MMHead: Towards Fine-grained Multi-modal 3D Facial Animation

Anonymous Authors

Figure 1: We present MMHead, the first multi-modal 3D facial animation dataset with hierarchical text annotations including abstract descriptions for overall actions and emotions, and fine-grained descriptions for expressions, head poses, as well as possible scenarios that may cause such emotions.

ABSTRACT

3D facial animation has attracted considerable attention due to its extensive applications in the multimedia field. Audio-driven 3D facial animation has been widely explored with promising results. However, multi-modal 3D facial animation, especially text-guided 3D facial animation is rarely explored due to the lack of multi-modal 3D facial animation dataset. To fill this gap, we first construct a large-scale multi-modal 3D facial animation dataset, MMHead, which consists of 49 hours of 3D facial motion sequences, speech audios, and rich hierarchical text annotations. Each text annotation contains abstract action and emotion descriptions, fine-grained facial and head movements (i.e., expression and head pose) descriptions, and three possible scenarios that may cause such emotion. Concretely, we integrate five public 2D portrait video datasets, and propose an automatic pipeline to 1) reconstruct 3D facial motion sequences from monocular videos; and 2) obtain hierarchical text annotations with the help of AU detection and ChatGPT. Based on the MMHead dataset, we establish benchmarks for two new tasks: text-induced 3D talking head animation and text-to-3D facial motion generation. Moreover, a simple but efficient VQ-VAE-based method named MM2Face is proposed to unify the multi-modal information and generate diverse and plausible 3D facial motions, which achieves competitive results on both benchmarks. Extensive experiments and comprehensive analysis demonstrate the significant potential of our dataset and benchmarks in promoting the development of multi-modal 3D facial animation.

CCS CONCEPTS

• Computing methodologies → Animation.

KEYWORDS

3D Facial Animation, Multi-modal Generation, Fine-grained Text Annotation, Dataset and Benchmark, VQ-VAE

1 INTRODUCTION

3D facial animation is becoming an increasingly popular topic in computer vision and multimedia due to its numerous applications in the multimedia field such as AR/VR content creation, games, and film production. The strong correlation between speech audio and facial movements makes it possible to automatically generate 3D facial motion (which can then be used to animate avatars) from audio, which will greatly simplify the animation production pipelines. Text, as another commonly used modality in human society, plays an important role in human-controlled AIGC. Recently, text guided image [3, 47], video [49, 62] and human motion [23, 54, 82] generation have achieved pleasing results with numerous multimedia applications. It would also be desirable to generate 3D facial motion...
from text descriptions. Further, it would be even more inspiring to achieve fine-grained 3D facial animation under multi-modal signal control. However, due to the lack of relevant datasets, fine-grained multi-modal 3D facial animation has been rarely explored.

In the field of 3D facial animation, impressive progress has been made in audio-driven 3D facial animation [1, 11, 41, 50, 56, 66, 70], making it possible to generate 3D facial motion sequences with synchronized lip movements based on the input talking audio. Moreover, some works [13, 42, 64] have incorporated motion labels to achieve explicit control over talking emotions. Recently, [37, 86] tried to use text descriptions in natural language rather than motion labels to guide the facial animation. However, they only consider textual descriptions of emotions during talking, while overlooking the actions other than talking, as well as text descriptions of fine-grained facial and head movements. Such limitations seriously affect the convenience and flexibility of 3D face animation in multimedia applications. A major reason for these limitations is the lack of open-source multi-modal 3D facial animation datasets with fine-grained text annotations, as shown in Tab. 1.

