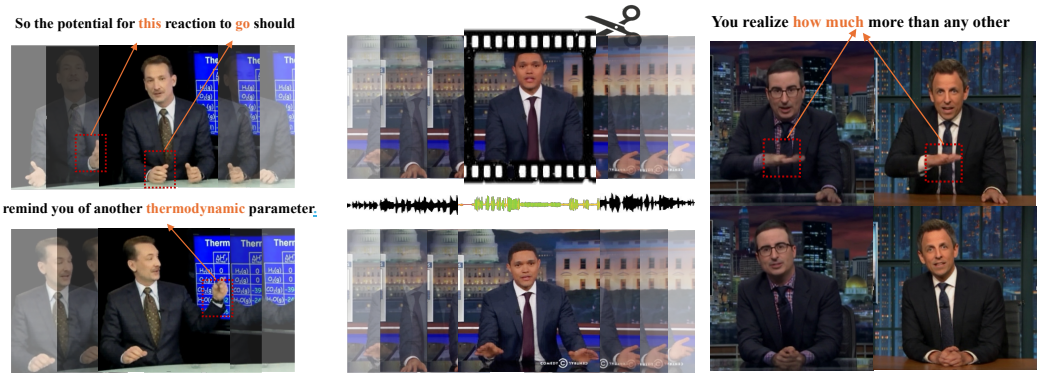
REALISTIC-GESTURE: CO-SPEECH GESTURE VIDEO GENERATION THROUGH CONTEXT-AWARE GESTURE REPRESENTATION

Pinxin Liu[1], **Pengfei Zhang**[2], **Hyeongwoo Kim**[3], **Pablo Garrido**[4], **Ari Sharpio**[4], **Kyle Olszewski**[4]
[1] University of Rochester, [2]University of California, Irvine, [3] Imperial College, London, [4] FlawlessAI
[1]pliu23@u.rochester.edu, [2]pengfz5@uci.edu, [3]hyeongwoo.kim@imperial.ac.uk,
[4]{pablo.garrido, ari.shapiro, kyle.olszewski}@flawlessai.com



Long Sequence Generation | Video Gesture Editing | Gesture Pattern Transfer

Figure 1: Realistic-Gesture achieves various fine-grained control over video-level gesture motion.

ABSTRACT

Co-speech gesture generation is crucial for creating lifelike avatars and enhancing human-computer interactions by synchronizing gestures with speech in computer vision. Despite recent advancements, existing methods often struggle with accurately aligning gesture motions with speech signals and achieving pixel-level realism. To address these challenges, we introduce Realistic-Gesture, a groundbreaking framework that transforms co-speech gesture video generation through three innovative components: (1) a speech-aware gesture tokenization that incorporate speech context into motion pattern representation, (2) a mask gesture generator that learns to map audio signals to gestures by predicting masked motion tokens, enabling bidirectional contextually relevant gesture synthesis and editing, and (3) a structure-aware refinement module that employs differentiable edge connection to link gesture keypoints to improve video generation. Our extensive experiments demonstrate that Realistic-Gesture not only produces highly realistic and speech-aligned gesture videos but also supports long-sequence generation and video gesture editing applications.

1 INTRODUCTION

In human communication, speech is often accompanied by gestures that enhance understanding and convey emotions De Ruiter et al. (2012). As these non-verbal cues play a vital role in effective interaction Burgoon et al. (1990), gesture generation a key component of natural human-computer interactions. As artificial intelligence advances, equipping virtual avatars with realistic gesture capabilities will become essential in creating immersive interactive experiences.

The relationships between the semantic and emotion content of speech context, the corresponding gestures, and the visual appearance of the speaker's performance are complex. As such, many recent works Yi et al. (2023); Liu et al. (2023; 2022d;a) address a reduced form of this problem by generating a simplified representation of the 3D motion, consisting of joints and body parts, that plausibly accompanies a given speech sample, which can then be rendered using standard rendering pipelines. Such representations capture basic motion patterns, yet they neglect the importance of the speaker's visual appearance, resulting in a lack of realism that hinders effective communication.

