

# Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model

Xu He<sup>1</sup> Qiaochu Huang<sup>1</sup> Zhensong Zhang<sup>2</sup> Zhiwei Lin<sup>1</sup> Zhiyong Wu<sup>✉,1,4</sup>  
 Sicheng Yang<sup>1</sup> Minglei Li<sup>3</sup> Zhiyi Chen<sup>3</sup> Songcen Xu<sup>2</sup> Xiaofei Wu<sup>2</sup>

<sup>1</sup> Shenzhen International Graduate School, Tsinghua University <sup>2</sup> Huawei Noah’s Ark Lab

<sup>3</sup> Huawei Cloud Computing Technologies Co., Ltd <sup>4</sup> The Chinese University of Hong Kong

{hex22, hqc22, lzw22, yangsc21}@mails.tsinghua.edu.cn zyw@sz.tsinghua.edu.cn  
 {zhangzhensong, liminglei29, chenzhiyi2, xusongcen, wuxiaofei2}@huawei.com



Figure 1. Examples of our generated gesture videos. White dashed arrows indicate gestures corresponding to bold words.

## Abstract

*Co-speech gestures, if presented in the lively form of videos, can achieve superior visual effects in human-machine interaction. While previous works mostly generate structural human skeletons, resulting in the omission of appearance information, we focus on the direct generation of audio-driven co-speech gesture videos in this work. There are two main challenges: 1) A suitable motion feature is needed to describe complex human movements with crucial appearance information. 2) Gestures and speech exhibit inherent dependencies and should be temporally aligned even of arbitrary length. To solve these problems, we present a novel motion-decoupled framework to generate co-speech gesture videos. Specifically, we first introduce a well-designed nonlinear TPS transformation to obtain latent motion features preserving essential appearance information. Then a transformer-based diffusion model is proposed to learn the temporal correlation between gestures and speech, and performs generation in the latent motion space, followed by an optimal motion selection module to produce long-term coherent and consistent gesture videos. For better visual perception, we further design a refinement network focusing on missing details of certain areas. Extensive experimental results show that our proposed framework significantly outperforms existing ap-*

*proaches in both motion and video-related evaluations. Our code, demos, and more resources are available at <https://github.com/thuhcsi/S2G-MDDiffusion>.*

## 1. Introduction

Co-speech gestures, as a typical form of non-verbal behavior [7], convey a wealth of information and play an important role in human communication. Appropriate gestures complement human speech and thus benefit comprehension, persuasion, and credibility [65]. Hence providing artificial agents with human-like and speech-appropriate gestures is crucial in human-machine interaction.

To achieve this goal, several methods have been developed for automatic co-speech gesture generation, with a particular focus on deep learning techniques. However, they mostly aim at generating gestures as 2D/3D human skeletons. While relatively easy to generate, skeletons totally discard appearance information and create a disparity with human perception [34]. As a result, they need to be further processed for better visualization. For example, some work binds skeletons to custom virtual avatars and manually renders them using software like Blender and Maya, consuming exhaustive human labor. Other studies [14, 41] train independent image synthesizers [4] to translate skeletons into animated images, which still rely on hand-crafted

annotations and yield noticeable inter-frame jitters.

Different from previous methods that only generate skeletons, we aim to generate audio-driven co-speech gesture videos directly in a unified framework, which is challenging due to the following two reasons: First, we need to find a suitable motion feature that can describe both intricate motion trajectories and complex human appearance. A straightforward way is to design a two-stage pipeline by first generating hand-crafted and pre-defined skeletons as motion features and then synthesizing animated images with them. However, skeletons only contain positions of sparse joints and will lead to texture loss and accumulated errors, making it unsuitable for our task. Another way is to customize popular conditional video generation methods [12, 22, 46, 61] to solve our problem. These methods usually encode videos into a latent space and then generate content within this space using UNet-based diffusion models [15, 17, 49, 68]. However, they primarily concentrate on general video generation with latent features derived from VAEs lacking well-defined meaning and struggling to filter and retain necessary video information effectively. Directly applying them to videos concerning human motion results in implausible movements and missing fine-grained parts [46]. Second, gesture videos should be temporally aligned with the input audio even of arbitrary length, while it is still difficult to capture the inherent temporal dependencies between gestures and speech. Besides, existing video generation methods [46, 78] can only generate videos of fixed length, for example, 2 seconds. Generating longer consistent videos is either time-consuming or even impossible, since it requires much more computational resources.

To address these challenges, in this paper, we propose a novel unified motion-decoupled framework for audio-driven co-speech gesture video generation. The overview of our method is shown in Fig. 2. To decouple motion from gesture videos while preserving critical appearance information of body regions, we first carefully design a thin-plate spline (TPS) [5, 77] transformation to model first-order motion, which is nonlinear and thus flexible enough to adapt to curved human body regions. To be specific, we predict several groups of keypoints to generate TPS transformations, subsequently employed for estimating optical flow and guiding image warping to generate corresponding gesture video frames. Note that, gathered keypoints are considered as latent motion features, which allow for the explicit modeling of motion while maintaining a small scale, easing the burden on the generation model. Then we introduce a transformer-based diffusion model for generation within the latent motion space, equipped with self-attention and cross-attention modules to better capture the temporal dependency between speech and motion. To further extend the duration of generated videos, we propose an optimal motion selection module, which considers both coherence and con-

sistency and helps to produce long-term gesture videos. Finally, for better visual quality, we present a UNet-like [45] refinement network supplemented with residual blocks [74] to capture local and global information of video frames, drawing more attention to certain regions and recovering missing details of appearance and textures.

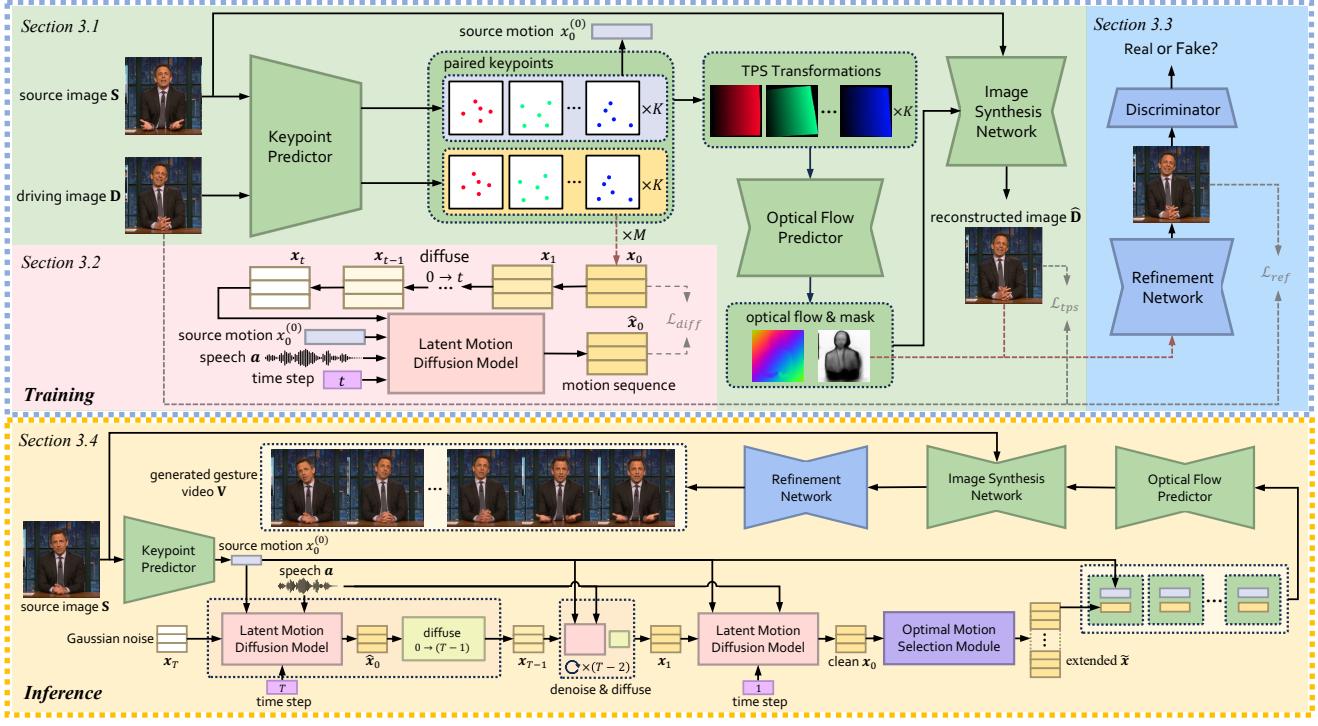
To summarize, the main contributions of our works are as follows:

- We present a novel motion-decoupled framework to directly generate co-speech gesture videos in an end-to-end manner independent of hand-crafted structural human priors, where a nonlinear TPS transformation is used to extract latent motion features and ultimately guide the synthesis of gesture video frames.
- We design a transformer-based diffusion model on latent motion features, capturing temporal correlation between speech and gestures, which is followed by an optimal motion selection module concerning coherence and consistency. With both modules, we can generate diverse long co-speech gesture videos.
- We introduce a refinement network to allocate additional attention to certain areas and enhancing appearance and texture details, which is crucial for human perception.
- Extensive experimental results show that our framework can generate vivid, realistic, speech-matched, and long-term stable gesture videos of high quality that significantly outperform existing methods.

## 2. Related Works

**Gesture generation on human skeletons.** Early works consider gesture generation as an end-to-end regression task [2, 36] and tend to generate averaged gestures without diversity. Subsequent insights into the many-to-many relationship between speech and gestures prompt the adoption of diverse generation methods including GANs [11], VAEs [25], and Flows [3]. More currently, diffusion models excel at modeling complex data distribution and have emerged as a promising approach to generate gestures [56, 63, 69, 76]. However, all these works depend on annotated datasets to generate human skeletons, including datasets labeled by pose estimators [1, 2, 71] and MoCap datasets [13, 33], suffering from error accumulation or insufficient data and totally devoid of appearance information. On the contrary, our framework generates gestures directly in the video form without relying on annotated skeleton priors.

**Gesture video generation.** To date, only a few works have made initial explorations into the problem of generating gesture videos directly. Zhou *et al.* [79] convert gesture video generation into a reenactment task and complete it in a rule-based way. They establish a motion graph with a reference video and search for a path matching the speech based on audio onset and a predefined keyword dictionary. However, it fails to generate novel gestures, and crafting



**Figure 2. Gesture video generation pipeline of our proposed framework** is composed of three core components: 1) the motion decoupling module (green) extracts latent motion features from videos with TPS transformations and synthesizes image frames; 2) the latent motion diffusion model (pink) generates motion features conditioned on the speech; 3) the refinement network (blue) restore missing details and produce the final fine-grained video.

rules is labor intensive. ANGIE [34] explicitly defines the problem of audio-driven co-speech gesture video generation, which utilizes an unsupervised feature, MRAA [51], to model body motion. Then a VQ-VAE [59] is leveraged to quantize common patterns, followed by a GPT-like network predicting discrete motion patterns to output gesture videos. However, as a coarse modeling of motion, MRAA is linear and fails to represent complex-shaped regions, limiting the quality of gesture videos generated by ANGIE. Differently, we carefully design a powerful latent motion feature and a matching generation model, enabling us to generate more realistic and stable gesture videos.