To alleviate the scarcity of multi-modal 3D facial animation datasets, we present MMHead, a large and diverse multi-modal 3D facial animation dataset with rich hierarchical text annotations. To the best of our knowledge, MMHead is the first 3D facial animation dataset with both abstract and fine-grained text annotations, including abstract action and emotion descriptions, fine-grained facial and head movement (i.e., expression and head pose) descriptions, as well as possible scenarios that may cause the person’s emotion (see Fig. 1 and Tab. 1). Concretely, we build our 3D facial animation dataset from 2D portrait video datasets [35, 63, 68, 76, 87], considering the huge amount of 2D videos and satisfactory monocular 3D face reconstruction results [12, 19, 20]. An automatic pipeline is also proposed to reconstruct 3D facial motion sequences from monocular videos, and then obtain abstract and fine-grained text annotations for these facial motions. The data construction pipeline can be easily scaled up to obtain larger datasets. To be specific, we first integrate and filter five public 2D portrait video datasets, i.e., CelebV-HQ [87], CelebV-Text [76], MEAD [63], RAVDESS [35], Multiface [68], to achieve richer facial actions and emotions. Then, we reconstruct 3D facial motion represented by FLAME parameters [32] via a state-of-the-art emotion-preserving monocular 3D reconstruction method named EMOCA [12, 19, 20]. Optimization and manual screening are then performed to obtain final 3D facial motion sequences that can achieve comparable precision compared to the lab-collected 3D facial animation datasets. As for text annotation, we explore the annotation capabilities of the large language model through well-designed prompts. Specifically, the action and emotion labels of the portrait video datasets, per frame activated Facial Action Units (AU) [36, 60] and head poses are conditionally combined with five different prompts to feed into ChatGPT [38] to obtain natural text descriptions of abstract action, abstract emotion, detailed expression, detailed head pose, and emotion scenarios, respectively.

Along with the proposed multi-modal 3D facial animation dataset, we benchmark two novel tasks: 1) text-induced 3D talking head animation (Benchmark I), which aims at generating 3D facial animations according to both speech audio and text instructions; 2) text-to-3D facial motion generation (Benchmark II), i.e., generating 3D facial motion sequences based on the given text descriptions. We define these tasks as non-deterministic generation tasks [73, 80] rather than regression tasks [11, 17, 70] to achieve robust and diverse 3D facial motion generation. Moreover, we propose a simple but efficient method named MM2Face to unify the multi-modal information and generate diverse and plausible 3D facial motions, which first utilizes a VQ-VAE [16, 58] to compress the 3D facial motions into a discrete codebook, and then generate plausible results.
in a constrained space by sampling the codebook in an autoregressive fashion through a transformer [59] equipped with specially designed attention maps. Based on MM2Face, we further explore the fusion and injection strategies of text and audio modalities to provide insight for future research.

In summary, our main contributions are:

- We propose the first multi-modal 3D facial animation dataset, MMHead, which contains rich hierarchical text annotations including abstract action and emotion descriptions, fine-grained facial and head movements descriptions, and possible emotion scenarios.
- We benchmark two new tasks: text-induced 3D talking head animation and text-to-3D facial motion generation to promote future research on multi-modal 3D facial animation.
- We propose a VQ-VAE-based method to unify the multi-modal information and generate diverse and plausible 3D facial motions. The proposed framework achieves competitive results on both benchmarks.

2 RELATED WORK

Audio-driven 3D Facial Animation. Audio-driven 3D facial animation aims at generating 3D facial motion sequences according to the input speech audio. In recent years, it has attracted much attention due to its potential use in virtual avatar animation [61, 67, 85], film, and game production. Audio-driven 3D facial animation methods can be roughly divided into two categories: rule-based methods [9, 15, 52, 72] and learning-based methods [1, 11, 13, 17, 41, 42, 46, 50, 51, 55, 66, 70]. Rule-based methods [9, 15, 52, 72] typically establish complex mapping rules between pronunciations and lip motions. This kind of method makes it easy to ensure the accuracy of the lip motion, however, requires a lot of manual effort and cannot generate the motion of the entire head. Learning-base methods solve the 3D facial animation problem in a data-driven way, and have made impressive progress in recent years [1, 11, 17, 41, 46, 50, 51, 55, 66, 70]. Furthermore, EmoTalk [42] and EMOTE [13] introduce emotion labels and achieve explicit control over talking emotions. ExpCLIP [86] goes one step further and uses emotion text descriptions rather than labels to control the emotion of audio-driven 3D facial animation. However, it only considers the textual control of emotions during talking, while ignoring the fine-grained control over facial and head movements. In addition, text-to-3D facial motion generation without input speech audio has never been explored except for expression generation [39, 88], which affects the convenience and flexibility of 3D face animation in multimedia applications. To fill this gap and promote the research on fine-grained multi-modal 3D facial animation, we present the first large-scale multi-modal 3D facial animation dataset with rich hierarchical text descriptions in this paper.