Other works, *e.g*. ANGIE Liu et al. (2022c) and S2G-diffusion He et al. (2024), employ image-warping techniques, constrained by keypoints obtained from optical-flow-based deformations, for co-speech video generation. However, such approaches encounter several critical issues. First, these keypoints only define large-scale transformations, and thus miss subtle movements of specific body parts (*e.g*. hands, fingers). Second, it is difficult to connect such broad and unconstrained motions representation to speech content. This makes it difficult to conditionally generate gestures that are appropriately responsive to the audio, which inhibits the gestures' naturalism and expressivity. Finally, the generated motion patterns are often unstructured and overly reactive to large motion, resulting in noisy and imprecise renderings, especially in the hands and shoulders. Collectively, these challenges significantly limit the overall quality and realism of the generated video content.

To address these challenges, we introduce *Realistic-Gesture*, a framework designed to generate speech-aligned gesture motions and high-fidelity speech video outputs. Our approach begins with refined gesture motion representations using keypoints from pretrained human pose estimators, allowing for clearer disentanglement of human motions across the face, body, and hands. To uncover the intrinsic temporal connections between gestures and speech, we employ contrastive learning to align these two modalities. This joint representation captures the triggers of gesture patterns influenced by speech. We incorporate speech-contextual features into the tokenization process of gesture motions through knowledge distillation, aiming to infuse the gesture representations with implicit intentions conveyed in the audio. This integration creates a clear linkage between the gestures and the corresponding speech, enabling the conditional generation of gestures that accurately reflect the speaker's intended meaning based on the speech input. For latent motion generation, inspired by Muse Chang et al. (2023) and MAGE Li et al. (2023), we introduce a masked gesture generator that refines the alignment of gesture motions with the speech signal through bidirectional mask pretraining, enabling long sequence generation and editing capabilities. Finally, for uplifting the latent motion generation into 2D animations, we propose a structure-aware image refinement module that generates heatmaps of edge connections from keypoints, providing image-level supervision to improve the quality of body regions with large motion. Extensive experiments demonstrate that our method outperforms the existing state-of-the-art approaches in both quantitative and qualitative metrics.

In summary, our primary contributions are:

1. a *speech-aware gesture motion representation* obtained through knowledge distillation from the gesture-speech aligned features from contrastive learning;
2. a *masked gesture motion generator*, carefully designed to enable high-quality gesture motion generation with long sequence generation and edit-ability support; and
3. a *pixel-level refinement module*, which uses a structure-aware edge heatmap as supervision to improve the final output fidelity.

2 Related Work

Co-speech Gesture generation Most recent works on co-speech gesture generation employ skeleton- or joint-level pose representations. Ginosar et al. (2019) use an adversarial framework to predict hand and arm poses from audio, and leverage conditional generation Chan et al. (2019) based on pix2pixHD Wang et al. (2018) for videos. Some recent works Liu et al. (2022d); Deichler et al. (2023); Xu et al. (2023) learns the hierarchical semantics or leverage contrastive learning to obtain joint audio-gesture embeding to assist the gesture pose generation. Rhythmic gesticulator Ao et al. (2022) construct high and low level audio-motion embedding based on lingustic theory for gesture generation. TalkShow Yi et al. (2023) estimates SMPL Pavlakos et al. (2019) poses, and models the body and hand motions for talk-show videos. CaMN Liu et al. (2022b) and EMAGE Liu et al. (2023) use large conversational and speech datasets for joint face and body modeling with

diverse style control. ANGIE Liu et al. (2022c) uses unsupervised 2D keypoints with image-warping features based on MRAA Siarohin et al. (2021) to model body motion. It leverages Vector Quantization van den Oord et al. (2018) to obtain common patterns, followed by a GPT-like network that outputs co-speech gesture videos. S2G-Diffusion He et al. (2024) uses TPS Zhao & Zhang (2022) and optical flow prediction to extract and refine latent motion features from videos. However, none of these works produce structure- and speech-aware motion patterns that are suitable for achieving natural and realistic gesture rendering.

