

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

Anonymous Authors

In the supplementary document, we first provide more details about MMHead dataset and MM2Face method in Section 1 and Section 2, respectively. Then, we show additional experimental results in Section 3. Finally, the limitations and broader impacts are discussed in Section 4. Please also refer to the accompanying video for more intuitive results.

1 MMHEAD DATASET

In this section, we first provide more details about how we obtain high precision 3D facial motion sequences and hierarchical text annotations from portrait videos, and then conduct comprehensive statistical analysis of the proposed MMHead dataset.

1.1 3D Facial Motion from Portrait Videos

We use sequences of expression and pose parameters of FLAME [7] to represent 3D facial motion. To obtain FLAME parameters from portrait videos, we utilize a SOTA monocular 3D face reconstruction method, EMOCA [2–4], to estimate FLAME parameters frame by frame. Then we optimize the obtained FLAME parameters to improve the quality of 3D facial motion. Specifically, we first calculate the mean value of shape, pose, expression, and bounding box size over time respectively, and then calculate L_2 distance of each value to the corresponding mean value and treat points whose distances are outside the interval $[Q_1 - s \times IQR, Q_3 + s \times IQR]$ as outliers and replace them, where $IQR = Q_3 - Q_1$, Q_1 and Q_3 denote lower quartile and upper quartile respectively, s is the scale factor. Then, we optimize the expression and pose parameters by minimizing the following loss function using Adam optimizer [6]:

$$\mathcal{L} = \|(\psi_{2:T} - \psi_{1:T-1})\|_2^2 + \lambda \|(\theta_{2:T} - \theta_{1:T-1})\|_2^2, \quad (1)$$

where, ψ and θ denote the FLAME expression and pose parameters respectively, T is the length of the facial motion sequence, λ is the balance weight.

Although the optimization step can remove the outliers and smooth the 3D facial motion sequences, motion sequences in which vast majority of frames cannot be correctly reconstructed are difficult to find and remove. To this end, we manually check the dataset and delete the bad reconstruction results, we summarized some common failure reconstruction results in Fig. 1. Finally, we obtain various great reconstruction results as shown in Fig. 2

1.2 Text Annotation

As described in the main paper, we design five different prompts for annotating abstract action and emotion descriptions, fine-grained head and facial movements (*i.e.*, head pose and expression) descriptions, and emotion scenarios, respectively. The prompts we used is shown in Fig. 3 and Fig. 4. Specifically, we follow [10] and use [8] to detect 41 facial action units (AU) frame by frame. As for head pose, we define 6 head pose descriptions according to the rotation



Figure 1: We remove such failed cases to ensure the quality of 3D facial motion.

Rotation Range	Description
$r_x < 0$	Head up
$r_x > 0$	Head down
$r_y < 0$	Head turns right
$r_y > 0$	Head turns left
$r_z < 0$	Head tilts left
$r_z > 0$	Head tilts right

Table 1: Head pose descriptions according to the rotation vector $r = (r_x, r_y, r_z)$ of the FLAME neck joint.

vector $r = (r_x, r_y, r_z)$ of the FLAME neck joint as illustrated in Tab. 1.

1.3 Dataset Statistics

We conduct data analysis of MMHead dataset. We gradually analyze the motion duration distribution, action distribution, emotion distribution and head pose distribution.

Motion Duration Distribution. The duration distribution is summarized in Fig. 5. MMHead contains diverse motions which cover a wide range of duration times (1 seconds to 8 seconds). The similar distributions of Bench I subset and Bench II subset also demonstrate that our talking animations and 3D facial motions both contains various motions.