**Conditional video generation.** Another related task is conditional video generation. A variety of methods have been developed to generate videos conditioned on text [38], pose [22, 61], and also audio [46]. Recently, Diffusion models are used to model video space and exhibit promising results, but their computational requirements are often substantial due to the large volume of video data. Some works [15, 17, 49, 68] adopt an auto-encoder to create a latent space for videos and subsequently, diffusion generative models can focus solely on the latent space. However, these methods concentrate on generating general videos. The meaning of latent features is not well-defined, which may not always preserve the desirable information such

as human motion. While LaMD [20] attempts to use two auto-encoders to separate content and motion, the separation is implicit and relies entirely on the design of the encoder network architecture. Additionally, the motion is represented as a vector without the time dimension, which may cause failure to model spatio-temporal variations in human gestures. In contrast, we design a time-aware diffusion model performing generation in a well-designed latent motion space tailored for gesture video generation and hence can generate gesture videos of high quality.

### 3. Our Approach

Given a speech audio  $a$  and a source image  $S$  of the speaker, our framework aims to generate an appropriate gesture video  $V$  (*i.e.* an image sequence). Due to the rich connotation of gesture videos, our overall concept is to decouple and generate motion information as a bridge in the video generation process. Therefore, the pipeline can be formulated as  $V = \mathcal{F}(\mathcal{G}(\mathcal{E}(S), a), S)$ , where  $\mathcal{E}(\cdot)$  means motion decoupling to extract the source motion feature, which will be used with the audio as conditions to facilitate the audio-to-motion conversion by the diffusion model  $\mathcal{G}(\cdot)$ , and finally the image synthesis and refinement network  $\mathcal{F}(\cdot)$  accomplish the refined motion-to-video generation.

In the following parts, we first explain the motion de-

coupling module with TPS transformation, which learns latent motion features from videos and guides the source image to warp to synthesize image frames containing desired gestures (Sec. 3.1). Then we elaborate the transformer-based diffusion model to perform generation within the latent motion space (Sec. 3.2). After that, we introduce the refinement network for better visual perception which focuses more on details of specific areas (Sec. 3.3). Finally, we present the inference process of the entire framework, where the optimal motion selection module helps to produce coherent and consistent long gesture videos (Sec. 3.4).

### 3.1. Motion Decoupling Module with TPS

To decouple human motion from videos, a straightforward method is to extract 2D poses with off-the-shelf pose estimators [8, 70]. However, as a zeroth-order model, poses completely discards appearance information around keypoints, making precise motion control and video rendering highly challenging. Furthermore, pre-training of pose estimators relies on hand-crafted annotations, suffering from error accumulation and often yielding jitters. The early work ANGIE [34] proposes to use MRAA [51] consisting of mean and covariance, which is linear and fails to model regions with intricate shapes. Besides, it is inappropriate to associate covariance directly with speech. Summarizing the above, we argue that an effective representation to decouple motion is crucial for the quality of generated gesture videos and their matching with speech. Therefore, we design a motion decoupling module based on a nonlinear transformation named TPS transformation, which deals well with curving edges and hence can model the motion of various-shaped body regions. Next, we will start by introducing TPS transformation as preliminary, followed by an exposition of the entire motion decoupling module.

**TPS transformation.** TPS transformation [5] aims to establish the mapping  $\mathcal{T}_{tps}(\cdot)$  from the origin space  $\mathbf{D}$  to the deformation space  $\mathbf{S}$  by utilizing known paired keypoints as control, which takes the following form:

$$\begin{aligned} \mathcal{T}_{tps}(p) &= A \begin{bmatrix} p \\ 1 \end{bmatrix} + \sum_{i=1}^N w_i U(\|p_i^{\mathbf{D}} - p\|_2), \\ \text{s.t. } \mathcal{T}_{tps}(p_i^{\mathbf{D}}) &= p_i^{\mathbf{S}}, \quad i = 1, 2, \dots, N, \end{aligned} \quad (1)$$

where  $p = (x, y)^{\top}$  denotes coordinate.  $p_i^{\mathbf{D}}$  and  $p_i^{\mathbf{S}}$  are the  $i^{th}$  paired keypoints from the origin and deformation space.  $U(r) = r^2 \log r^2$  is a radial basis function.  $A \in \mathbb{R}^{2 \times 3}$  and  $w_i \in \mathbb{R}^{2 \times 1}$  are solvable parameters as introduced in [5].

In our setting, given a driving and a source image corresponding to the origin space  $\mathbf{D}$  and the deformation space  $\mathbf{S}$  separately, TPS transformation can establish local connections between the two frames, which will be further used to estimate a global optical flow  $\mathcal{T}(\mathbf{D}) = \mathbf{S}$  [77]. It serves as the foundation for our motion decoupling module and

offers two advantages: 1) as a flexible, non-linear transformation, it is suitable for modeling the motion of complex-shaped human bodies. 2) it relies solely on paired keypoints, whose movements are closely related to speech and thus can be more accurately controlled. Note that, unlike the keypoints of 2D poses only labeling certain joints, keypoints for TPS transformation come from adaptive boundary detection, involving both motion and crucial appearance information (*i.e.* region shapes), and can be easily used for operation at pixel level and further generating video frames.

The motion decoupling module is depicted as green in Fig. 2, which takes  $\mathbf{S}$  and  $\mathbf{D}$  as input, and outputs the constructed  $\hat{\mathbf{D}}$  for end-to-end self-supervised training.

**Keypoints predictor.** To generate TPS transformation, we first design a keypoint predictor to predict  $K \times N$  keypoints, which will subsequently be used for producing  $K$  TPS transformations with  $N$  points for each. The keypoints in  $\mathbf{S}$  and  $\mathbf{D}$  are estimated separately and then pairwise. The collection of keypoints  $\{p_{k,i}\}$  is very small in scale while being capable of generating a compact optical flow to animate images. So we take it as the latent motion feature.

**Optical flow predictor.** Now that we have  $K$  TPS transformations from predicted keypoint pairs modeling local motion, we can warp the source image  $\mathbf{S}$  to obtain  $K$  deformed images. The optical flow predictor processes the stacked deformed images and finally outputs a pixel-level optical flow indicating global motion. Following [77], occlusion masks are also predicted, which will be fed into the image synthesis network together with the optical flow.

**Image synthesis network.** Due to misaligned pixels and occlusions in  $\mathbf{S}$  and  $\mathbf{D}$ , direct warping fails to generate a valid reconstructed image  $\hat{\mathbf{D}}$ . Hence, we propose an image synthesis network of encoder-decoder architecture, with which  $\mathbf{S}$  is encoded into feature maps in different scales. The warping operation is performed on these feature maps, and occlusion masks then guide them to be masked. Subsequently, the decoder synthesizes the constructed image  $\hat{\mathbf{D}}$ .

**Training losses.** From previous work [50, 51, 77], we use the perceptual construction loss, equivariance loss, and warping loss to train the whole module in an unsupervised manner. The final loss is the sum of the above:

$$\mathcal{L}_{tps} = \mathcal{L}_{per} + \mathcal{L}_{eq} + \mathcal{L}_{warp}. \quad (2)$$

For more details about training and the architecture, please refer to our supplementary material.

### 3.2. Latent Motion Diffusion Model

Since we have decoupled motion from gesture videos, our idea is to employ a diffusion model [18, 54, 55] for generation in the latent space by denoising pure Gaussian noise. Given a real video clip, we utilize the trained keypoint predictor to obtain the keypoint sets for all frames as

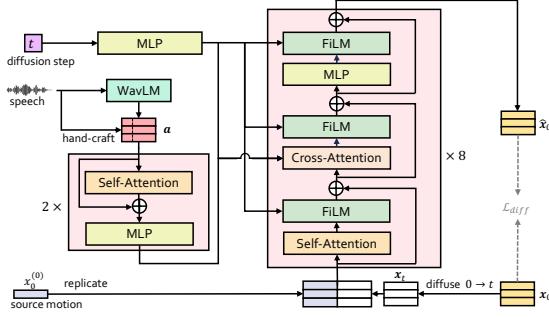


Figure 3. Noised motion features  $\mathbf{x}_t$  are concatenated with replicated source  $x_0^{(0)}$  and fed into our **transformer-based latent motion diffusion model** to predict the clean motion feature  $\hat{\mathbf{x}}_0$  conditioned on the audio feature  $\mathbf{a}$ . The attention mechanism captures inherent connections between latent motion features and speech.

$\{p_{ki} \in \mathbb{R}^2\}^{(1:M)}$ , where  $M$  is the frame number. We flatten the keypoints of each frame into a  $C = K \times N \times 2$ -dimensioned latent motion feature and finally get a feature sequence  $\mathbf{x}_0 = x_0^{(1:M)} \in \mathbb{R}^{M \times C}$ . Following [18],  $\mathbf{x}_0$  will be diffused  $t$  times to get noised  $\mathbf{x}_t$  and finally be cleaned.

**Model.** Per [43], our diffusion model predicts the clean motion feature sequence  $\hat{\mathbf{x}}_0$  from noised  $\mathbf{x}_t$  given noising step  $t$  and conditions  $\mathbf{c} = \{\mathbf{a}, x_0^{(0)}\}$ , where  $\mathbf{a}$  denotes the audio feature, and  $x_0^{(0)} \in \mathbb{R}^C$  is the source motion feature extracted from the source image  $\mathbf{S}$ , *i.e.* the first video frame.

During training,  $t$  is sampled from a uniform distribution  $\mathcal{U}\{1, 2, \dots, T\}$ , and noised sequence  $\mathbf{x}_t \in \mathbb{R}^{M \times C}$  is obtained by adding noise to  $\mathbf{x}_0$  following DDPM [18]. Concerning speech audio features, [69] reveals that WavLM [9] features contain semantic information and are beneficial to the generation of co-speech motion. So we stack features generated from WavLM Large [9] with hand-crafted audio features to form a complete speech audio feature  $\mathbf{a} \in \mathbb{R}^{M \times C_a}$ . The former is interpolated to be aligned with the latter temporally, and  $\mathbf{a}$  is also aligned with  $\mathbf{x}_t$ .

