study experience rating among the users