Action Distribution. The action distribution indicates the distribution of action words extracted from abstract action descriptions in our dataset, reflecting the diversity of our abstract action text descriptions. The results are summarized in Fig. 7. We can observe that (1) "talk" is the word with the highest frequency in MMHead dataset because most of our data are collected from talking head datasets. (2) Compared to the action distributions in MMHead and Bench I datasets, the distribution in Bench II subset is quite different, which contains many high frequency novel actions such as "make a face", "eat", "chew", "weep" etc. This demonstrate that 3D facial motion is quite different from 3D talking head because benchmark II are mainly focusing on facial expression motion generation instead of talking motions with corresponding audios.

Emotion Distribution. The Fig. 6 reflects the diverse emotion distribution in our MMHead dataset.

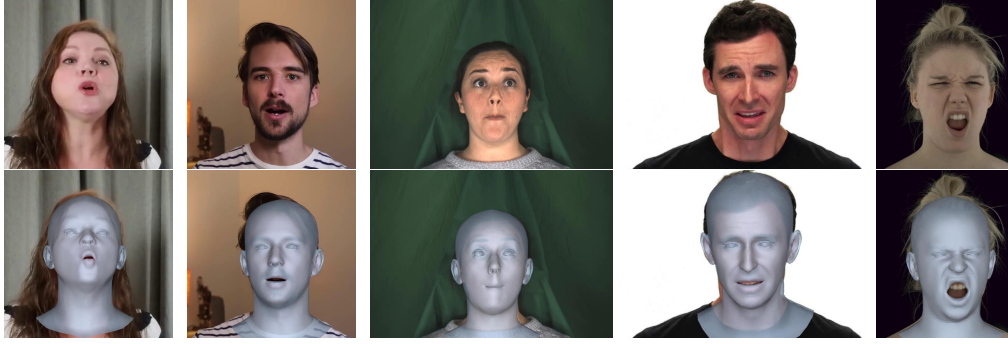


Figure 2: Key frames of 3D facial motion reconstructed from portrait videos, which demonstrates the high precision of constructing 3D facial animation dataset by monocular 3D face reconstruction.

I'll give you an [action label set] (in the format of {frown, head_wagging, talk}) which consists of the actions that the portrait in a video clip have done, and your task is to use natural language to string the action labels in the [action label set] together into a descriptive statement that describes the actions of the portrait in the video clip according to the chronological order. To help you figure out the chronological order of the portrait's actions, I'll also give you a [sequence of activated facial action units and head pose descriptions] (in the format of <frame id, Lid tightener: 0.79, Upper lip raiser: 0.59, Left Inner brow raiser: 0.69, Head down: 0.25, Head turns left: 0.20, Head tilts right: 0.15>, where behind frame id is all the activated facial action units with its corresponding intensity ranging from 0~1, followed by head pose descriptions with its corresponding rotation angle expressed in radians ranging from 0~1.57) that reflects the details of the portrait's facial and head movements. Note again that your task is to naturally string the action labels in the [action label set] into a descriptive statement (no more than 30 words) to describe the portrait's actions in chronological order without the use of facial action units and head pose descriptions. Do not mention any useless or redundant statements. Make sure there are no numbers in your answer. Make sure there are no explicit references to activated facial action units and head pose descriptions in your answer. Make sure your answer is grammatically correct and described in the present simple tense.