The latent motion diffusion model is in a transformer-like [57, 60] architecture as illustrated in Fig. 3, which is temporally aware and well-proven for modeling motion sequences [69]. The encoder takes the audio feature  $\mathbf{a}$  as input and yields hidden speech embeddings. The decoder is a transformer decoder equipped with feature-wise linear modulation (FiLM) [39]. The source motion feature  $x_0^{(0)}$  is replicated  $M$  times to have the same temporal length as  $\mathbf{x}_t$ , which are then concatenated together and fed into the self-attention network, capturing the temporal interactions within the motion sequence. After that, speech embeddings are projected to the cross-attention layer together with the output of self-attention, which facilitates learning the inherent relationship between the motion and speech sequence.

**Training losses.** We design the first term of loss to be common “simple” objective [18]. Besides, in the domain of

motion generation, geometric losses [48, 56] are commonly used, which serve to constrain physical attributes and promote naturalness and coherence. Concerning the discussion in Sec. 3.1 that our latent features represent the motion, it is natural and reasonable to introduce geometric losses within the latent space. Here we use losses for velocity [53, 56] and acceleration [53]. The final training loss is as follows:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{simple}} + \lambda_{\text{vel}} \mathcal{L}_{\text{vel}} + \lambda_{\text{acc}} \mathcal{L}_{\text{acc}}. \quad (3)$$

Details can be found in the supplementary material.

### 3.3. Refinement Network

Guided by the motion features, the image synthesis network can generate speech-matched image frames according to the optical flow. However, we observe that the synthesized frames exhibit some blurs with missing details, especially in two types of regions: 1) occluded areas labeled by the occlusion masks, and 2) regions with complex textures such as hands and the face. As the image synthesis network is jointly trained with the motion decoupling module, to address this issue without disrupting the balance of motion modeling, we propose an independent refinement network.

We use a Unet-like architecture [45] equipped with residual blocks [74] to capture both global and local information. To draw more attention to occluded areas, the synthesized image frame is concatenated with the mask of the corresponding resolution mentioned in Sec. 3.1 and then fed into our refinement network. Additionally, in order to focus more on certain regions, we utilize MobileSAM [75] to segment hands and the face, and assign larger weights to both hands, face, and occluded areas in L1 reconstruction loss. Please refer to our supplementary material for more details.

### 3.4. Inference

As shown in Fig. 2, given a source image  $\mathbf{S}$  and speech as inputs, keypoints of  $\mathbf{S}$  are first detected with the keypoint predictor and gathered to form the source motion feature  $x_0^{(0)}$ . Conditioned on  $x_0^{(0)}$  and extracted audio features  $\mathbf{a}$ , we randomly sample a Gaussian noise  $\mathbf{x}_T \in \mathbb{R}^{M \times C}$  from  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  and denoise it via the DDPM reverse process. At each time step  $t$ , the denoised sample is predicted as  $\hat{\mathbf{x}}_0 = \mathcal{G}(\mathbf{x}_t, t, \{\mathbf{a}, x_0^{(0)}\})$  and noised back to  $\mathbf{x}_{t-1}$ . After  $T$  steps, we obtain a clean sample  $\mathbf{x}_0$ . Repeating this procedure, we can get a consistent and coherent long sequence of motion features  $\tilde{\mathbf{x}}$  with a novel optimal motion selection module, detailed further below. For each frame of  $\tilde{\mathbf{x}}$ , we can rearrange it to get  $K \times N$  pairs of keypoints, producing  $K$  TPS transformations along with  $x_0^{(0)}$  to estimate optical flow and occlusion masks. They are then fed into the image synthesis network to generate image frames, which will go through the refinement network together with corresponding masks and finally convert into fine-grained results. All frames gather to form a complete co-speech gesture video.

**Optimal motion selection module.** For the fact that meaningful co-speech gesture units last between 4-15 seconds [6, 65], it is crucial to generate motion feature sequences of any desired length. However, the transformer-based diffusion model, designed for fixed-length inputs, struggles with direct sampling of longer noise for generation due to both poor performance and high computational costs. A naive solution is to generate fixed-length segments for concatenation, where the source motion feature  $x_0^{(0)}$  is replaced by the last frame of the previous segment to ensure continuity. However, in practice we notice that a single-frame condition cannot ensure the coherence and consistency between two segments, leading to flickers from position changes or jitters from direction changes of velocity.

To solve this problem, we propose an optimal motion selection module leveraging the diverse generative capability of the diffusion model, which operates solely at the inference stage. To be specific, from the second segment on, we generate  $P$  candidate sequences for the same audio segment. Then a lower-better score is calculated for each candidate according to two basic assumptions: within a small time interval of a real motion sequence, 1) keypoint positions are close; 2) keypoint velocity directions are similar. Finally, the candidate motion segment with the lowest score will be selected to extend the motion sequence. Details can be found in the supplementary material.

## 4. Experiments

### 4.1. Experimental Settings

**Dataset and preprocessing.** Data of our experiments is sourced from PATS dataset [1, 2, 14], consisting of transcribed poses with aligned audios and text transcriptions, containing around 84,000 clips from 25 speakers with a mean length of 10.7s, 251 hours in total. Similar to ANGIE [34], we perform our experiments on subsets of 4 speakers, including Jon, Kubinec, Oliver, and Seth. We download raw videos and audios to get clips according to PATS and conduct the following preprocessing steps: 1) Invalid clips with excessive audience applause, significant camera motion, or side views are excluded. 2) Clip lengths are limited to 4-15 seconds for meaningful gestures and resampled at 25 fps. 3) Frames are cropped with square bounding boxes, centering speakers, and resized to  $256 \times 256$ . 4) We extend these subsets with hand-crafted onset and chromagram features and WavLM [9] features. Finally, we obtain 1,200 valid clips for each speaker, randomly divided into 90% for training and 10% for evaluation, 4,800 in total.

**Evaluation metrics.** For motion-related metrics, we first extract 2D human poses with off-the-shelf pose estimator MMPose [47]. On this basis, we consider the quality, diversity, and alignment between gestures and speech, and choose: 1) **Fréchet Gesture Distance (FGD)** [72] to mea-

sure the distribution gap between real and generated gestures in the feature space, 2) **Diversity (Div.)** [36] which calculates feature distance between generated gestures on average. For these two metrics, we train an auto-encoder on poses from PATS. Also, we compute the average distance between closest speech beats and gesture beats as 3) **Beat Alignment Score (BAS)** following [28]. For video-related metrics, we utilize 4) **Fréchet Video Distance (FVD)** [58] to assess the overall quality of gesture videos. I3D [62] classifier pre-trained on Kinetics-400 [23] is used to compute FVD in the feature space.

### 4.2. Comparison to Existing Methods

We compare our method to: 1) the SOTA work ANGIE [34] in gesture video generation, and 2) MM-Diffusion [46], the SOTA work in video generation proven to be able to generate audio-driven human motion videos with experiments on AIST++ [28] human dance dataset.

The quantitative results are reported in Tab. 1. According to the comparison, our proposed approach significantly outperforms existing methods on motion-related metrics of FGD (56.44%) and Diversity (8.54%), which reveals that our motion-decoupled and diffusion-based generation framework is capable of generating realistic and diverse gestures in the motion space. Also, we achieve better performance on FVD than the best compared baseline MM-Diffusion, indicating that our method holds an advantage of ensuring the overall quality over the general audio-to-video method in gesture-specific settings. We notice that ANGIE with motion refinement tends to generate tremors synchronized with audio beat, leading to better results on BAS but at the expense of motion and visual quality. Fig. 4 presents frames of our generated videos compared with other methods, emphasizing the capacity of our method to generate videos with rich and realistic gestures matching the speech. On the contrary, limbs in ANGIE are modeled coarsely and vulnerable to abnormal deformations and absence from autoregressive error accumulation. MM-Diffusion struggles to capture body structures, leading to more or no hands.

Additionally, owing to the capability of TPS transformation to model complex-shaped regions and the close association between motion and speech established by the diffusion model, our method excels in generating precise and diverse fine-grained hand movements. As shown in Fig. 5, directly generated videos by MM-Diffusion entirely fail to produce reasonable hand morphology. While ANGIE attempts to utilize MRAA to represent motion, this linear affine transformation coarsely models curved body regions with Gaussian distribution, resulting in hand movements presented as the translation (controlled by the mean), rotation and scaling (controlled by PCA parameters of the covariance) of an “ellipse” in ANGIE’s results. In contrast, our method generates hand movements matching the

Table 1. Quantitative results on test set. Bold indicates the best and underline indicates the second. For ANGIE [34] we reproduce the code. For MM-Diffusion [46] we use the officially published code. Subjective evaluation is results of MOS with 95% confidence intervals.

Name	Objective evaluation					Subjective evaluation		
	FGD ↓	Div. ↑	BAS ↑	FVD ↓	Realness ↑	Diversity ↑	Synchrony ↑	Overall quality ↑
Ground Truth (GT)	8.976	5.911	0.1506	1852.86	4.76±0.05	4.70±0.06	4.77±0.05	4.73±0.06
ANGIE	55.655	5.089	<b>0.1504</b>	2965.29	2.07±0.08	2.53±0.08	2.19±0.08	2.00±0.07
MM-Diffusion	41.626	5.189	0.1098	<u>2656.06</u>	1.77±0.08	2.02±0.09	1.69±0.08	1.47±0.07
Ours	<b>18.131</b>	<b>5.632</b>	0.1273	<b>2058.19</b>	<b>3.79±0.08</b>	<b>3.91±0.07</b>	<b>3.90±0.08</b>	<b>3.77±0.07</b>

Table 2. Ablation study results. Bold indicates the best and underline indicates the second. ‘w/o’ is short for ‘without’.

Name	Objective evaluation					Subjective evaluation		
	FGD ↓	Div. ↑	BAS ↑	FVD ↓	Realness ↑	Diversity ↑	Synchrony ↑	Overall quality ↑
w/o TPS + MRAA	288.378	4.625	0.1200	3034.71	2.59±0.09	2.50±0.09	2.59±0.09	1.96±0.07
w/o WavLM	37.072	5.344	0.1253	<b>2053.44</b>	3.44±0.08	3.45±0.08	3.43±0.08	3.38±0.07
w/o refinement	26.125	5.549	<b>0.1288</b>	2154.00	3.67±0.08	<u>3.75±0.08</u>	<u>3.74±0.07</u>	3.49±0.06
LN Samp.	46.055	4.871	0.1250	2236.72	2.65±0.09	2.25±0.09	2.45±0.09	2.70±0.09
Concat.	<u>20.964</u>	<u>5.596</u>	0.1250	2085.50	3.66±0.07	3.64±0.08	3.71±0.08	<u>3.67±0.07</u>
Ours	<b>18.131</b>	<b>5.632</b>	0.1273	<u>2058.19</u>	<b>3.79±0.08</b>	<b>3.91±0.07</b>	<b>3.90±0.08</b>	<b>3.77±0.07</b>

speech, featuring intricate and plausible variations in hand shapes, which is crucial for high-quality human gestures.