Text-guided Generation. Incorporating text descriptions into generative models has been a popular topic in AIGC areas such as text-to-image/video/motion generation. In the common setting of text to image task, generative models [21, 28, 31, 58] are widely used to capture the multi-modal distributions of images, and can generate diverse and plausible images compared to the regression models. Early work [25, 30, 53, 57, 69, 78] often uses GANs or VAEs with well-designed networks for text-guided generation. Recently, denoising diffusion models [5, 14, 47, 49, 54, 74, 77, 92] and VQ-VAE have been explored with superior performance. VQ-VAE is first explored in text to image with numerous researches [4, 16, 75]. In recent years, lots of works [22, 24, 29, 44, 81] also focus on text-to-3D human motion generation and achieve promising results. Despite the impressive progress of text-guided generation in numerous areas, text-guided 3D facial animation has been rarely explored before. To this end, we explore text-to-3D facial motion generation based on the proposed MMHead dataset in this paper.

3 MMHEAD DATASET

3.1 Data Collection

Dataset Integration. We integrate five commonly used portrait video datasets with rich facial action and emotion to obtain the candidate video set for our 3D facial motion. (1) CelebV-HQ [87], a large-scale and high-quality in-the-wild face video dataset with manually labeled facial action and emotion labels. We remove videos with actions that are hard to see on 3D head, including "kiss", "listen to music", "play instrument", "smoke", and "whisper",
Figure 2: Dataset construction pipeline. We first integrate five portrait video datasets, and filter the data to obtain the candidate videos for constructing our 3D facial animation dataset. Then, high-precision 3D facial motion sequences are obtained from candidate videos through monocular 3D face reconstruction, FLAME parameter optimization, and manual screening in turn. Finally, we utilize ChatGPT with well-designed prompts to obtain abstract and fine-grained text annotations.

Figure 3: Text annotation pipeline with well-designed prompts. We use different prompts and input information to annotate each data with five types of text descriptions separately by ChatGPT.

reconstruction results may exhibit instability, such as incorrect shapes and jittery motions. To solve this problem, we first sieve out the outliers of the estimated FLAME parameters through Box Plot and replace them with the values of neighboring points, and then smooth the expression and pose parameters via per-video optimization. All 3D facial motion sequences are manually checked to guarantee quality. For more details, please refer to the Sup. Mat. Text Annotation. We annotate each 3D facial motion with rich text descriptions including abstract action and emotion descriptions, fine-grained facial and head movements descriptions, and three possible scenarios that may cause such emotion. These five types of text descriptions are annotated separately as illustrated in Fig. 3. For full content and format of the well-designed prompts, please refer to the Sup. Mat. (1) Abstract action. Since the lab-collected part of our candidate video set only contains one action in “talking”, “singing” and “making a facial expression”, we predetermine seven text descriptions for each of these three actions, and then randomly select one of the seven predefined texts as the abstract action text description for each 3D facial motion from lab-collected videos. As for 3D facial motions from in-the-wild videos, multiple actions may occur in one motion sequence and we only have the action labels that are not strictly in chronological order, so we feed ChatGPT additional information including per frame activated AU and head pose labels with their intensity values, and let ChatGPT deduce the chronological order of the action labels, and then give

and reuse the remaining action and emotion labels in our automatic text annotation pipeline for abstract text descriptions. (2) CelebV-Text [76], a large-scale and high-quality in-the-wild dataset with videos for constructing our 3D facial animation dataset. Then, high-precision 3D facial motion sequences are obtained from portrait videos through 3D face reconstruction, FLAME parameter optimization, and manual screening in turn. Finally, we utilize ChatGPT with well-designed prompts to obtain abstract and fine-grained text annotations.