Here is the [action label set]:
{talk, gaze, smile}

Here is the [sequence of activated facial action units and head pose descriptions]:
<0, Lips part: 0.81, Head up: 0.30>,
<3, Lips part: 0.81, Head up: 0.30>,
<6, Lid tightener: 0.68, Dimpler: 0.73, Chin raiser: 0.71, Lips part: 0.72, Head up: 0.30>,
<9, Inner brow raiser: 0.82, Lid tightener: 0.75, Upper lip raiser: 0.61, Dimpler: 0.68, Chin raiser: 0.75, Lips part: 0.83, Head up: 0.22>,
<12, Inner brow raiser: 0.92, Outer brow raiser: 0.70, Lid tightener: 0.74, Upper lip raiser: 0.87, Lips part: 0.91, No head pose>,
<15, Inner brow raiser: 0.86, Outer brow raiser: 0.64, Lid tightener: 0.79, Upper lip raiser: 0.90, Lip corner puller: 0.65, Lips part: 0.96, No head pose>,
<18, Inner brow raiser: 0.70, Lid tightener: 0.85, Upper lip raiser: 0.91, Lip corner puller: 0.90, Lips part: 0.97, No head pose>,
<21, Inner brow raiser: 0.64, Cheek raiser: 0.61, Lid tightener: 0.88, Upper lip raiser: 0.91, Lip corner puller: 0.91, Lips part: 0.97, No head pose>,
<24, Cheek raiser: 0.68, Lid tightener: 0.92, Upper lip raiser: 0.93, Lip corner puller: 0.93, Lips part: 0.97, Right Inner brow raiser: 0.64, No head pose>,
<27, Cheek raiser: 0.71, Lid tightener: 0.93, Upper lip raiser: 0.93, Lip corner puller: 0.94, Lips part: 0.97, Right Inner brow raiser: 0.83, No head pose>,
<30, Cheek raiser: 0.72, Lid tightener: 0.95, Upper lip raiser: 0.93, Lip corner puller: 0.94, Dimpler: 0.69, Lips part: 0.98, Right Inner brow raiser: 0.79, No head pose>,
<33, Cheek raiser: 0.68, Lid tightener: 0.95, Upper lip raiser: 0.95, Lip corner puller: 0.93, Lips part: 0.98, No head pose>,
<36, Cheek raiser: 0.64, Lid tightener: 0.96, Upper lip raiser: 0.94, Lip corner puller: 0.94, Dimpler: 0.66, Lips part: 0.98, Right Inner brow raiser: 0.90, Right Outer brow raiser: 0.64, Head up: 0.13, Head turns right: 0.11>,
<39, Cheek raiser: 0.72, Lid tightener: 0.96, Upper lip raiser: 0.95, Lip corner puller: 0.89, Lips part: 0.98, Head up: 0.16, Head turns right: 0.15>,
<42, Cheek raiser: 0.62, Lid tightener: 0.95, Upper lip raiser: 0.94, Lip corner puller: 0.79, Lips part: 0.98, Head up: 0.17, Head turns right: 0.20>,
<45, Lid tightener: 0.94, Upper lip raiser: 0.93, Lip corner puller: 0.68, Lips part: 0.97, Right Inner brow raiser: 0.76, Head up: 0.18, Head turns right: 0.22>,
<48, Lid tightener: 0.90, Upper lip raiser: 0.91, Lips part: 0.94, Right Inner brow raiser: 0.67, Head up: 0.18, Head turns right: 0.23>,
<51, Lid tightener: 0.92, Upper lip raiser: 0.87, Chin raiser: 0.68, Lips part: 0.87, Right Inner brow raiser: 0.75, Head up: 0.18, Head turns right: 0.24>,
...

I'll give you an [emotion label set] (in the format of {happy, natural}) which consists of the emotions of the portrait in a video clip, and your task is to use natural language to string the emotion labels in the [emotion label set] in order into a descriptive statement that describes the emotions of the portrait in the video clip. Note that there may be one or more than one emotion labels in the [emotion label set]. If there is only one emotion label in the [emotion label set], it means that the emotion of the portrait remains unchanged throughout the video, and you need to describe this unchanging status with appropriate natural language. If there are more than one emotion labels in the [emotion label set], it means that the portrait's emotion changed during the video, and the order of emotion change is the order of the emotion labels in the [emotion label set], so that you need to describe this change with appropriate natural language. Do not mention any useless or redundant statements, and make sure the answer is no more than 15 words. Make sure your answer is grammatically correct and described in the present simple tense.

Here is the [emotion label set]:
{happy}

Figure 3: Prompts used to annotate abstract action and emotion descriptions.

Head Pose Distribution. In addition, the distribution of head pose indicates the diversity of the dataset in head movements. As shown in Fig. 8, MMHead has a uniform distribution, which indicates that the motions have diverse head movements.

