Supplementary Materials: Speech-Driven 3D Facial Animation with Head Pose Using Decoupled Speech Feature

Anonymous Authors

1 MODULE DETAIL

1.1 Facial Animation Module

For the VQ-VAE network designed to model facial motion space, a motion block is initially processed by a linear embedding layer. This is followed by a encoder consisting of a convolutional layer with a kernel size of 5, stride of 1, and padding of 2 and a Transformer with 6 layers and 8 heads. The quantized embeddings are then fed into a decoder that mirrors the encoder architecture but in reverse order.

For the upper branch network tasked with predicting global motion based on semantic features, a 12-layer, 4-head Transformer decoder is employed.

For the embedding interpreter, a 2-layer LSTM is utilized.

1.2 Diverse Pose Module

For the audio encoder, the pre-trained wav2vec2.0 encoder is utilized to extract semantic feature. Additionally, we design a low-level feature encoder consists of six Conv1d layers followed by a linear layer to extract rythm feature. Then, we feed them together into a fusion network composed of four linear layers and LeakyReLU activation layers.

For the VQ-VAE network of pose, the encoder consists of a transformer block with 6 layers and 4 heads. The quantized embeddings are then fed into a decoder that mirrors the encoder architecture but in reverse order.

The pose autoregressive generator consists of three parts: Periodic Positional Encoding (PPE), a learnable style embedding layer, and a six-layer transformer decoder with four heads each.

2 TRAINING DETAIL

2.1 Dataset

For the reconstructed MEAD dataset, we selected the FLAME parameters of each subject performing at level 3 intensity across 8 emotions for reconstruction and divided them into motion blocks of $T_b = 8$ frames each for VQ-VAE training. Following this division, we obtained 226,621 motion blocks, from which we randomly selected 7,609 blocks for training. We believe that 8 frames (at 30fps, approximately 27ms) span a duration sufficient for a human to produce a phoneme and complete a mouth shape movement.

For the VoxCeleb2, we used seed 42 in PyTorch to randomly shuffle the dataset, then selected the first 70% of the audio-visual sequences as training data, the subsequent 15% as validation data, and the final 15% as test data. For the convenience of subsequent training and metric calculation, we further segmented the data to maintain videos at 200 frames in the temporal dimension, while the audio was divided into segments of 8 seconds in length. We believe that 8 seconds of audio is sufficient to contain a complete dialogue, with rich contextual information, allowing for better extraction of semantic information.

For the vocaset, where the training data includes 8 subjects, and testing and validation data include separate subjects each. Each subject performed 40 sentences, each approximately 4 seconds in length. VOCASET's video frame rate is 60fps, which we downsample to 30fps for our use.

2.2 Implementation

We train the facial motion VQ-VAE model on NVIDIA A100 for 250 epochs with AdamW optimizer($\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 1e-8$), the learning rate is 10^{-4}. And the batchsize is set to 64. We train the dual-path network on a single 4090(24GB graphic memory) for 250 epochs with AdamW optimizer($\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 1e-8$), the learning rate is 10^{-4}.

We train the head poses VQ-VAE model on a single 4090(24GB graphic memory) for 200 epochs with same hyper-parameters as lip sync module, but set the batchsize to 32. And then we changed the batchsize to 16 to train the Transformer-liked autoregressive model.