**User study.** In practice, objective metrics may not always be consistent with human subjective perceptions [69], especially in the novel setting of co-speech gesture video generation. To gain further insights into the visual performance of our method, we conduct a user study to evaluate the gesture videos generated by each method alongside the ground truth. For each method, we sample 24 generated videos from the PATS test set between 3.2-12.8 seconds. 20 participants are invited to conduct the Mean Opinion Scores (MOS). Participants are asked to rate the videos in four aspects: 1) **Realness**, 2) **Diversity**, 3) **Synchrony** between speech and gestures, and 4) **Overall quality**. The first three focus on motion, while the last places more emphasis on visual perceptions. The rating scale ranges from 1 to 5 with a 1-point interval, where 1 means the poorest and 5 means the best. The results are reported in the last four columns in Tab. 1. Our method significantly surpasses other methods in all dimensions, which reveals that our framework can generate better gesture videos considering both motion and overall visual effects. It is noteworthy that the slight advantage of ANGIE on BAS does not translate into better gesture-speech synchrony in human subjective evaluation, where excessive tremors are not considered in sync with the speech. Please refer to the supplement for the effectiveness and robustness analysis of BAS and other objective metrics. According to the feedback from participants, “before seeing the ground truth”, our generated gesture videos are already “natural and well-matched to the speech enough to be mistaken as real”. Besides, there is an interesting finding that despite our emphasis on excluding irrelevant fac-

tors like textures and facial expressions in motion-related evaluations, participants express that “when compared with the ground truth containing rich details, although generated motion is realistic, they are inevitably influenced by appearance factors”. This demonstrates that human perception of motion and appearance are interrelated. Hence generating co-speech gesture videos with visual appearance is a meaningful problem in the field of human-machine interaction.

### 4.3. Ablation Study

We conduct an ablation study to demonstrate the effectiveness of different components in our framework. The results are shown in Tab. 2. We explore the effectiveness of the following components: 1) the TPS-based motion decoupling module, 2) WavLM features, 3) the refinement network, and 4) the optimal motion selection module.

Supported by the results in Tab. 2, when we replace TPS-based motion features with MRAA following ANGIE, FGD and FVD severely deteriorate by 1490% and 47.4%. When WavLM features are removed, FGD, Diversity, and BAS all deteriorate for the fact that WavLM features contain rich high-level information like semantics and emotions, crucial for driving gestures. However, WavLM brings a slight increase in FVD by 4.75, although not significantly (0.23%), demonstrating that the positive impact of WavLM is evident in motion while having subtle effects in visual aspects. The refinement network brings improvements in FGD, Diversity, and FVD, especially FVD decreased by 95.81 (4.4%). Detailed analysis and visual comparisons of our ablation study can be found in the supplementary material.

For the optimal motion selection module, we replace it with two simple strategies to generate longer videos as men-



Figure 4. **Visual comparison with SOTAs.** Our method generates gestures with a broader range of motion (dashed boxes) matching both beats (green words) and semantics (purple words). Red boxes denote unrealistic gestures generated by ANGIE [34] and MM-Diffusion [46].



Figure 5. **Visualization results of fine-grained hand variations.** Our generated gesture videos are more plausible and diverse.

tioned in Sec. 3.4: 1) long noise sampling (LN Samp.), and 2) direct concatenation (Concat.). According to Tab. 2, our method equipped with the optimal motion selection module achieves the best performance across all dimensions.

**User study.** Similarly, we conduct a user study for ablations as described in Sec. 4.2. Results in Tab. 2 indicate that the final performance of our model decreases without any module. Consistent with our expectations, removing TPS has the most significant impact on the results of Realness. This reiterates the crucial significance of employing an appropriate motion feature to decouple motion. Besides, we also conduct another user study in the context of longer

video generation and report the results in the supplement.

## 5. Conclusion

In this paper, we present a novel motion-decoupled framework for co-speech gesture video generation without structural human priors. Specifically, we carefully design a nonlinear TPS transformation to obtain latent motion features, which describe motion trajectories while retaining crucial appearance information. Then, a transformer-based diffusion model is used within this latent motion space to model the intricate temporal relationship between gestures and speech, followed by an optimal motion selection module to generate diverse long gesture videos. Besides, a refinement network is leveraged to draw more attention to certain details and bring better visual effects. Extensive experiments demonstrate that our framework produces long-term realistic, diverse gesture videos appropriate to the given speech, and significantly outperforms existing approaches.

## Acknowledgments

This work is supported by National Natural Science Foundation of China (62076144), Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004) and Shenzhen Science and Technology Program (WDZC20220816140515001, JCYJ20220818101014030).

## References

- [1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020. [2](#), [6](#), [10](#)
- [2] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 248–265. Springer, 2020. [2](#), [6](#), [10](#)
- [3] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, pages 487–496. Wiley Online Library, 2020. [2](#)
- [4] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8340–8348, 2018. [1](#)
- [5] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. [2](#), [4](#), [1](#)
- [6] Peter Bull. Gesture: Visible action as utterance. *Journal of Language and Social Psychology*, 25(3):339–341, 2006. [6](#)
- [7] Judee K Burgoon, Thomas Birk, and Michael Pfau. Nonverbal behaviors, persuasion, and credibility. *Human communication research*, 17(1):140–169, 1990. [1](#)
- [8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. [4](#)
- [9] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. [5](#), [6](#), [7](#)
- [10] Philippe G Ciarlet and Pierre-Arnaud Raviart. A mixed finite element method for the biharmonic equation. In *Mathematical aspects of finite elements in partial differential equations*, pages 125–145. Elsevier, 1974. [1](#)
- [11] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. [2](#)
- [12] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. [2](#)
- [13] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, pages 206–216. Wiley Online Library, 2023. [2](#)
- [14] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. [1](#), [6](#), [10](#)
- [15] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022. [2](#), [3](#)
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [4](#)
- [17] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. [2](#), [3](#)
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [4](#), [5](#), [2](#), [3](#)
- [19] Li Hu et al. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. [9](#), [10](#)
- [20] Yaosi Hu, Zhenzhong Chen, and Chong Luo. Lamd: Latent motion diffusion for video generation. *arXiv preprint arXiv:2304.11603*, 2023. [3](#)
- [21] Qiaochu Huang, Xu He, Boshi Tang, Haolin Zhuang, Liyang Chen, Shuochen Gao, Zhiyong Wu, Haozhi Huang, and Helen Meng. Enhancing expressiveness in dance generation via integrating frequency and music style information. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8185–8189. IEEE, 2024. [3](#)
- [22] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023. [2](#), [3](#)
- [23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. [6](#)
- [24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [4](#)
- [25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. [2](#)
- [26] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In *26th international conference on intelligent user interfaces*, pages 11–21, 2021. [9](#)

- [27] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 792–801, 2023. 9
- [28] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv preprint arXiv:2101.08779*, 2(3), 2021. 6, 5, 7
- [29] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10234–10243, 2023. 4, 7
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5
- [31] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 9
- [32] Guilin Liu, Fitzsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 4
- [33] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, pages 612–630. Springer, 2022. 2
- [34] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems*, 35:21386–21399, 2022. 1, 3, 4, 6, 7, 8, 5, 9, 10
- [35] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 5
- [36] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 2, 6
- [37] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 9
- [38] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 3
- [39] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5
- [40] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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*, pages 484–492, 2020. 10
- [41] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11077–11086, 2021. 1, 5
- [42] Weize Quan, Ruisong Zhang, Yong Zhang, Zhifeng Li, Jue Wang, and Dong-Ming Yan. Image inpainting with local and global refinement. *IEEE Transactions on Image Processing*, 31:2405–2420, 2022. 3, 4
- [43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 5, 3
- [44] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 10
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 5, 3
- [46] Luhuan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 2, 3, 6, 7, 8, 5, 9
- [47] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 20(17):10032–10044, 2020. 6, 5
- [48] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40(1):1–15, 2020. 5
- [49] Gaurav Shrivastava and Abhinav Shrivastava. Diverse video generation using a gaussian process trigger. *arXiv preprint arXiv:2107.04619*, 2021. 2, 3
- [50] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 4, 2

- [51] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. [3](#), [4](#), [2](#), [6](#)
- [52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [2](#), [4](#)
- [53] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. [5](#), [3](#)
- [54] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. [4](#)
- [55] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. [4](#)
- [56] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. [2](#), [5](#), [3](#)
- [57] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. [5](#)
- [58] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. [6](#), [5](#)
- [59] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. [3](#)
- [60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [5](#)
- [61] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Li-juan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. [2](#), [3](#)
- [62] Xianyuan Wang, Zhenjiang Miao, Ruyi Zhang, and Shanshan Hao. I3d-lstm: A new model for human action recognition. In *IOP Conference Series: Materials Science and Engineering*, page 032035. IOP Publishing, 2019. [6](#)
- [63] Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Ditt-Yan Yeung, Qiang Xu, et al. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. *arXiv preprint arXiv:2403.13304*, 2024. [2](#)
- [64] Zilin Wang, Haolin Zhuang, Lu Li, Yinmin Zhang, Junjie Zhong, Jun Chen, Yu Yang, Boshi Tang, and Zhiyong Wu. Explore 3d dance generation via reward model from automatically-ranked demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 301–309, 2024. [3](#)
- [65] Jason R Wilson, Nah Young Lee, Annie Saechao, Sharon Hershenson, Matthias Scheutz, and Linda Tickle-Degnen. Hand gestures and verbal acknowledgments improve human-robot rapport. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22–24, 2017, Proceedings* 9, pages 334–344. Springer, 2017. [1](#), [6](#)
- [66] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8858–8867, 2019. [4](#)
- [67] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *arXiv preprint arXiv:2208.06677*, 2022. [4](#)
- [68] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. [2](#), [3](#)
- [69] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919*, 2023. [2](#), [5](#), [7](#)
- [70] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. [4](#)
- [71] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. [2](#)
- [72] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. [6](#), [5](#)
- [73] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 736–747, 2022. [7](#)
- [74] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. [2](#), [5](#), [3](#)
- [75] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. [5](#), [3](#)
- [76] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motondif-

- fuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2
- [77] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 2, 4
- [78] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2
- [79] Yang Zhou, Jimei Yang, Dingzeyu Li, Jun Saito, Deepali Aneja, and Evangelos Kalogerakis. Audio-driven neural gesture reenactment with video motion graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3418–3428, 2022. 2, 5

# Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model

## Supplementary Material

In the supplementary document, we will introduce the following contents: 1) details of **TPS transformation** (Sec. A); 2) more details of our proposed **framework** (Sec. B), including the motion decoupling module (Sec. B.1), the latent motion diffusion model (Sec. B.2), the refinement network (Sec. B.3), the optimal motion selection module (Sec. B.4), and other implementation details (Sec. B.5); 3) the selection of **objective metrics** (Sec. C); 4) more details and analysis of **comparison to existing methods** (Sec. D); 5) results and analysis of the **ablation study** (Sec. E); 6) capability of generating **long gesture videos** (Sec. F); 7) **user study** details (Sec. G); 8) analysis of the **robustness and effectiveness of objective metrics** (Sec. H); 9) **generalization ability** analysis (Sec. I); 10) **time and resource consumption** (Sec. J); 11) **limitations and future work** (Sec. K); 12) **dataset license** (Sec. L). Since more mathematical expressions are included, we choose a single-column format in this supplementary document instead of two-column for readability. All demos, code, and more resources can be found at <https://github.com/thuhcsi/S2G-MDDiffusion>.