Video Filtering. We first unify all the portrait videos we got above into 25 FPS, and then remove videos with more than 200 frames. Moreover, we filter out videos with poor audio quality, since the synchronization between audio and facial motion, especially mouth motion, is of great importance, however, out-of-sync sound and picture, excessive background noise or background music, and multiple people speaking together in one video are common in in-the-wild videos. We follow VoxCeleb [7] and utilize SyncNet [8], a pre-trained model for determining the audio-video synchronization, to get the confidence score of each in-the-wild portrait video, and then filter out videos with confidence score lower than a specified threshold.

3.2 Automatic Annotation Pipeline

3D Facial Motion Reconstruction. We utilize sequences of expression (50 dimensions) and pose (6 dimensions) parameters of a commonly used parametric head model, FLAME [32], to represent 3D facial motion. To reconstruct FLAME parameters from portrait videos, we select a state-of-the-art emotion-preserving monocular 3D face reconstruction method, EMOCA [12, 19, 20], to estimate FLAME parameters for each frame. However, due to the shaking of the people and complex background in the videos, frame-by-frame

Optimize

Manual Check

Text Annotator

Mono. 3D Face Recon.

Activated AUs

Filtering

Candidate Videos

AU Detection

3D Facial Motion

Text Annotations

Portak Video Dataset

<action> The portrait looks around and then talks while...
<emotion> The portrait looks happy...
<head pose> The portrait starts with a slight tilt to the left...
<expression> The portrait starts with a raised inner brow...
<scenarios> Recognizing with a long lost friend...

ChatGPT

I’ll give you an [action label]... and your task is to string the action labels together into a statement that describes the portrait’s actions in chronological order. To help you figure out the chronological order, I’ll also give you a sequence of activated facial action units and head pose descriptions that reflects the detail of the portrait’s facial and head movement...

I’ll give you an [emotion label]... and your task is to string the emotion labels in order into a descriptive statement that describes the emotions of the portrait in the video clip...

I’ll give you a [sequence of head pose descriptions]... and your task is to summarize the head movements of the portrait into a sequence of head poses which describes the head pose dynamics in chronological order...

I’ll give you a [sequence of activated facial action descriptions]... and your task is to summarize the portrait’s detailed facial movements into a description which describes the facial movement dynamics in chronological order...

I’ll give you a [text description] which describes the emotion of the person, and your task is to imagine three possible scenarios that could cause such emotion...
3.3 Dataset Analysis

3.3.1 Data Statistics. As described above, we construct MMHead from five online portrait video datasets and end up with 35903 3D facial motion sequences up to 49 hours in total. Specifically, there are 10505, 11128, 11313, 2452, 505 motion sequences from CelebV-HQ [87], CelebV-Text [76], MEAD [63], RAVDESS [35], MultiFace [68], respectively. Each 3D facial motion is annotated with five types of text annotations including abstract action and emotion, fine-grained facial and head movements, and three possible scenarios. For more details about the motion duration, action and emotion categories statistics, please refer to the Sup. Mat.

3.3.2 Dataset partitioning for Two Benchmarks. We benchmark two tasks based on the MMHead dataset. For text-induced talking head animation (Benchmark I), we use a total of 28466 3D facial motion sequences related to “talk”, “sing”, and “read”. For text-to-3D facial motion generation (Benchmark II), we use 7937 sequences in total, including 505 sequences with various facial expressions from the Multiface dataset, 250 talking, and 250 singing sequences randomly sampled from the data used in benchmark I, and a total of 6932 sequences for all actions other than “talk”, “sing”, and “read”.

4 MM2FACE METHOD

We explore a simple but efficient two-stage method MM2Face that can handle both newly proposed text-induced 3D talking head animation and text-to-3D facial motion generation tasks. MM2Face can synthesize diverse and plausible 3D facial motions given a speech audio sequence (optional) and a fine-grained text description.

Our MM2Face framework consisting of two stages is illustrated in Fig. 4. In the first stage (section 4.1), given a 3D facial motion sequence, we utilize a motion tokenizer to extract discrete and compact 3D facial motion tokens. In the second stage (section 4.2), we conduct an auto-regressive token modeling task in discrete token space using our MM2Face transformer [59]. During inference, MM2Face iteratively predicts the likelihood of motion tokens in each step hence allows for non-deterministic motion generation.