2 MM2FACE METHOD

2.1 Network Architecture

Our MM2Face method consists of a motion encoder \mathcal{E} , a motion decoder \mathcal{D} , our MM2Face transformer model \mathcal{G} . \mathcal{E} and \mathcal{D} are trained in stage 1, and \mathcal{G} is trained in stage 2.

I'll give you a [sequence of head pose descriptions] that is sampled every three frames from a 25fps portrait video (in the format of <frame id, Head down: 0.25, Head turns left: 0.20, Head tilts right: 0.15>, where behind frame id is all the head pose descriptions with its corresponding rotation angle expressed in radians ranging from 0~1.57), and your task is to summarize the head movements of the portrait into a fluent natural language description (no more than 100 words) which contains semantically rich information that accurately describes the head pose dynamics according to the chronological order. Do not mention any useless or redundant statements. Do not mention specific frame and radians value explicitly. Do not include any overall or summary statements in your answer, directly describe the portrait's head movements in chronological order. For example, if the chronological text description of the portrait's head movements is: (<XXXXXX>), the answer should be: (<XXXXXX>). Make sure your answer is grammatically correct and described in the present simple tense. Make sure there are no numbers in your answer.

Here is the [sequence of head pose descriptions]:

<0, Head up: 0.30>,
<3, Head up: 0.30>,
<6, Head up: 0.30>,
<9, Head up: 0.22>,
<12, No head pose>,
<15, No head pose>,
<18, No head pose>,
<21, No head pose>,
<24, No head pose>,
<27, No head pose>,
<30, No head pose>,
<33, No head pose>,
<36, Head up: 0.13, Head turns right: 0.11>,
<39, Head up: 0.16, Head turns right: 0.15>,
<42, Head up: 0.17, Head turns right: 0.20>,
<45, Head up: 0.18, Head turns right: 0.22>,
<48, Head up: 0.18, Head turns right: 0.23>,
<51, Head up: 0.18, Head turns right: 0.24>,
...

I'll give you a [sequence of activated facial action descriptions] that is sampled every three frames from a 25fps portrait video (in the format of <frame id, Lid tightener: 0.79, Upper lip raiser: 0.59, Left Inner brow raiser: 0.69>, where behind frame id is all the activated facial action units with its corresponding intensity ranging from 0~1), and your task is to summarize the detailed facial movements of the portrait into a fluent natural language description (no more than 100 words) which contains semantically rich information that accurately describes the facial movement dynamics according to the chronological order. Do not mention any useless or redundant statements. Do not mention specific frame and intensity value explicitly. Make sure your answer is grammatically correct and described in the present simple tense.

Here is the [sequence of activated facial action units descriptions]:

<0, Lid tightener: 0.55, Lips part: 0.81>,
<3, Lid tightener: 0.57, Lips part: 0.81>,
<6, Inner brow raiser: 0.59, Lid tightener: 0.68, Dimpler: 0.73, Chin raiser: 0.71, Lips part: 0.72>,
<9, Inner brow raiser: 0.82, Outer brow raiser: 0.55, Lid tightener: 0.75, Upper lip raiser: 0.61, Dimpler: 0.68, Chin raiser: 0.75, Lips part: 0.83, Jaw drop: 0.55>,
<12, Inner brow raiser: 0.92, Outer brow raiser: 0.70, Lid tightener: 0.74, Upper lip raiser: 0.87, Lips part: 0.91, Jaw drop: 0.55>,
<15, Inner brow raiser: 0.86, Outer brow raiser: 0.64, Lid tightener: 0.79, Upper lip raiser: 0.90, Lip corner puller: 0.65, Lips part: 0.96>,
<18, Inner brow raiser: 0.70, Cheek raiser: 0.56, Lid tightener: 0.85, Upper lip raiser: 0.91, Lip corner puller: 0.90, Lips part: 0.97>,
<21, Inner brow raiser: 0.64, Cheek raiser: 0.61, Lid tightener: 0.88, Upper lip raiser: 0.91, Lip corner puller: 0.91, Lips part: 0.97>,
<24, Cheek raiser: 0.68, Lid tightener: 0.92, Upper lip raiser: 0.93, Lip corner puller: 0.93, Lips part: 0.97, Right Inner brow raiser: 0.64>,
<27, Cheek raiser: 0.71, Lid tightener: 0.93, Upper lip raiser: 0.93, Lip corner puller: 0.94, Dimpler: 0.59, Lips part: 0.97, Right Inner brow raiser: 0.83, Right Outer brow raiser: 0.56>,
<30, Cheek raiser: 0.72, Lid tightener: 0.95, Upper lip raiser: 0.93, Lip corner puller: 0.94, Dimpler: 0.69, Lips part: 0.98, Right Inner brow raiser: 0.79, Right Outer brow raiser: 0.56>,
<33, Inner brow raiser: 0.57, Cheek raiser: 0.68, Lid tightener: 0.95, Upper lip raiser: 0.95, Lip corner puller: 0.93, Dimpler: 0.59, Lips part: 0.98, Jaw drop: 0.59>,
<36, Cheek raiser: 0.64, Lid tightener: 0.96, Upper lip raiser: 0.94, Lip corner puller: 0.94, Dimpler: 0.66, Lips part: 0.98, Right Inner brow raiser: 0.90, Right Outer brow raiser: 0.64>,
<39, Cheek raiser: 0.72, Lid tightener: 0.96, Upper lip raiser: 0.95, Lip corner puller: 0.89, Lips part: 0.98>,
<42, Inner brow raiser: 0.57, Cheek raiser: 0.62, Lid tightener: 0.95, Upper lip raiser: 0.94, Lip corner puller: 0.79, Lips part: 0.98>,
<45, Lid tightener: 0.94, Upper lip raiser: 0.93, Lip corner puller: 0.68, Lips part: 0.97, Right Inner brow raiser: 0.76>,
<48, Inner brow raiser: 0.57, Lid tightener: 0.90, Upper lip raiser: 0.91, Lips part: 0.94, Right Inner brow raiser: 0.67>,
<51, Lid tightener: 0.92, Upper lip raiser: 0.87, Lip corner puller: 0.58, Chin raiser: 0.68, Lips part: 0.87, Right Inner brow raiser: 0.75>,
...

I'll give you a [text description] (in the format of <The portrait's emotion shifts from happy to sad, then back to a natural expression. >) which describes the emotion of the person, and your task is to imagine 3 possible scenarios that could cause the person to experience such emotion. Do not mention any useless or redundant statements, just list the 3 scenarios one by one. Make sure your answer is grammatically correct and in the third person singular. Make sure each scenario is described in no more than 15 words.

Here is the [text description]:

<The portrait looks happy throughout the video clip.>

Figure 4: Prompts used to annotate fine-grained head and facial movements (i.e., head pose and expression) descriptions, and possible emotion scenarios.

Concretely, the motion encoder \mathcal{E} is comprised of ResNet based 1D CNN blocks, each block contains 1D CNN layers and normalization layers, we adopt relu function as activation function. Similarly, the motion decoder \mathcal{D} is comprised of ResNet [5] based 1D CNN blocks with upsample layers. We upsample the network features with nearest neighbor search strategy. The transformer model \mathcal{G} is a typical layers transformer decoder models. We select Wav2Vec

2.0 [1] as our audio encoder and freeze the parameters of its feature extractor. We select pre-trained distilbert [9] model as our text encoder.