### A. Details of TPS Transformation

In the main paper, we employ TPS transformation [5] to establish pixel-level optical flow relying solely on sparse keypoint pairs from driving and reference images, thereby achieving precise control over the motion of human body regions. This is the foundation of our approach to decoupling motion while retaining crucial appearance information. Here we give a more detailed explanation of TPS transformation.

TPS transformation is a type of image warping algorithm. It takes as input corresponding  $N$  pairs of keypoints  $(p_i^D, p_i^S), i = 1, 2, \dots, N$  (referred to as control points) from a driving image  $\mathbf{D}$  and a source image  $\mathbf{S}$ , and outputs a pixel coordinate mapping  $\mathcal{T}_{tps}(\cdot)$  from  $\mathbf{D}$  to  $\mathbf{S}$  (referred to as backward optical flow). This process is grounded in the foundational assumption that the 2D warping can be emulated through a thin plate deformation model. TPS transformation seeks to minimize the energy function necessary to bend the thin plate, all while ensuring that the deformation accurately aligns with the control points, and the mathematical formulation is as follows:

$$\begin{aligned} & \min \iint_{\mathbb{R}^2} \left( \left( \frac{\partial^2 \mathcal{T}_{tps}}{\partial x^2} \right)^2 + 2 \left( \frac{\partial^2 \mathcal{T}_{tps}}{\partial x \partial y} \right)^2 + \left( \frac{\partial^2 \mathcal{T}_{tps}}{\partial y^2} \right)^2 \right) dx dy, \\ & \text{s.t. } \mathcal{T}_{tps}(p_i^D) = p_i^S, \quad i = 1, 2, \dots, N, \end{aligned} \quad (\text{A1})$$

where  $p_i^D$  and  $p_i^S$  denotes the  $i^{th}$  keypoints paired in  $\mathbf{D}$  and  $\mathbf{S}$ . According to [5], it can be proven that TPS interpolating function is a solution to Eq. (A1):

$$\mathcal{T}_{tps}(p) = A \begin{bmatrix} p \\ 1 \end{bmatrix} + \sum_{i=1}^N w_i U(\|p_i^D - p\|_2), \quad (\text{A2})$$

where  $p = (x, y)^\top$  is the origin coordinate in  $\mathbf{D}$ , and  $p_i^D$  is the  $i^{th}$  keypoint in  $\mathbf{D}$ .  $U(r) = r^2 \log r^2$  is a radial basis function. Actually,  $U(r)$  is the fundamental solution of the biharmonic equation [10] that satisfies

$$\Delta^2 U = \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^2 U \propto \delta_{(0,0)}, \quad (\text{A3})$$

where the generalized function  $\delta_{(0,0)}$  is defined as

$$\delta_{(0,0)} = \begin{cases} \infty, & \text{if } (x, y) = (0, 0) \\ 0, & \text{otherwise} \end{cases}, \quad \text{and } \iint_{\mathbb{R}^2} \delta_{(0,0)}(x, y) dx dy = 1, \quad (\text{A4})$$

which means that  $\delta_{(0,0)}$  is zero everywhere except at the origin while having an integral equal to 1.

We use  $p_i^X = (x_i^X, y_i^X)^\top$  to denote the  $i^{th}$  keypoint in image  $\mathbf{X}$  (*i.e.*  $\mathbf{D}$  or  $\mathbf{S}$ ), and denote:

$$r_{ij} = \|p_i^D - p_j^D\|, \quad i, j = 1, 2, \dots, N,$$

$$K = \begin{bmatrix} 0 & U(r_{12}) & \cdots & U(r_{1N}) \\ U(r_{21}) & 0 & \cdots & U(r_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ U(r_{N1}) & U(r_{N2}) & \cdots & 0 \end{bmatrix}, \quad P = \begin{bmatrix} 1 & x_1^D & y_1^D \\ 1 & x_2^D & y_2^D \\ \vdots & \vdots & \vdots \\ 1 & x_N^D & y_N^D \end{bmatrix},$$

$$L = \begin{bmatrix} K & P \\ P^T & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} x_1^S & x_2^S & \cdots & x_N^S & 0 & 0 & 0 \\ y_1^S & y_2^S & \cdots & y_N^S & 0 & 0 & 0 \end{bmatrix}^\top.$$

Then we can solve the affine parameters  $A \in \mathcal{R}^{2 \times 3}$  and TPS parameters  $w_i \in \mathcal{R}^{2 \times 1}$  as:

$$[w_1, w_2, \dots, w_N, A]^\top = L^{-1}Y. \quad (\text{A5})$$

In fact, in Eq. (A2), the first term  $A \begin{bmatrix} p \\ 1 \end{bmatrix}$  is an affine transformation for the alignment in the linear space of paired control points  $(p_i^D, p_i^S)$ . The second term  $\sum_{i=1}^N w_i U(\|p_i^D - p\|_2)$  introduces non-linear distortions for elevating or lowering the thin plate. With both the linear and nonlinear transformations, TPS transformation allows for precise deformation which is important to describe the motion without discarding crucial appearance information in our framework.

## B. More Details of Our Proposed Framework

### B.1. Motion Decoupling Module

**Training losses.** The motion decoupling module is trained end-to-end in an unsupervised manner. From previous works [50, 51, 77], we use a pretrained VGG-19 network [52] to calculate the perceptual construction loss in different resolutions as the main driving loss:

$$\mathcal{L}_{per} = \sum_j \sum_i \left| \text{VGG19}_i(\text{DS}_j(\mathbf{D})) - \text{VGG19}_i(\text{DS}_j(\hat{\mathbf{D}})) \right|, \quad (\text{B6})$$

where  $\text{VGG19}_i$  means the  $i^{th}$  layer of the VGG-19 network, while  $\text{DS}_j$  represents  $j$  downsampling operations. Also, equivariance loss is used to enhance the stability of the keypoint predictor as:

$$\mathcal{L}_{eq} = \left| E_{kp}(\tilde{\mathcal{A}}(\mathbf{S})) - \tilde{\mathcal{A}}(E_{kp}(\mathbf{S})) \right|, \quad (\text{B7})$$

where  $E_{kp}$  is the keypoint predictor, and  $\tilde{\mathcal{A}}$  is a random geometric transformation operator.

In addition, as introduced in [77], we also encode  $\mathbf{D}$  into feature maps with the encoder of the image synthesis network, compared with warped reference feature maps to calculate the warping loss:

$$\mathcal{L}_{warp} = \sum_i \left| \tilde{\mathcal{T}}^{-1}(E_i(\mathbf{S})) - E_i(\mathbf{D}) \right|, \quad (\text{B8})$$

where  $E_i$  is the  $i^{th}$  layer of the encoder of the image synthesis network, and  $\tilde{\mathcal{T}}^{-1}$  denotes the inverse function of the estimated optical flow, *i.e.* the forward optical flow from  $\mathbf{R}$  to  $\mathbf{D}$ .

The final loss is the sum of the above terms:

$$\mathcal{L}_{tps} = \mathcal{L}_{per} + \mathcal{L}_{eq} + \mathcal{L}_{warp}. \quad (\text{B9})$$

### B.2. Latent Motion Diffusion Model

**Framework.** The framework of our latent motion diffusion model is based on DDPM [18], where diffusion is defined as a Markov noising process.  $\mathbf{x}_0 \sim p(\mathbf{x})$  is sampled from the real data distribution (*i.e.*  $\mathbf{x}_0$  is a sequence of latent motion features drawn from a real gesture video). Given constant hyper-parameters  $\alpha_t \in (0, 1)$  decreasing with  $t$ , the forward diffusion process is to add Gaussian noise to the sample:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I}). \quad (\text{B10})$$

When the maximum time step  $T$  is sufficiently large and  $\alpha_t$  is small enough, we can use standard Gaussian distribution  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  to approximate  $\mathbf{x}_T$ . This indicates that it is possible to estimate real posterior  $q(\mathbf{x}_{t-1} | \mathbf{x}_t)$  following the reverse denoising process:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \quad (\text{B11})$$

where  $\mu_\theta(\cdot)$  and  $\Sigma_\theta(\cdot)$  mean estimating the mean and covariance via a neural network with learnable parameters  $\theta$ . From DDPM [18], the network predicts the noise  $\epsilon_\theta(\mathbf{x}_t, t)$  and thus we can use  $\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(\mathbf{x}_t, t) \right)$  added by randomly sampled noise to estimate  $\mathbf{x}_{t-1}$ . In our context, we take speech audio and the seed motion feature of the reference frame as conditions  $c$ , and aim to model the conditional distribution  $p_\theta(\mathbf{x}_0 | c)$  by gradually removing the noise. Following [43], we predict  $\mathbf{x}_0$  itself instead of noise  $\epsilon$ . The neural network of the diffusion network can be represented as  $\hat{\mathbf{x}}_0 = \mathcal{G}(\mathbf{x}_t, t, c)$ .

**Training losses.** We follow [18] to use *simple* objective as the first term of losses:

$$\mathcal{L}_{\text{simple}} = E_{\mathbf{x}_0 \sim q(\mathbf{x}|c), t \sim [1, T]} \left[ \|\mathbf{x}_0 - \mathcal{G}(\mathbf{x}_t, t, c)\|_2^2 \right]. \quad (\text{B12})$$

Besides, as mentioned in the main paper, we use the velocity loss and the acceleration loss to constrain the physical attributes of the motion features that describe the trajectories of the keypoint movements. Velocity and acceleration are respectively defined as the first and second-order time derivatives of the keypoint positions, and here, differential methods are employed to represent derivatives [21, 53, 56, 64]:

$$\mathcal{L}_{\text{vel}} = \frac{1}{M-1} \sum_{m=1}^{M-1} \left\| \left( \mathbf{x}_0^{(m+1)} - \mathbf{x}_0^{(m)} \right) - \left( \hat{\mathbf{x}}_0^{(m+1)} - \hat{\mathbf{x}}_0^{(m)} \right) \right\|_2^2, \quad (\text{B13})$$

$$\mathcal{L}_{\text{acc}} = \frac{1}{M-2} \sum_{m=1}^{M-2} \left\| \left[ \left( \mathbf{x}_0^{(m+2)} - \mathbf{x}_0^{(m+1)} \right) - \left( \mathbf{x}_0^{(m+1)} - \mathbf{x}_0^{(m)} \right) \right] - \left[ \left( \hat{\mathbf{x}}_0^{(m+2)} - \hat{\mathbf{x}}_0^{(m+1)} \right) - \left( \hat{\mathbf{x}}_0^{(m+1)} - \hat{\mathbf{x}}_0^{(m)} \right) \right] \right\|_2^2. \quad (\text{B14})$$

The final training loss is as follows:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{simple}} + \lambda_{\text{vel}} \mathcal{L}_{\text{vel}} + \lambda_{\text{acc}} \mathcal{L}_{\text{acc}}. \quad (\text{B15})$$

**Guidance.** Following [56], we train our diffusion model with classifier-free guidance. In training, we randomly mask the speech audio with a certain probability of 25%, *i.e.* replacing the condition  $c = \{\mathbf{a}, \mathbf{x}_0^{(0)}\}$  with  $c_\emptyset = \{\emptyset, \mathbf{x}_0^{(0)}\}$ . Then, we can strike a balance between diversity and fidelity by weighting the two results with  $\gamma$ :

$$\hat{\mathbf{x}}_0 = \gamma \mathcal{G}(\mathbf{x}_t, t, c) + (1 - \gamma) \mathcal{G}(\mathbf{x}_t, t, c_\emptyset), \quad (\text{B16})$$

where we can use  $\gamma > 1$  for extrapolating to enhance the speech condition.