4.1 Facial Motion Tokenization

In stage I, we utilize VQ-VAE to extract meaningful discrete facial motion representations. We utilize a 1D CNN-based network [26] to train a facial motion VQ-VAE \( \mathcal{V} \) on diverse 3D facial motion data, which consists of a motion encoder \( \mathcal{E} \), a vector quantization network \( \mathcal{Q} \), a codebook \( \mathcal{B} \), a motion decoder \( \mathcal{D} \). Given a motion sequence \( F_{t:T} = (F_1, F_2, ..., F_T) \in \mathbb{R}^{T \times d} \), where \( T \) denotes the motion frame number and \( d \) denotes the dimension of FLAME parameters. The encoder \( \mathcal{E} \) extract meaningful feature vectors \( z = (z_1, z_2, ..., z_T) \) from \( F \), where \( T' = T/r \) and \( r \) is the temporal downsampling rate (\( r = 4 \) in our paper). Then the feature vector \( z \) is processed by the vector quantization network to get quantized features \( \hat{z} = (\hat{z}_1, ..., \hat{z}_{T'}) \) via searching nearest neighbors in the
where \( \mathcal{B} \), which can be written as:

\[
\hat{z}_j = \arg \min_{b_k \in \mathcal{B}} ||z_j - b_k||.
\]  

Finally, the quantized feature \( \hat{z} \) is fed into the motion decoder \( D \) to reconstruct human motions \( F \). Through the process, we can convert facial motion data to discrete features and their corresponding codebook indices \( i \), i.e., motion tokens.

**Training Loss.** We adopt the motion reconstruction loss, embedding commitment loss for training. The reconstruction loss is:

\[
L_{\text{rec}} = L_{\text{mse}}(F_{1:T}, \hat{F}_{1:T}) + \alpha L_{\text{mse}}(V(F_{1:T}), V(\hat{F}_{1:T})),
\]

where \( L_{\text{mse}} \) denotes the MSE loss function, \( V(F_{1:T}) = F_{2:T} - F_{1:T-1} \) denotes the velocity of \( F \). \( \alpha \) is a hyper-parameter.

The commitment loss is:

\[
L_{\text{commit}} = ||sg(\hat{z}) - z|| + \beta ||sg(z) - \hat{z}||,
\]

where \( \beta \) is a hyper-parameter, \( sg \) denotes the stop gradient operation. Finally, we combine these losses \( L_{\text{eq}} = L_{\text{rec}} + L_{\text{commit}} \) to optimize the VQ-VAE network.

### 4.2 Auto-regressive Token Modeling

After converting all the facial motions to discrete tokens, auto-regressive token modeling is conducted to train our MM2Face transformer model \( \mathcal{G} \). The MM2Face transformer, as shown in Fig. 4, consists of an audio encoder \( \mathcal{A} \), a full text encoder \( T_{f{}\text{ull}} \), hierarchical text encoders \( T_{\mathcal{H}{}1:5} \), and a MM2Face transformer \( C \).

Specifically, the audio \( A_{1:T} \) is fed into \( \mathcal{A} \) to extract dense audio features, we concatenate all fine-grained text descriptions into a full text and feed it into \( T_{f{}\text{ull}} \) to obtain full text feature, fine-grained texts are separately fed into hierarchical text encoders \( T_{\mathcal{H}{}1:5} \) to obtain hierarchical text features, then audio features and hierarchical text features are concatenated together as cross attention inputs.

The MM2Face transformer \( C \) takes the ground truth motion tokens, audio, full and hierarchical text features as inputs and sequentially predicts the generated motion tokens \( \hat{I} \). Then \( \hat{I} \) can be decoded to final generated facial motions \( \hat{F} \).

**Architecture.** Our audio encoder \( \mathcal{A} \) is a pre-trained state-of-the-art speech model wav2vec 2.0 [2], which is composed of a temporal convolution-based network and a multi-layer transformer encoder. \( \mathcal{A} \) takes the audio inputs \( A_{1:T} \) and outputs the audio feature. The text encoder \( T \) is a pre-trained distilbert [48] model. The MM2Face transformer \( C \) is a multi-layer transformer decoder. Each layer of \( C \) is composed of normalization layers, a self-attention layer, a feed-forward layer, and a cross-attention layer.