2.2 Implementation Details

We use PyTorch to implement our MM2Face model, and utilize Adam optimizer [6] with $[\beta_1, \beta_2] = [0.5, 0.999]$ for training. All the audio signals are normalized by Wav2Vec2.0 [1] audio processor before training. Training the stage I model takes about 10 hours on

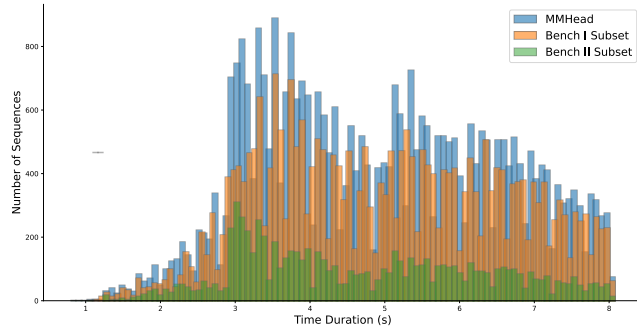


Figure 5: Histograms of facial motion time duration of entire MMHead dataset, benchmark I subset, and benchmark II subset, respectively.

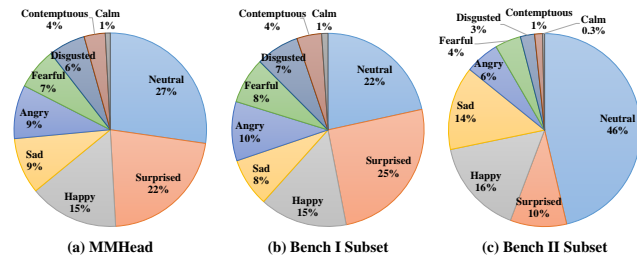


Figure 6: Pie charts of emotion categories of entire MMHead dataset, benchmark I subset, and benchmark II subset, respectively.

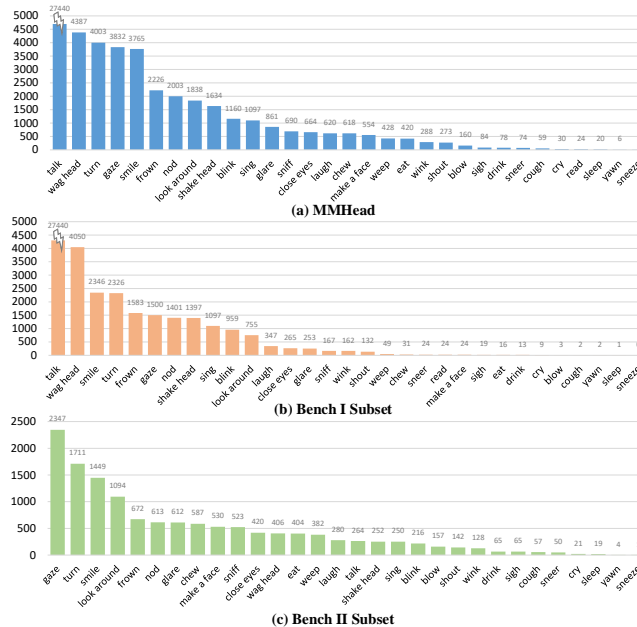


Figure 7: Distributions of all actions that occur in the entire MMHead dataset, benchmark I subset, and benchmark II subset, respectively.

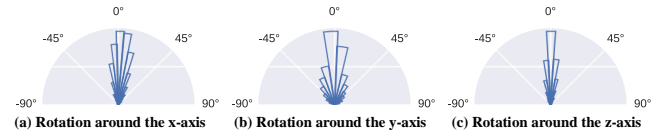


Figure 8: Histograms of the head rotation around x-axis (i.e., up/down), y-axis (i.e., turn left/right), and z-axis (i.e., tilt left/right), respectively.

1 Nvidia 3090 GPUs, stage II takes about 26 hours on 1 Nvidia 3090 GPUs.

3 ADDITIONAL EXPERIMENTAL RESULTS

3.1 Diverse Generation

We show the diverse generation results in the video of our supplementary materials.

3.2 Text Controlled 3D Talking Head Animation Results

We specifically evaluate the influence of text descriptions on generated 3D talking head.