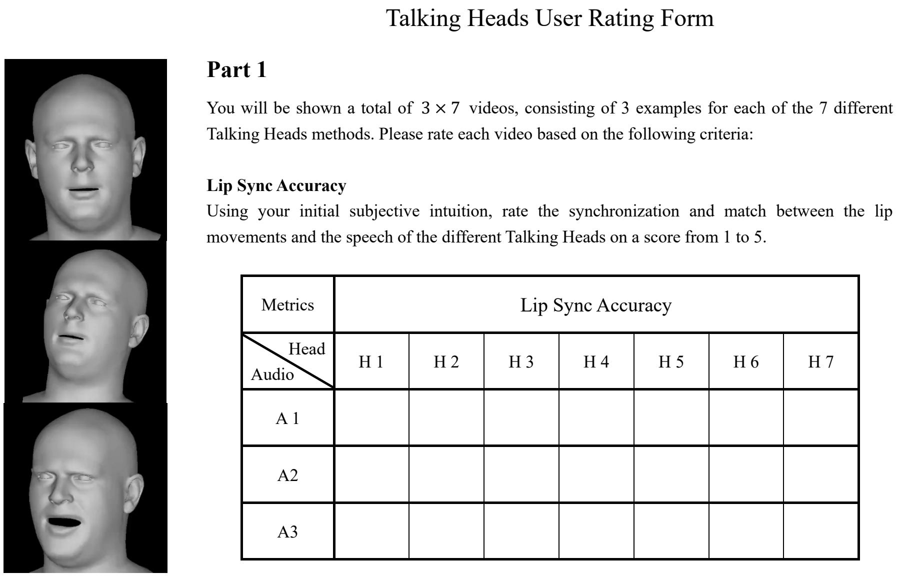
3 EXPERIMENTS DETAIL

3.1 User Study

For user study, we compare our results with 5 methods and ground truth by presenting three separate audio examples of each to the participants, one at a time. For the different audio tracks, the sequence in which the seven videos are presented will be randomized. This ensures that the order does not influence the assessment of each method's performance. Participants will be asked to rate each video based on the Lip Sync Accuracy and the Overall Realism on a score of 1 to 5. Regarding the diversity of head movements, we will present five different audio examples from each method simultaneously to the participants. They will be asked to compare the diversity of head movements across these examples and rate them on a score of 1 to 5.

3.2 FID Extractor

When applying the concept of FID to head pose generation problems, there is no readily available universal feature extractor for head pose data. Therefore, we trained an autoencoder-based feature extractor that can be trained in an unsupervised manner. The feature extractor consists of a convolutional encoder and decoder. The encoder encodes a series of direction vectors d into latent feature, and then the decoder reconstructs the original pose sequence from the latent feature. Specifically, the encoder consists of three one-dimensional convolutional layers, three linear layers, and two batch normalization layers. The first two one-dimensional convolutions maintain the temporal dimension unchanged, while the final one-dimensional convolution performs downsampling in the temporal dimension. The decoder network mirrors the encoder, with one-dimensional convolutions replaced by transposed one-dimensional convolutions.



Talking Heads User Rating Form

Part 1

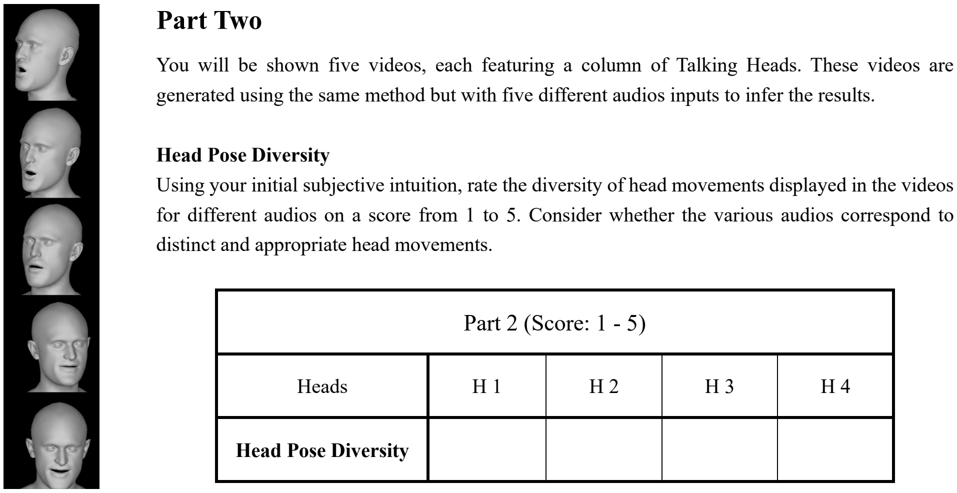
You will be shown a total of 3×7 videos, consisting of 3 examples for each of the 7 different Talking Heads methods. Please rate each video based on the following criteria:

Lip Sync Accuracy

Using your initial subjective intuition, rate the synchronization and match between the lip movements and the speech of the different Talking Heads on a score from 1 to 5.

Metrics	Lip Sync Accuracy						
Head Audio	H 1	H 2	H 3	H 4	H 5	H 6	H 7
A 1							
A2							
A3							

Figure 1: User Study of Lip Sync Accuracy. Each row of the table represents different types of audio, and each column of the table represents different methods. The evaluation table for Overall Realism follows the same format, where participants are asked to rate both of them after watching one video segment.



Part Two

You will be shown five videos, each featuring a column of Talking Heads. These videos are generated using the same method but with five different audios inputs to infer the results.

Head Pose Diversity

Using your initial subjective intuition, rate the diversity of head movements displayed in the videos for different audios on a score from 1 to 5. Consider whether the various audios correspond to distinct and appropriate head movements.

Part 2 (Score: 1 - 5)				
Heads	H 1	H 2	H 3	H 4
Head Pose Diversity				

Figure 2: Head Pose Diversity. Each column represents a specific method.