Appendices

A. Method Details

A.1. PQ-VAE

The PQ-VAE processes motion sequences, denoted as $M_{1:N}$, utilizing an encoder-decoder architecture tailored for temporal data [59]. The encoder comprises four residual blocks, each featuring a series of three temporal convolution layers. These layers are parameterized with kernel size, stride, and padding set to 3, 1, and 1, respectively, and are followed by batch normalization [25] and Leaky ReLU activation [39] for non-linear transformations. Additionally, temporal convolutions (kernel size 4, stride 2, padding 1) are placed between residual blocks to adjust the temporal resolution, setting the temporal window w to 8. A fully-connected layer precedes the quantization step, serving dimensionality reduction. The decoder mirrors the encoder's architecture, ensuring symmetry in information reconstruction.

A.2. Predictor

The Predictor leverages dual condition encoders and a transformer-based decoder. The audio encoder employs 3 temporal convolution layers (kernel size 4, stride 2, padding 1), followed by batch normalization [25] and Leaky ReLU activation [39], focusing on audio feature extraction. In contrast, the motion context encoder utilizes 10 gated convolution layers, catering to motion context comprehension [12, 43]. The transformer-based decoder consists of an embedding layer and six decoder blocks, each integrating a self-attention layer, a cross-attention layer, and a linear layer, followed by an AdaIN layer for style normalization. A fully-connected layer finalizes the structure, adjusting output dimensions to match the target motion specifications.

A.3. Refiner

Employing a transformer-based architecture akin to the Predictor's decoder (excluding the embedding layer), the Refiner fine-tunes motion predictions. During training phase, input to the Refiner combines ground truth (GT) motion $M^{gt}_{1:N}$ and PQ-reconstructed motion $M^{pq}_{1:N}$ via:

$$\overline{M}_{1:N} = I \odot M^{gt}_{1:N} + (1 - I) \odot M^{pq}_{1:N}, \tag{8}$$

where $\odot$ denotes element-wise multiplication. The Refiner outputs refined motion $M^{r}_{1:N}$, guided by input audio $A_{1:N}$, mask I and speaker identity $D_{1:N}$:

$$M^{r}_{1:N} = \text{Refiner}(\overline{M}_{1:N}; A_{1:N}, I, D_{1:N}). \tag{9}$$

The Refiner is optimized using the loss function $\mathcal{L}_{refine}$, which is formulated as follow:

$$\mathcal{L}_{refine} = \mathcal{L}_1(I \odot M^{gt}_{1:N}, I \odot M^{r}_{1:N}) + \mathcal{L}_1(V^{gt}_{1:N-1}, V^{r}_{1:N-1}). \tag{10}$$

Method	FGD ↓	MAE ↓	BC (GT=0.847)
Habibie et al.	44.60	8.59	0.964 (GT+0.117)
TalkShow	6.60	9.39	0.885 (GT+0.038)
ProbTalk (Ours)	3.98	7.79	0.818 (GT-0.029)

Table 7. More metrics.

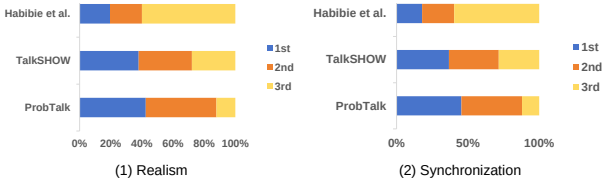

Figure 5. User study.

	FGD↓	BC	Diversity↑	MSE↓	LVD↓
Ground Truth	0	6.856	13.05	0	0
FaceFormer [13]	-	-	-	7.787	7.593
CodeTalker [56]	-	-	-	8.026	7.766
S2G [15]	28.15	4.683	5.971	-	-
Trimodal [61]	12.41	5.933	7.724	-	-
HA2G [36]	12.32	6.779	8.626	-	-
DisCo [33]	9.417	6.439	9.912	-	-
CaMN [34]	6.644	6.769	10.86	-	-
DiffStyleGesture [57]	8.811	7.241	11.49	-	-
Habibie et al. [21]	9.040	7.716	8.213	8.614	8.043
TalkShow [59]	6.209	6.947	13.47	7.791	7.771
EMAGE [35]	5.512	7.724	13.06	7.680	7.556
ProbTalk (Ours)	6.170	8.099	10.43	8.990	8.385

Table 8. Quantitative evaluation on BEAT-X.

Here, $\mathcal{L}_1$ is L1 reconstruction loss, while $V^{gt}_{1:N-1}$ and $V^{r}_{1:N-1}$ refer to the velocity of GT and generated holistic body motion, respectively.

B. More Comparison

More Metrics. We add the **Mean Absolute Error** (MAE) to quantitatively assess the difference between the ground truth and the motion generated by our model. Additionally, we introduced the **Beat Consistency** (BC) metric to evaluate the synchronization between the generated motion and the corresponding audio. The outcomes of these evaluations are presented in Tab. 7. The superior performance in both the MAE and BC metrics demonstrates that our model's generated outputs exhibit the highest degree of fidelity to the ground truth.

User Study. We conduct a user study to compare the realism and synchronization of our method with existing works. We randomly sample 20 audios. 12 participants were asked to rank videos generated by three methods based on their realism and synchronization. Results in Fig. 5 show that our method outperforms the others in both realism and synchronization.

Experiments on Beat-X Dataset. Our experimental evaluation conducted on the Beat-X dataset [35] is shown in Tab. 8. We follow the experimental configuration from [35],

Figure 6. Motion sequences selected based on their minimal Mean Squared Error (MSE), incorporating symmetry by treating movements of corresponding body parts (e.g., left and right hands) as equivalent in the evaluation.

employing data from **Speaker 2** for model training and validation. This ensures a direct comparison with benchmarks and aligns with prior research. However, it is important to acknowledge the potential for **overfitting** due to the limited data. As such, the outcomes of these experiments should be interpreted as indicative rather than definitive.

More Qualitative Comparison. To further validate the accuracy of the generated motions, we generate 32 samples using each method. Subsequently, we select the samples exhibiting the minimal Mean Squared Error (MSE) for each method. The corresponding results are presented in Fig. 6. The results demonstrate that our method generates a closer approximation to the true motion sequences, thus highlighting the enhanced accuracy of our generated samples.