Conditional Video Generation Conditional Video Generation has undergo significant progress for various modalities, like text Blattmann et al. (2023), pose Karras et al. (2023); Wang et al. (2023b), and audio Ruan et al. (2023). Diffusion Models Ho et al. (2020) improve generation qualities. AnimateDiff Guo et al. (2024) presents an efficient motion adaptation module based on low-rank adaptation Hu et al. (2022) (LoRA) to adapt image diffusion model for video motion generation. AnimateAnyone Hu et al. (2023) construct referencenet for fine-grained control based on skeleton. Make-Your-Anchor Huang et al. (2024) improves avatar video generation through disentangled face and body based on SMPL-X conditions. Champ Zhu et al. (2024) introduces human SMPL models for guidance. EMO Tian et al. (2024) leverages audio as control signal for talking head generation. However, these methods are based on large amount of training data and slow in inference speed. None of them focus on the speech-gesture pixel-level video generation.

Masked Representation Learning for Generation Masked Representation Learning has been demonstated an effective representation learning for various modalities. Devlin (2018); He et al. (2022) Some works explored the generation capabilities using this paradigm. MAGE Li et al. (2023) achieves high-quality image generation through iterative remasking. Muse Chang et al. (2023) extends this idea to leverage language with region masking for image editing and achieve fine-grained control. Recent Masking Models Pinyoanuntapong et al. (2024); Wang (2023); Mao et al. (2024) bring this strategy to the motion and gesture domain and improves the motion generation speed, quality, and editing capability. Inspired by these work, we propose the masked gesture generation conditioned the audio to learn the gesture-speech correspondence during generation.

3 PRELIMINARY

Warping-Based Image Animation. Warping-based image animation methods have risen to prominence recently Siarohin et al. (2021; 2019); Zhao & Zhang (2022). They leverage keypoint predictor to identify pairwise corresponding keypoints between a source image and a driving image. This information is used to warp the source image to match the driving image, thereby producing a deformation that aligns with the driving scene. Following this, pixel-level optical flow and occlusion masks are estimated from the deformed images to capture global motion and handle occlusions for achieving driving image reconstruction. We defer additional details to the Appendix.

Image-Animation Based Co-Speech Gesture Video Synthesis. In the context of co-speech gesture video synthesis, recent advancements have employed warping-based image animation techniques to derive motion patterns and learn the correspondence between these patterns and audio, facilitating speech-driven generation. Given a video clip $V = \{I_0, I_1, \ldots, I_N\}$ and an accompanying audio sequence $A = \{a_1, a_2, \ldots, a_N\}$, the objective is to predict motion representations $\hat{M}$ based on the initial frame I_0 and the audio input. The image animation module reconstructs all video frames $\hat{I}_1$ through unsupervised learning to derive motion representations and transformations. The audio sequence serves as guidance for reconstructing motion patterns across the entire sequence of frames following the initial frame.

However, this approach faces three significant challenges: (1) keypoints derived from global optical-flow-based transformations, learned through unsupervised methods, often fail to capture subtle movements of specific body parts; (2) the motion representations do not include contextual information from the speech, making it difficult to generate gestures that are conditionally responsive to audio; and (3) the lack of structural awareness in the motion representations leads to blurry and noisy predictions, particularly affecting the hands and shoulders, while also rendering the system sensitive to large motion patterns. To address these challenges, we propose the following methods to enhance control over co-speech gesture video generation.

4 REALISTIC-GESTURE

As shown in Fig. 2, our framework targets at generating realistic gesture videos. To achieve this goal, we first learn gesture-speech alignment to build speech-aware gesture motion representation through contrastive learning. (Sec. 4.1) To achieve fine-grained control over the gesture motion generation, we propose a Masking-based Gesture Generator, with long sequence and editing capabilities. (Sec. 4.2). To improve the noisy hand and shoulder movement during the uplifting of latent motion to pixel space, we propose a structure-aware image refinement through differentiable edge heatmaps for guidance. (Sec. 4.3).

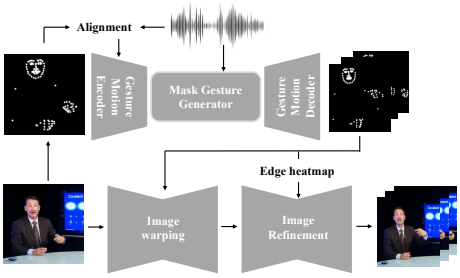


Figure 2: An overview of our framework.