### B.3. Refinement Network

**Architecture details.** Inspired by [42], we use a Unet-like [45] architecture to restore missing details of synthesized image frames. In specific, we use eight “convolution – LeakyReLU – batch norm” downsampling blocks and eight “upsample – convolution – LeakyReLU – batch norm” upsampling blocks with long skip connections, which prevent the information loss during downsampling while maintaining a large receptive field. Additionally, we insert two residual blocks [74] into the final two layers respectively, whose shallow architecture leads to a small receptive field and processes the feature maps in a sliding window manner. Simultaneously possessing large and small receptive fields enables the refinement network to capture both global and local information, thus better recovering missing details. Also, to ensure authenticity, we employ a patch-based discriminator [42] trained with GAN discriminator loss  $\mathcal{L}_D$  for adversarial training. Both the ground truth and refined image are converted into feature maps, with each element being discriminated as real or fake.

**Training losses.** Firstly, we train the refinement network with the common L1 reconstruction loss. Note that, as mentioned in the main paper, we utilize MobileSAM [75] to segment hands and the face to get the masks, and assign larger weights to both hands, face, and occluded areas using the masks in L1 reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{valid}} + \lambda_{\text{occ}} \mathcal{L}_{\text{occ}} + \lambda_{\text{hand}} \mathcal{L}_{\text{hand}} + \lambda_{\text{face}} \mathcal{L}_{\text{face}}, \quad (\text{B17})$$

where we use the complement of the occlusion masks from the optical flow predictor to compute  $\mathcal{L}_{valid}$ .

Then similar to [32, 42, 66], VGG-16 [52] is used to compute the perceptual loss and style loss in the feature space as:

$$\mathcal{L}_{per} = \sum_i \left| \text{VGG16}_i(\mathbf{D}) - \text{VGG16}_i(\hat{\mathbf{D}}_{ref}) \right|, \quad (\text{B18})$$

$$\mathcal{L}_{style} = \sum_i \left| \text{VGG16}_i(\mathbf{D}) \cdot [\text{VGG16}_i(\mathbf{D})]^\top - \text{VGG16}_i(\hat{\mathbf{D}}_{ref}) \cdot [\text{VGG16}_i(\hat{\mathbf{D}}_{ref})]^\top \right|, \quad (\text{B19})$$

where  $\hat{\mathbf{D}}_{ref}$  and  $\mathbf{D}$  represent the refined image frame and the real image frame respectively.  $\text{VGG16}_i$  means the  $i^{th}$  layer of the VGG-16 network, and we select  $i = 5, 10, 17$  in this work. In addition, following [32, 42], the total variation (TV) loss is used as:

$$\mathcal{L}_{tv} = \sum_i \sum_j \left( \left| \hat{\mathbf{D}}_{ref}^{i+1,j} - \hat{\mathbf{D}}_{ref}^{i,j} \right| + \left| \hat{\mathbf{D}}_{ref}^{i,j+1} - \hat{\mathbf{D}}_{ref}^{i,j} \right| \right), \quad (\text{B20})$$

where  $\hat{\mathbf{D}}_{ref}^{i,j}$  denotes the  $(i, j)$  pixel of the refined image frame.

The final loss is the weighted sum of the above terms, along with GAN generator loss  $\mathcal{L}_G$ :

$$\mathcal{L}_{ref} = \mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{style}\mathcal{L}_{style} + \lambda_{tv}\mathcal{L}_{tv} + \lambda_G\mathcal{L}_G. \quad (\text{B21})$$

#### B.4. Optimal Motion Selection Module

We employ a segment-wise generation approach to generate motion feature sequences of arbitrary length. Inspired by [29], starting from the second segment, leveraging the diversity generation capability of diffusion, we generate  $P$  candidates for each segment conditioned on the current audio and the end frame of the preceding segment. The scores are computed using the last five frames of the preceding segment and the first five frames of the candidate.

Specifically, by reorganizing the motion features back into keypoint positions, we calculate two scores: 1) Position coherency score calculates the **L1 distance** between the mean positions of the preceding segment and all candidates over five frames. 2) Velocity consistency score calculates the **angle of velocity directions** in average between the preceding and candidate segments over five frames, where velocity is computed through the differential of position. These two scores are summed to obtain the final score. A lower final score indicates fewer abrupt changes in position and velocity direction between two segments, thereby reducing flickers and jitters. So the candidate segment with the lowest score is chosen to extend the motion feature sequence. The frames at the transition points are eventually filled using cubic spline interpolation.

#### B.5. Other Implementation Details

We train our overall framework on four speakers jointly in three stages. 1) For the motion decoupling module: The number of TPS transformations  $K$  is set to 20, each with  $N = 5$  paired keypoints. We select ResNet18 [16] as the keypoint predictor for its simplicity and modify its output dimension to  $20 \times 5 \times 2$  to match the number and dimension of keypoints. Following [77], the optical flow predictor and the image synthesis network are 2D-convolution-based and produce  $64 \times 64$  weight maps to generate optical flow and four occlusion masks of different resolutions (32, 64, 128, 256) to synthesize image frames. We conduct training using Adam optimizer [24] with learning rate of  $2 \times 10^{-4}$ ,  $\beta_1 = 0.5$ ,  $\beta_2 = 0.999$ . 2) For the latent motion diffusion model: Keypoints are gathered and unfolded into the motion feature  $x \in \mathcal{R}^{200}$  for each frame. Motion features and audios are clipped to  $M = 80$  frames (3.2s) with stride 10 (0.4s) for training. The 35-dimension hand-crafted audio features include MFCC, constant-Q chromagram, tempogram, on-set strength and on-set beat, which are concatenated with 1024-dimension WavLM features to form  $a \in \mathcal{R}^{1059}$ . For Eq. (B15), we set  $\lambda_{vel} = \lambda_{acc} = 1$  and use Adan optimizer [67] with learning rate of  $2 \times 10^{-4}$  and 0.02 weight decay for 3,000 epochs training. The maximum sampling step  $T$  is 50. 3) For the refinement network: We set  $\lambda_{occ} = 3$ ,  $\lambda_{hand} = \lambda_{face} = 5$  in Eq. (B17). Following the hyper-parameter search results in [32], we set  $\lambda_{per} = 0.05$ ,  $\lambda_{style} = 120$ ,  $\lambda_{tv} = 0.1$ , and  $\lambda_{GAN} = 0.1$  in Eq. (B21). Adam optimizer [24] with learning rate of  $2 \times 10^{-4}$ ,  $\beta_1 = 0.5$ ,  $\beta_2 = 0.999$  is used for the refinement generator and learning rate of  $4 \times 10^{-5}$  for the discriminator. The whole framework is trained on 6 NVIDIA A10 GPUs for 5 days. In inference,  $\gamma$  in Eq. (B16) is set to 2 for extrapolating to augment the speech condition. Candidate number  $P$  is set to 5 for the balance between quality and inference time.

#### C. Selection of Objective Metrics

As a relatively unexplored task, co-speech gesture video generation lacks effective means of objective evaluation. Pioneering work ANGIE [34] simplifies the evaluation process by degrading their generation framework to 2D human skeletons

Table D1. **Subjective evaluation results on test set with two generation schemes for MM-Diffusion.** Bold indicates the best and underline indicates the second. Results of MOS are presented with 95% confidence intervals. Only the favorable results of MM-Diffusion-C are reported in the main paper.

Name	Subjective evaluation			
	Realness $\uparrow$	Diversity $\uparrow$	Synchrony $\uparrow$	Overall quality $\uparrow$
Ground Truth (GT)	4.76 $\pm$ 0.05	4.70 $\pm$ 0.06	4.77 $\pm$ 0.05	4.73 $\pm$ 0.06
ANGIE	<u>2.07</u> $\pm$ 0.08	<u>2.53</u> $\pm$ 0.08	2.19 $\pm$ 0.08	<u>2.00</u> $\pm$ 0.07
MM-Diffusion-D	1.63 $\pm$ 0.09	1.98 $\pm$ 0.09	1.54 $\pm$ 0.08	1.46 $\pm$ 0.08
MM-Diffusion-C	1.77 $\pm$ 0.08	2.02 $\pm$ 0.09	1.69 $\pm$ 0.08	1.47 $\pm$ 0.07
Ours	<b>3.79</b> $\pm$ <b>0.08</b>	<b>3.91</b> $\pm$ <b>0.07</b>	<b>3.90</b> $\pm$ <b>0.08</b>	<b>3.77</b> $\pm$ <b>0.07</b>

before leveraging the objective metrics common in skeleton generation, which, however, only assesses the performance of the generation module in structural skeletons without considering the effectiveness of the entire framework for gesture video generation. [79] employs metrics such as LPIPS popular in image evaluation and MOVIE for video evaluation to assess gesture reenactment. However, these general visual metrics only operate in the pixel or pixel-derived feature space, neglecting the crucial body movements in gesture videos. Therefore, we propose to use both motion and video-related metrics to evaluate gesture videos. Specifically, we use **Fréchet Gesture Distance (FGD)** [72], **Diversity (Div.)** [35], and **Beat Alignment Score (BAS)** [28] to evaluate the motion quality, and use **Fréchet Video Distance (FVD)** [58] to evaluate the video quality.