**Biased Cross-attention.** Giving the concatenated hierarchical text features and audio features as inputs for cross-attention calculation, we specifically design a biased cross-attention mask \( \text{mask} \) in Fig. 4 to align the motion tokens with inputs:

\[
\text{Att} = \text{Softmax} \left( \frac{Q^T (K)^T \times \text{mask}}{\sqrt{d_k}} \right),
\]

where \( K \in \mathbb{R}^{T \times d_k} \) is the key of the concatenated hierarchical text features and audio features.

**Training Loss.** During training, we adopt the teacher-forcing strategy and directly maximize the log-likelihood of the facial motion token representations:

\[
L_{\text{MM2Face}} = E_{x \sim p(x)} [-\log p(x|z)].
\]
datasets [23, 33, 34, 71], demonstrating that our text descriptions are accurate enough to conduct multi-modal generation tasks.

**Audio-to-3D Facial Motion Alignment Evaluation.** Similarly, we slightly modify the text-to-motion retrieval method in Fig. 5 by changing the text encoder to a Wav2Vec [2] audio encoder, and removing the 3D head pose as motion inputs. After conducting the experiments, the Top-1 to Top-3 R-Precision accuracy is 0.885, 0.942, and 0.962, demonstrating the great audio-3D face motion matching performance.

### 5.2 Text-induced 3D Talking Head Animation

The provided fine-grained textual annotations combined with corresponding audio 3D facial motions lead to a new task, i.e. text-induced 3D talking head animation. Since there are no methods designed for this task, we select 5 state-of-the-art 3D talking head methods, i.e. Faceformer [17], CodeTalker [70], Facediffuser [50], Imitator [56], and Selftalk [41] and modify them by adding a text encoder. Concretely, for mesh-based methods [17, 41, 56, 70], we replace the template feature with the text feature, for parameter-based methods [50], we concatenate the text feature with the audio feature.

**Experiment setup.** We adopt the common protocol to split our dataset into training, validation, and test sets with a ratio of 0.8, 0.05, and 0.15. We select the 56 dimensions FLAME parameters as the input of our method and evaluate the generation performance on generated mesh sequences for fair comparison with other mesh-based methods.

**Metrics.** Previous approaches [17, 70] treat this task as a deterministic prediction task instead of a generative task, hence only the accuracy-related metrics are calculated. To evaluate the diversity and plausibility of generated motions, we follow the previous text to human motion task [23, 24] and select Frechet Inception Distance (FID), text R-Precision, audio matching score, diversity and lip vertices error [17] as our evaluation metrics: (1) **FID**, which is the distribution distance between the extracted features of generated motion and real motion by our pre-trained text-motion retrieval network introduced in Fig. 5. (2) **Text-R-Precision**: we calculate the top-1 and top-3 text to motion retrieval accuracy as reported metrics. We specifically compute them by ranking the Euclidean distance between the facial motion and text embeddings in a batch of 32 motion-text pairs. (3) **Audio Matching Score** is the average Euclidean distances between each gt audio feature and the generated motion feature extracted by our audio-motion retrieval network trained by the method mentioned in section 5.1. (4) **Diversity.** We randomly select 300 motion pairs from generated motions. Then we extract motion features and compute the average Euclidean distances of each motion pair to compute motion diversity in test set. (5) **Lip Vertex Error (LVE)** [46] calculates the maximum L2 error across all lip vertices for each frame. (6) **Face Vertex Error (FVE)** calculates the average L2 error across all face region vertices for each frame.

**Quantitative Evaluation.** The quantitative results are summarized in Tab. 3 on MMHead dataset. For motion reconstruction, our VQ-VAE achieves considerable reconstruction performance against the real facial motion, demonstrating the excellent discrete representations. For motion generation, we can observe that our MM2Face achieves the best performance against other state-of-the-art methods in almost all metrics. Concretely, the excellent Text R-Precision accuracy demonstrates that MM2Face is able to synthesize plausible 3D facial motions while correctly following the fine-grained text descriptions. The excellent Audio matching score, LVE and FVE demonstrate MM2Face’s ability for precise motion generation. The FID score also indicates the generative distribution modeling performance of our MM2Face method.