4.1 SPEECH-AWARE GESTURE MOTION REPRESENTATION

Unlike ANGIE Liu et al. (2022c) and S2G-Diffusion He et al. (2024), which rely on unsupervised learning of keypoints for warping-based on optical flow, we utilize 2D poses extracted from images. While using 2D poses for image warping may slightly decrease fidelity, it significantly enhances the perceptual quality of the generated video, shown in Sec. 5.4. With poses, gestures can be decomposed into facial and body movements. We represent a gesture motion sequence as $G = [F; B] = [f_t; b_t]_{t=1}^{T}$, where T denotes the length of the motion, f represents the 2D facial landmarks, and b denotes the 2D body landmarks. Further details on gesture representations can be found in the appendix. For speech representation, we extract audio embeddings from WavLM Chen et al. (2022). In addition, we extract Mel spectrogram features Rabiner & Schafer (2010) and beat information using librosa McFee et al. (2015). These features are concatenated to form the speech representation. For image-warping transformation, we select TPS Zhao & Zhang (2022).

Speech-Gesture Alignment. To align gesture motion patterns with the content of speech and beats, we draw inspiration from image-language contrastive learning Radford et al. (2021). We first project both speech and gesture modalities into a shared embedding space to enhance the speech content awareness of gesture features. As illustrated in Fig. 3, we separately train two gesture content encoders, $\mathcal{E}_f$ for face motion and $\mathcal{E}_b$ for body motion, alongside two speech encoders, $\mathcal{E}_{S_f}$ and $\mathcal{E}_{S_b}$, to map face and body movements and speech signals into this joint embedding space. For simplicity, we represent the general gesture motion sequence as G. We then apply mean pooling to aggregate content-relevant information from each feature sequence, resulting in the embeddings z^s and z^g for speech and gestures, respectively. We leverage CLIP-style contrastive learning to train these content encoders. Given a batch of paired embeddings $\mathcal{B} = \{(z_i^t, z_i^g)\}_{i=1}^{B}$, we optimize the following loss with τ as the temperature:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{B}\sum_{i=1}^{B}\left(\log\frac{\exp(z_i^g \cdot z_i^t/\tau)}{\sum_{j=1}^{B}\exp(z_i^g \cdot z_j^t/\tau)} + \log\frac{\exp(z_i^t \cdot z_i^g/\tau)}{\sum_{j=1}^{B}\exp(z_i^t \cdot z_j^g/\tau)}\right) \tag{1}$$

Unlike previous methods Ao et al. (2022); Liu et al. (2022d); Deichler et al. (2023), which primarily capture sequence-level alignment and may overlook local temporal dynamics, to mitigate this limitation, we randomly mask 30% of segments from both speech and gesture sequences within the same temporal regions during training. Furthermore, we apply a linear classifier on the gesture embedding to predict speech beats, enhancing the temporal alignment between gestures and speech. We defer additional details of temporal-level improvement by our strategy in the Appendix.

Speech-Pattern Learning Through Knowledge Distillation. For gesture motion tokenization, we utilize Residual Vector Quantization Lee et al. (2022) (RVQ) to capture the high diversity and complexity of facial and body motions. To construct context-aware motion representations, we directly encode alignment information into the gesture motion codebook. This allows the semantics and contextual triggers from speech (e.g., pronouns like "this" or "they") to be fused into the motion embedding, and enables the generator to easily identify the corresponding motion representation in response to speech triggers. To achieve this goal, we leverage gesture content encoder as the teacher and distill knowledge to codebook latent representation. We aim to maximize the cosine similarity