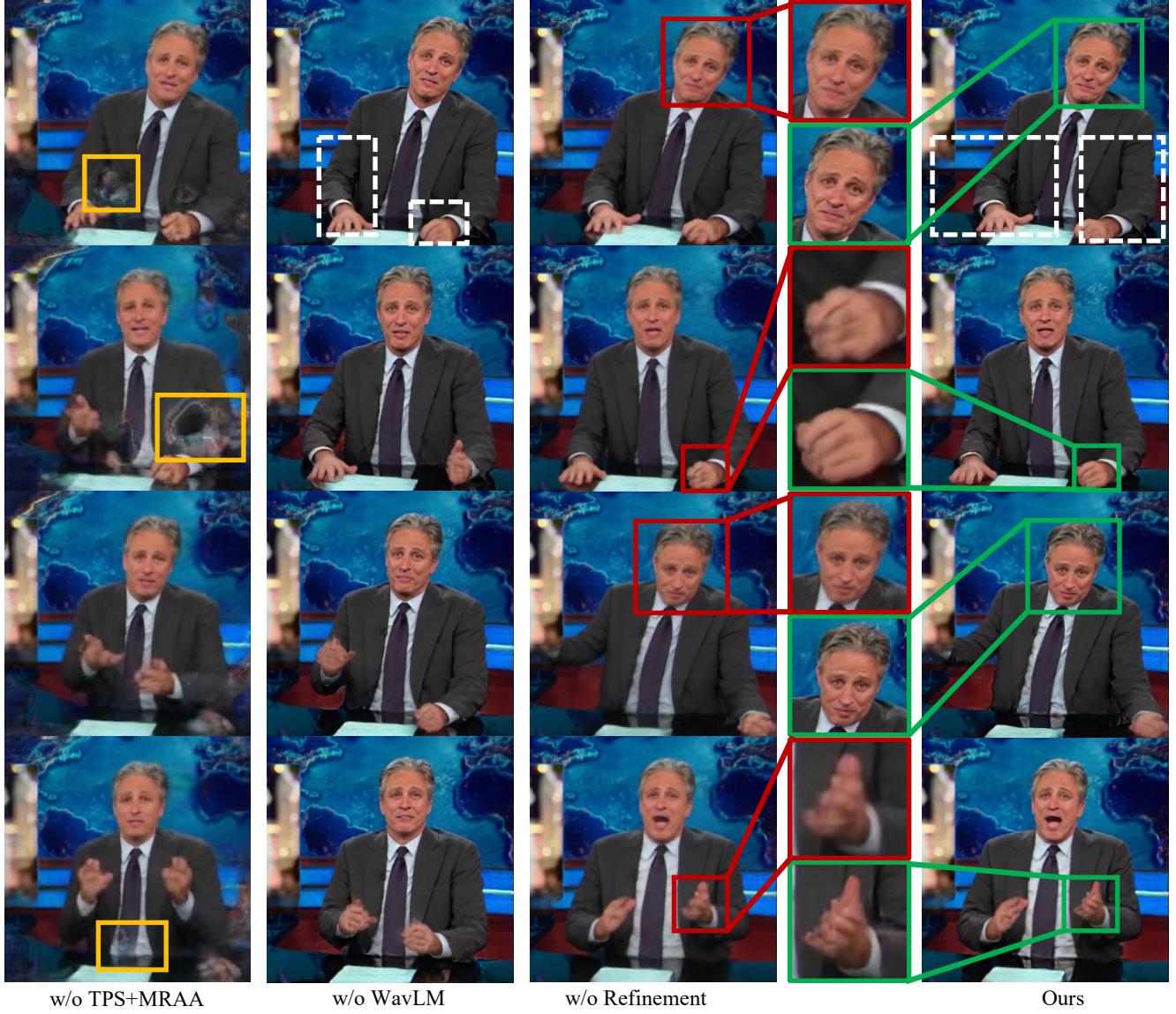
**Details of motion-related metrics.** We first extract 2D human poses with off-the-shelf pose estimator MMPose [47]. Extracting poses after generating gesture videos avoids the degradation of our original generation framework, allowing for effective measurements of the gesture motion quality in the videos. For the feasibility of calculating metrics, we performed normalization on raw poses: 1) We preserve 13 keypoints for the upper body and 21 keypoints for each hand, 55 keypoints in total [30, 47]. 2) We align the wrist points from body detection with those from hand detection. 3) For frames where the body is not detected, all keypoints are defined as centered at (128, 128). 4) For frames where hands are not detected, 21  $\times$  2 hand keypoints are assigned to the corresponding body wrist points.

Then, BAS can be directly computed using the audio and the normalized poses. For FGD and Diversity metrics, we follow [41] to train an auto-encoder on pose sequences from PATS train set to encode poses into a feature space. During training, pose sequences are clipped to 80 frames without overlapping. Each clip is then encoded into a 32-dimension feature. For FGD, we compute the Fréchet Distance between features of generated videos and all real videos, including both train set and test set. For Diversity, we calculate the average Euclidean distance of generated videos in the feature space following [35].

## D. Comparison to Existing Methods

As stated in the main paper, we compare our method with ANGIE [34] and MM-Diffusion [46]. For both our method and ANGIE, we use the audio and the initial frame image from PATS test set as inputs to generate corresponding 25fps gesture videos with a resolution of  $256 \times 256$ . Given that MM-Diffusion is trained solely conditioned on audio segments to generate 1.6s video segments of 10fps, we implement it with two generation schemes: 1) directly sampling long noise to generate videos of corresponding audio length (MM-Diffusion-D) and, 2) generating 1.6s segments for concatenation (MM-Diffusion-C). For both schemes, the generated gesture videos are resampled to 25fps. Additionally, considering that our method and MM-Diffusion-C generate fixed-length sub-clips (3.2s and 1.6s respectively) to form the full videos, both ground truth and generated videos are cropped to multiples of 3.2s for fair comparison.

User study results, including both of the two generation schemes of MM-Diffusion, are presented in Tab. D1. Due to space constraints, only the favorable results (MM-Diffusion-C) are reported in the main paper as “MM-Diffusion”. It is important to note that MM-Diffusion does not use the initial frame image as a condition, thus lacking control over the appearance of the speaker in the generated videos, resulting in inconsistent speakers between concatenated segments. So, in the user study, participants are instructed to evaluate the videos generated by MM-Diffusion-C only within each 1.6s segment, neglecting the overall quality of the full-length video. This, in fact, is a lenient evaluation for disregarding the inherent limitation of MM-Diffusion in generating consistently long videos. Nonetheless, the experimental results still demonstrate the superiority of our method over MM-Diffusion in all dimensions. Despite some setting differences, this concessive evaluation is sufficient to prove that our method surpasses MM-Diffusion when generating short segments in gesture-specific scenarios, not to mention the capability of our method to generate consistent long gesture videos.



**Figure E1. Visualization results of the ablation study.** Replacing TPS with MRAA leads to ghost effects (yellow boxes). WavLM brings greater amplitude of hand motion (dashed boxes) given an impassioned speech. Refinement restores the details especially in hands and the face (red and green boxes).

Constrained by computational resources and referring to the result of our user study in Tab. D1, only the favorable MM-Diffusion-C is used to generate 480 test videos for objective evaluation and reported as “MM-Diffusion” in the main paper.

## E. Ablation Study

Visualization results of the ablation study are shown in Fig. E1, where an impassioned speech is given as the condition. From the first column, we observe that the generated videos exhibit severe ghost effects (labeled by yellow boxes) when we replace the TPS-based motion features with MRAA [51]. We will give an explanation in the following part. According to [51], MRAA is a PCA-based affine transformation that represents motion features as the mean  $\mu$  and the covariance  $\Sigma$  of the probability distribution of body regions. While it is appropriate to infer  $\mu$  as the region translation from speech, the interaction between speech and the region shape represented by  $\Sigma$  is quite unclear. Unlike ANGIE [34] which uses a cross-condition GPT to connect  $\Sigma$  with  $\mu$  and speech, our diffusion model emphasizes the interactions between speech and motion

Table F2. **Results of generating long gesture videos.** Bold indicates the best and underline indicates the second.

Name	Effective duration $\uparrow$
Ground Truth (GT)	27.8s
ANGIE	4.1s
LN Samp.	3.5s
Concat.	<u>15.9s</u>
Ours	<b>21.0s</b>

features, with less focus on relating  $\Sigma$  to  $\mu$ . Thus the prediction of  $\Sigma$  is unstable. Although we impose constraints on  $\Sigma$  to be symmetric positive definite using Cholesky decomposition as mentioned in [34] for valid gestures, it still tends to output near-singular matrices, resulting severe errors in heatmaps for the estimation of the optical flow and occlusion masks. This, in turn, causes undesirable visual effects.

The second column shows the results of removing WavLM [9] features with only hand-crafted audio features used. Given an impassioned speech, the generated gestures with WavLM display greater amplitude and heightened intensity, because WavLM contains rich high-level information such as emotions and semantics [69]. The final three columns of Fig. E1 show that textures are restored after refinement, especially in hands and the face.

Please refer to our [homepage](#) for more visualization results of comparison with other methods and the ablation study.

## F. Capability of Generating Long Gesture Videos

To better assess the effectiveness of the optimal motion selection module and the capability of our framework to generate long gesture videos, we conduct another user study following [29]. We sample 10 long audios from the original PATS dataset as conditions to generate videos of 28s, and compare the generated results of 1) our complete framework, 2) long noise sampling (LN Samp.), 3) direct concatenation (Concat.), 4) ANGIE, and 5) the ground truth. 20 participants are asked to evaluate the effective duration of the videos, *i.e.* to decide how many seconds of the videos are effective. The average effective duration for each method is shown in Tab. F2. The results show that, although based on an easy-to-make hand-crafted rule, the optimal motion selection module benefits our method to generate longer videos with better coherency and consistency compared to only seed motion used and other methods. Directly sampling long noise and the autoregressive generation approach of ANGIE both face challenges in generating effective videos over 10 seconds.