**Qualitative Evaluation.** We also provide the qualitative results in Fig. 6. Our approach exhibits superior text and face alignment with accurate audio-visual correspondence. More qualitative results can be seen in supplementary material.

**Ablation Study.** Since our MM2Face method focuses on the multi-modal generation of facial motions, we perform an ablation study of various modality fusion strategies. Considering self-attention and cross-attention are the two most popular fusion mechanisms, we conducted a detailed study. The experiment results are summarized in Table 4. First, we can observe that merely utilizing full text results...
We re-implement these methods to adapt to our dataset format.

Table 5: Experimental Results for benchmark II: Text-to-3D Facial Motion Generation.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Text R-Precision ↑</th>
<th>FID ↓</th>
<th>Text-Match ↓</th>
<th>Diversity →</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real motion</td>
<td>0.678 0.842 0.904</td>
<td>0</td>
<td>19.14</td>
<td>33.39</td>
</tr>
<tr>
<td>VQ-VAE (Recons.)</td>
<td>0.492 0.667 0.763</td>
<td>21.29</td>
<td>21.72</td>
<td>33.48</td>
</tr>
<tr>
<td>TM2T [24]</td>
<td>0.316 0.481 0.559</td>
<td>77.21</td>
<td>25.35</td>
<td>32.17</td>
</tr>
<tr>
<td>T2M-GPT [81]</td>
<td>0.531 0.495 0.598</td>
<td>22.36</td>
<td>24.62</td>
<td>32.91</td>
</tr>
<tr>
<td>MDM [54]</td>
<td>0.214 0.327 0.404</td>
<td>64.45</td>
<td>27.82</td>
<td>31.82</td>
</tr>
<tr>
<td>Our MM2Face</td>
<td>0.405 0.601 0.698</td>
<td>35.80</td>
<td>23.17</td>
<td>33.30</td>
</tr>
</tbody>
</table>

Table 5: Experimental Results for benchmark II: Text-to-3D Facial Motion Generation. The best and runner-up performances are bold and underlined, respectively.

5.3 Text-to-3D Facial Motion Generation

MMHead contains diverse facial expression sequences from specific face expression datasets and general facial motion datasets. Hence we explore a new text-to-3D facial motion generation task. We conduct experiments with the state-of-the-art text-to-3D human motion methods, i.e., MDM [54], TM2T [24] and T2M-GPT [81]. We re-implement these methods to adapt to our dataset format.

Experiment Setup and Metrics. Similar to benchmark I, we select 56 dimensions FLAME parameters as motion representations and adopt the Frechet Inception Distance, R-precision, diversity, and text matching score for evaluation. We re-train a text-to-3D facial motion retrieval network as introduced in Section 5.1 in benchmark II subset for metric calculation.

Quantitative Evaluation. The quantitative results are summarized in Tab. 5 on the benchmark II subset of the MMHead dataset. We can derive that our method achieves the best performance against all metrics. T2M-GPT [81] achieves the second-best performance in FID, matching score, diversity and R-precision. MDM [54] achieves considerable FID performance but fails to capture the text-to-motion consistency according to R-precision. From the results, we derive that our MMHead dataset has the potential for further explorations.

Qualitative Evaluation. We also provide the qualitative results in Fig. 7. Our approach exhibits superior text and face alignment. More results are in the supplementary material.

6 CONCLUSION

In this paper, we push forward the 3D facial animation task, and present MMHead, the first multi-modal 3D facial animation dataset with rich hierarchical text annotations including abstract action and emotion descriptions, fine-grained facial and head movements descriptions, and possible emotion scenarios. With MMHead, we benchmark two new tasks: text-induced 3D talking head animation and text-to-3D facial motion generation. Moreover, we propose a simple but efficient VQ-VAE-based method named MM2Face to explore the multi-modal information fusion strategies and generate diverse and plausible 3D facial motions, which achieves competitive performance on both benchmarks. We hope that the MMHead dataset and the corresponding benchmarks will promote in-depth research works on multi-modal 3D facial animation.


5100

5100