Text guided Emotion Control Results. We conduct experiments to evaluate text control ability on emotion control. We demonstrate the brilliant emotion control ability of our MM2Face in Fig. 9. From Fig. 9, given an audio of a 3D talking head with neutral faces, we show different generation results by changing the abstract emotion and detailed expression descriptions. We select various descriptions from 7 test sequences belonging to 7 different emotions and our MM2Face merges the excellent emotion control ability.

Text guided Head Pose Control Results. We also provides the text guided head pose results in Fig. 10. From Fig. 10, we can observe that through changing head pose text descriptions, we can obtain head pose control results given the same audio input.

4 LIMITATIONS AND BROADER IMPACTS

Limitations: While our method contributes a satisfactory frame work for both text-induced talking head animation and text-to-3D facial motion generation, our method can still fail to generate some results. First, the reason is probably that for talking head animation, MM2Face model may not disentangle the text conditions and audio conditions perfectly, causing the model sometimes may focus on global text consistency instead of audio consistency. Discovering a more efficient ways to disentangle these two useful control signals is a promising future direction. Second, our method can only animate the FLAME mesh, it is also interesting to animate a high quality head mesh with detailed hairs and faces.

Boarder Social Impact: There can be also negative impact from this work. Our MMHead dataset and MM2Face model can be used to generate visually realistic and plausible 3D facial motions. And the 3D motions can be further utilized to animate a vivid human avatar. It may be misused to create fake face videos using Deepfake-like technology, which is horrible for society, we sincerely encourage more researchers focusing on deep fake detection.

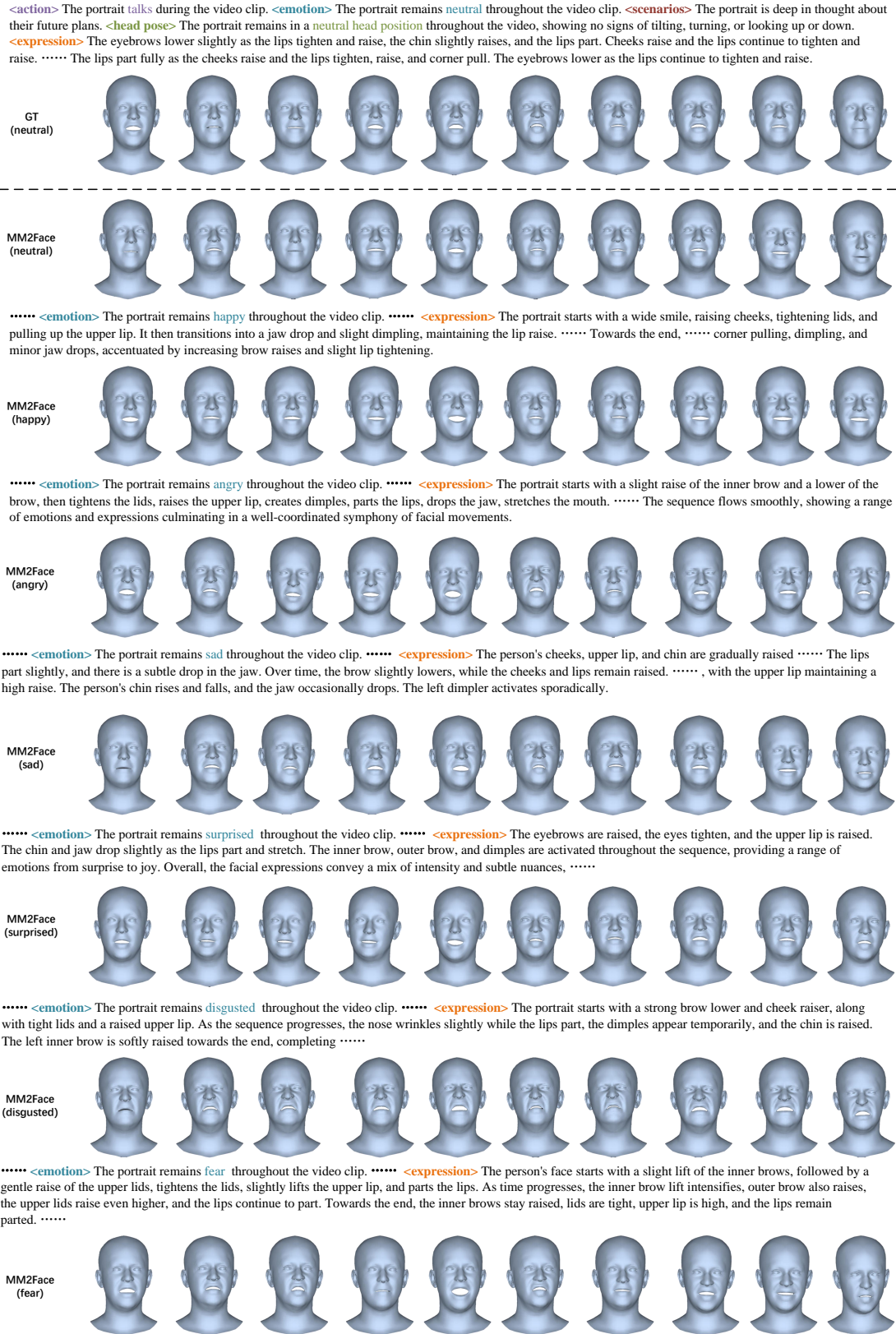
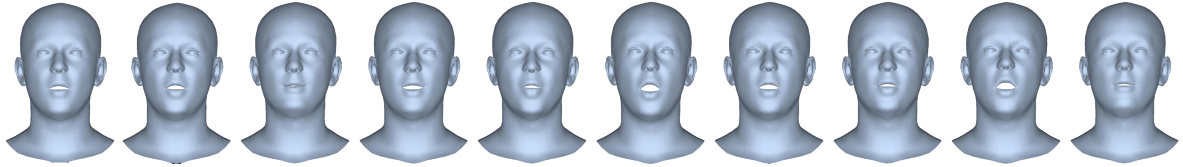