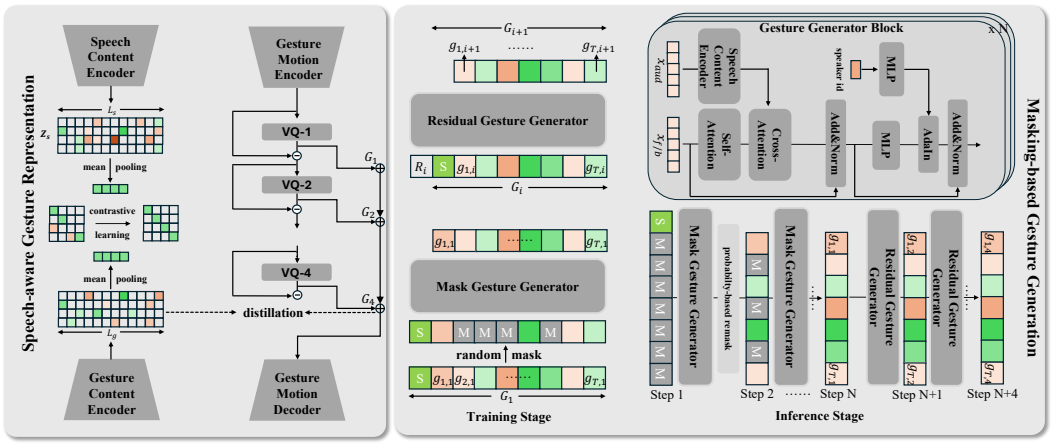


Figure 3: Left: Contrastive Learning for gesture-speech alignment. We distill the joint speech contextual-aware feature into latent codebook. Right: We use speech for generating gesture motion tokens with Mask Gesture Generator. We apply random mask for reconstruction during training and iterative remask based on probability for inference. Residual Gesture Generator based on the base VQ-tokens to predict the residuals.

over time between the final RVQ quantization output and the representation from the gesture content encoder, formulated as follows:

$$\mathcal{L}_{\text{distill}} = \sum_{t=1}^{T} \cos\left(p(Q_R)^t, Es(G)^t\right) \tag{2}$$

where p denotes a linear projection layer, Q_R is the final quantization output from the RVQ-VAE, $Es(G)$ represents the output from the gesture content encoder, and T is the total time frames. The overall training objective for the RVQ-VAE is defined as:

$$\mathcal{L}_{\text{rvq}} = \mathbb{E}_{x\sim p(x)}\left[\|x-\hat{x}\|^2\right] + \alpha \sum_{r=1}^{R} \mathbb{E}_{z_r\sim q(z_r|x)}\left[\|e_r - \text{sg}\left(z_r - e_r\right)\|^2\right] + \beta\mathcal{L}_{\text{distill}} \tag{3}$$

where $\mathcal{L}_{\text{rvq}}$ combines a motion reconstruction loss, a commitment loss van den Oord et al. (2018) for each layer of quantizer with a distillation loss, with α and β weighting the contributions.

4.2 Speech-conditioned Gesture Motion Generation

To enhance the generation of gesture motions across different layers of the quantized codebooks, we draw inspiration from VALL-E Wang et al. (2023a) to design **Masked Gesture Generator** for jointly decoding facial and body motions for the base-layer outputs of quantizers, and **Residual Gesture Generator** for the face and body tokens from the subsequent R residual quantization layers.

Masked Gesture Generator. As shown in Fig. 3, during training, we derive motion tokens by processing raw gesture sequences through both body and face tokenizers. The motion token corresponding to the source image acts as the conditioning for all subsequent frames. For speech control, we initialize the audio content encoder from alignment pre-training as described in Sec. 4.1. This pre-alignment of gesture tokens with audio encoder features enhances the coherence of gesture generation. We employ cross-attention, using the audio input as keys and values while the gesture representation serves as the query, integrating audio information with gesture feature. To refine control over gesture patterns, we apply Adaptive Instance Normalization (AdaIN) Huang & Belongie (2017) after the feed-forward layers, enabling diverse gesture styles based on the speaker's identity.

Residual Gesture Generator. The Residual Gesture Generator shares a similar architecture with the Masked Gesture Generator, but it includes R separate embedding layers corresponding to each RVQ residual layer. During training, we randomly select a quantizer layer $j \in [1, R]$ for learning. All tokens from the preceding layers $t^{0:j-1}$ are embedded and summed to form the token embedding input. After generating the base layer predictions of discrete tokens from the Masked Gesture Generator, these tokens are fed into the Residual Gesture Generator. This module iteratively predicts the tokens from the base layers, ultimately producing the final quantized output.