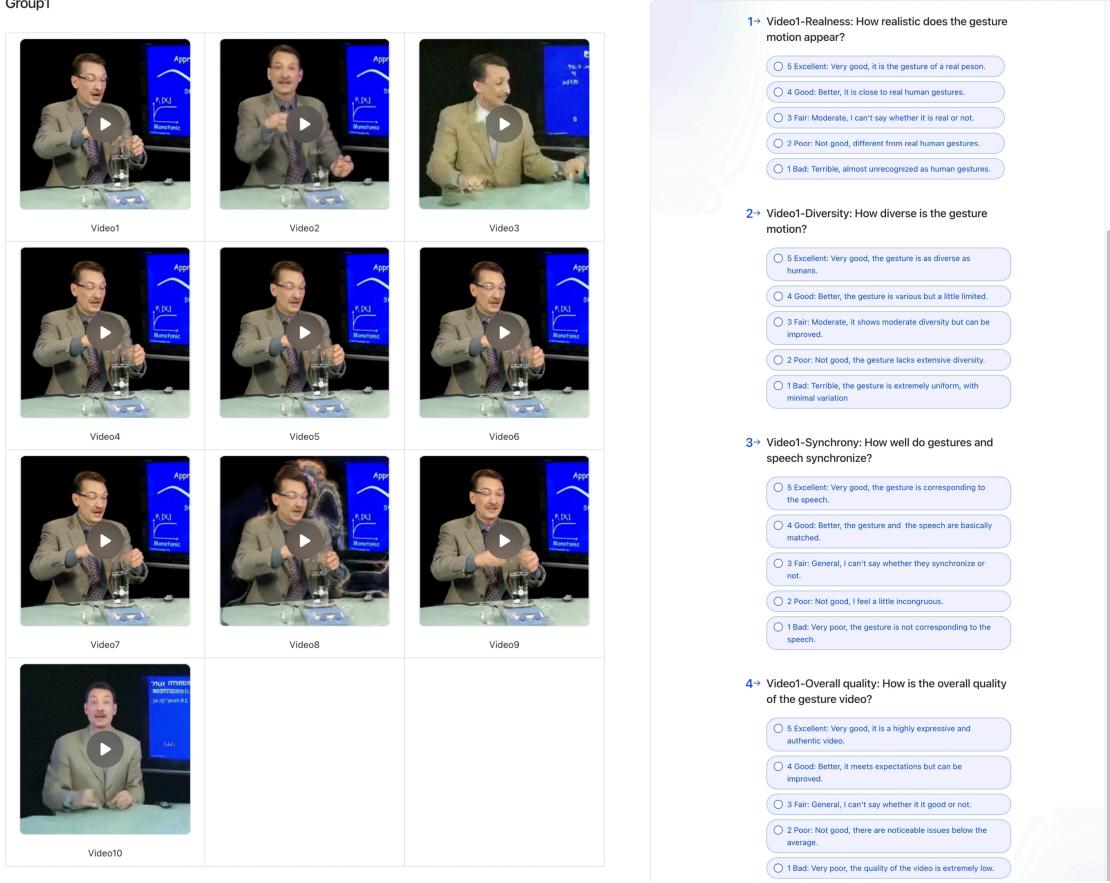
## G. Details of User Study

The user study is conducted by 20 participants with good English proficiency, involving 15 males and 5 females. Each participant is remunerated about 15 USD for a rating of 40-50 minutes, which is approximately at the average wage level [73]. Screenshots of the rating interface used for comparison, the ablation study, and the evaluation of long video generation are presented in Fig. G2.

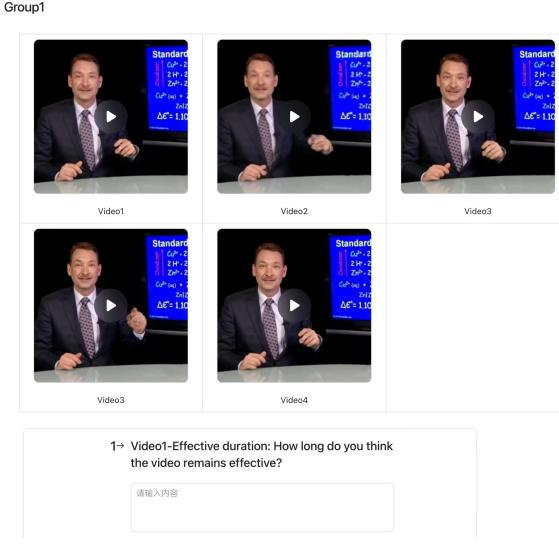
## H. Robustness and Effectiveness of Objective Metrics

From the main paper, we observe: 1) ANGIE [34] achieves higher BAS than ours. 2) Refinement brings lower BAS. 3) Sampling long noise and concatenation strategies have similar BAS. All these observations regarding BAS are inconsistent with subjective perceptions. Actually, BAS considers the distance between each audio beat with its nearest gesture beat, while gesture beats are defined as local velocity minima of 2D pose sequences filtered with a Gaussian kernel [28]. In practice, we encounter unavoidable inter-frame jitters when extracting 2D poses for evaluation with the off-the-shelf pose estimator. Tremors such as those in ANGIE, blurred images without refinement, or almost stationary long noise sampling results could amplify the jitters of estimated poses and cause incorrect identification as denser gesture beats, reducing the distance between gesture and speech beats and thus incorrectly increasing BAS, which can be seen from Fig. H3. In summary, BAS is susceptible to unrelated factors, making it a less robust objective metric. FGD, Diversity, and FVD are calculated in the feature space, making them somewhat more robust compared to BAS.

Another interesting finding is that despite other metrics of our method being closer to the GT, FGD still exhibits a noticeable discrepancy. However, user study results strongly indicate the authenticity of our generated motion. One plausible explanation is that for FGD, we take the entire data, including the training and testing sets, as the real reference to calculate distribution distances. Given the rich diversity of gestures, there are inherent distribution gaps between the training and test-



(a) User study interface for comparison and ablation.



(b) User study interface for rating effective duration.

Figure G2. Screenshots of the user study interface.

ing sets. Our model learns the data distribution from the training set, slightly deviating from the entire, while the GT of the testing set constitutes a portion of the overall distribution. This results in a noticeable difference in FGD. Referring to the training distribution reduces the difference (GT: 8.976 to 10.327 vs. ours: 18.131 to 13.285), providing supporting evidence.

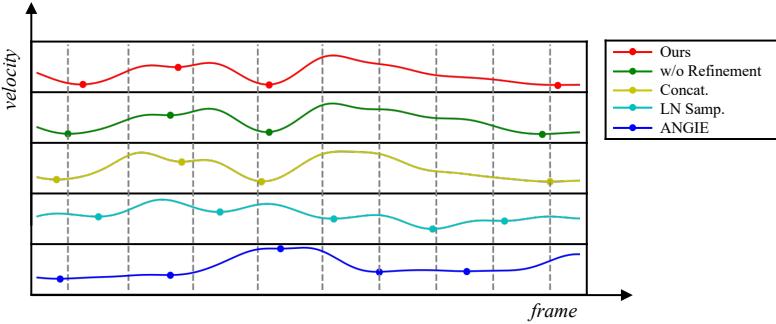


Figure H3. **Examples of velocity-frame curves of motion sequences** generated by each method for BAS analysis. For clear visualization, velocity is normalized and displayed without overlap. Dots represent gesture beats and dashed lines signify speech beats. ‘‘Concat.’’ is short for concatenation. ‘‘LN Samp.’’ is short for long noise sampling. ‘‘LN Samp.’’ and ‘‘ANGIE’’ exhibit more gesture beats but are not aligned with speech beats.

Actually, previous studies [26, 27] indicate that co-speech gesture generation still lacks objective metrics perfectly consistent with human subjective perception. To summarize the above, we have to demonstrate that subjective evaluation remains the gold standard for co-speech gesture video generation just like any other technology in the field of human-machine interaction [26].

## I. Generalization Ability

Gestures vary greatly between different speakers, so previous work typically trains an independent model for each person to capture individual styles. In contrast, we train a unified model jointly with the four speakers to ensure the scalability of our method. Experimental results indicate that even in this more challenging setting, our approach still generates gestures matching individual styles. Besides, we notice joint training brings about generalization ability to the speech of unseen speakers, which can be seen on our [homepage](#). However, it is still hard to generalize to any given portrait at present. Yet, given two critical facts: 1) our method can animate unseen dressing appearances of the four given speakers, for the dataset contains various appearances of the same speaker, and 2) efforts like [19] on extensive multi-person datasets show stronger generalization ability to unseen portraits, we believe that our approach exhibits generalization potential, and a high-quality multi-speaker gesture video dataset may help to enhance it, which will be explored in our future work.

## J. Time and Resource Consumption

Tab. J3 indicates that our training and inference time are comparable to ANGIE [34] and significantly shorter than MM-Diffusion [46]. Therefore, to the best of our knowledge, we achieve an optimal trade-off between time consumption and generation quality with distinct superiority in the latter. Although motion decoupling takes longer time, it greatly reduces the overall time and resource commitment compared to MM-Diffusion and other video generation works, *e.g.* [19] taking 14 days on 4 NVIDIA A100 GPUs for training<sup>1</sup>, providing a relatively efficient solution. Notably, our proposed diffusion model in the latent motion space achieves competitive generation results with relatively less time consumption, highlighting its necessity in the audio-to-motion process. Undeniably, repetitive diffusion denoising introduces extra inference time, and we will further explore methods like LCM [37] and Flow Matching [31] for acceleration.

Table J3. **Time consumption comparison** of training (6 NVIDIA A10 GPUs) and inference (1 NVIDIA GeForce RTX 4090 GPUs).

Name	Training	Training Breakdown	Inference
			(Generate a video of ~10 sec)
ANGIE	~5d	Motion Representation ~3d + Quantization ~0.2d + Gesture GPT ~1.8d	~30 sec
MM-Diffusion	~14d	Generation ~9d + Super-Resolution ~5d	~600 sec
Ours	~5d	Motion Decoupling ~3d + Motion Diffusion ~1.5d + Refinement ~0.5d	~35 sec

<sup>1</sup>Experimental results from our reproduced code instead of official resources.

## K. Limitations and Future Work

As research towards a relatively unexplored problem, there is still room for improvements in the following areas.

Despite significant superiority to existing methods, our generated videos still exhibit some accuracy issues of blurs and flickering, especially in hand details. This arises from the intricate structures of hands, characterized by varying movements like intersections and overlaps, which actually presents an unresolved challenge in the field of image and video generation [19, 44]. TPS-based motion decoupling effectively captures curved hand contours, making our method more adaptable to complex hand shapes than ANGIE [34], but still struggles to model structural details. The limited presence of hands in the frame drawing insufficient attention, coupled with the relatively weak inpainting capability of the image synthesis network, also leads to inaccurate hands. In addition, we observe that PATS dataset sourced from in-the-wild videos is of limited quality with noticeable hand motion blur, influencing the network’s performance to some extent. Therefore, in our future work, we will: 1) refine our method, *e.g.* prioritizing attention to hands and inpainting occlusion with more powerful pre-trained image generation models like SD model [44], and 2) collect high-quality gesture video data with clearer representations of hands to further enhance the generation quality.

Our current solution is unable to effectively synthesize the lip shape because there is a gap in the relationship between lips and gestures with speech. A unified framework for generating co-speech gestures and the lip shape simultaneously remains a valuable research problem, which we will explore in future work. In some showcases of the supplementary video, we use the off-the-shelf Wav2Lip [40] to synthesize lip shapes. Note that, the lip shape is not within the scope of this work, and generating lip shapes is just for better visual effects in the demo video.

For videos of bad quality, the accuracy of 2D poses from the pose estimator is compromised, leading to significant uncertainty when calculating all objective metrics regarding motion, especially BAS. Up until now, human subjective evaluation remains the most effective means of assessing generated gesture videos. Further exploration is needed to develop more robust and effective objective metrics.

## L. Dataset License

We download the YouTube videos and perform preprocessing according to the video links in the metadata provided by the PATS dataset [1, 2, 14]. Video license “CC BY - NC - ND4.0 International” allows for non-commercial use. Although the video data includes personal identity information, we adhere to the data usage license, and our processed data, models, and results will be used only for academic purposes and not be permitted for commercial use.