Figure 9: Visualization Results of Text guided Emotion Control of 3D Talking Head.

<action> The portrait talks in the video clip. **<emotion>** The portrait remains neutral throughout the video clip. **<scenarios>** The portrait is focused on a task, showing no emotion. **<head pose>** The portrait's head **remains still** for a while, then gradually starts to **lift upwards**. **<expression>** The portrait begins with the inner brow raising, tightening of the lids, raising of the upper lip, and lifting of the chin. The lips part slightly, and the jaw drops. Finally, the inner brow lowers slightly, while the lid tightens and the lips part. The jaw drops, and the left dimple is prominent.

GT



MM2Face



..... **<head pose>** The portrait gradually **tilts his head to the right**.

MM2Face
(right)



..... **<head pose>** The portrait starts with the head in a **neutral position**. As time progresses, the head **leans down**.

MM2Face
(down)



Figure 10: Visualization Results of Text guided Head Pose Control of 3D Talking Head.

REFERENCES

- [1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.
- [2] Radek Daněček, Michael J Black, and Timo Bolkart. 2022. EMOCA: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20311–20322.
- [3] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH* 40, 8 (2021). <https://doi.org/10.1145/3450626.3459936>
- [4] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. 2022. Visual Speech-Aware Perceptual 3D Facial Expression Reconstruction from Videos. *arXiv preprint arXiv:2207.11094* (2022).
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [6] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [7] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.
- [8] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. 2022. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782* (2022).
- [9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
- [10] Duomin Wang, Bin Dai, Yu Deng, and Baoyuan Wang. 2023. Agentavatar: Disentangling planning, driving and rendering for photorealistic avatar agents. *arXiv preprint arXiv:2311.17465* (2023).