Inference. While existing works Liu et al. (2023); Yi et al. (2023); Chen et al. (2024) leverage auto-regressive next-token prediction or diffusion-based generation process, these strategies hinder the fast synthesis for real-time applications. To resolve this problem, as in Fig. 3, we employ an iterative mask prediction strategy to decode motion tokens during inference. Initially, all tokens are masked except for the first token from the source frame. Conditioned on the audio input, the Mask Gesture Generator predicts probabilities for the masked tokens. In the l-th iteration, the tokens with the lowest confidence are re-masked, while the remaining tokens stay unchanged for subsequent iterations. This updated sequence continues to inform predictions until the final iteration, when the base-layer tokens are fully generated. Upon completion, the Residual Gesture Generator uses the predicted base-layer tokens to progressively generate sequences for the remaining quantization layers. Finally, all tokens are transformed back into motion sequences via the RVQ-VAE decoder.

Training Objective. To train our gesture generation models, $\mathcal{L}_{mask}$, and $\mathcal{L}_{res}$ functions for two generaors respectively by minimizing the categorical cross-entropy loss, as illustrated below:

$$\mathcal{L}_{mask} = \sum_{i=1}^{T} -\log p_\phi(t_i | Es(S), \text{MASK}), \quad \mathcal{L}_{res} = \sum_{j=1}^{V}\sum_{i=1}^{T} -\log p_\phi(t_i^j | t_i^{1:j-1}, Es(S), j). \quad (4)$$

In this formulation, $\mathcal{L}_{mask}$ predicts the masked motion tokens t_i at each time step i based on the input audio and the special [MASK] token. Conversely, $\mathcal{L}_{res}$ focuses on learning from multiple quantization layers, where t_i^j represents the motion token from quantizer layer j and $t_i^{1:j-1}$ includes the tokens from preceding layers. We also feed the predicted tokens into the RVQ decoder for gesture reconstructions, with velocity and acceleration losses Tevet et al. (2022); Siyao et al. (2022).

4.3 Structure-Aware Image Refinement

To transfer gesture generation to pixel-level video synthesis, we leverage TPS Zhao & Zhang (2022) to achieve portrait animation based on gesture pattern keypoints from Sec. 4.2 through image warping. To address the uncertainties by optical-flow-based deformation, particularly in large motion regions such as the hands and shoulders, we propose a Semantic-Aware Generator. Auto-Link He et al. (2023) demonstrates that the learning of keypoint connections for image reconstruction aids the model in understanding image semantics. Based on this, we leverage keypoint connections as semantic guidance for image refinement.

Learnable Edge Heatmaps. Using the gesture motion keypoints, we establish linkages between them to provide structural information. To optimize computational efficiency, we limit the number of keypoint connections to those defined by body joint relationships Wan et al. (2017), rather than considering all potential connections in He et al. (2023).

For two keypoints $\boldsymbol{k}_i$ and $\boldsymbol{k}_j$ within predefined connection groups, we create a differentiable edge map $\boldsymbol{S}_{ij}$. This edge is modeled as a Gaussian function extending along the line connecting the keypoints. Formally, the edge map $\boldsymbol{S}_{ij}$ for keypoints $(\boldsymbol{k}_i, \boldsymbol{k}_j)$ is defined as:

$$\boldsymbol{S}_{ij}(\boldsymbol{p}) = \exp\left(v_{ij}(\boldsymbol{p}) d_{ij}^2(\boldsymbol{p})/\sigma^2\right), \quad (5)$$

where σ is a learnable parameter controlling the edge thickness, and $d_{ij}(\boldsymbol{p})$ is the L_2 distance between the pixel $\boldsymbol{p}$ and the edge defined by keypoints $\boldsymbol{k}_i$ and $\boldsymbol{k}_j$:

$$\boldsymbol{d}_{ij}(\boldsymbol{p}) = \begin{cases} \|\boldsymbol{p} - \boldsymbol{k}_i\|_2 & \text{if } t \le 0, \\ \|\boldsymbol{p} - ((1-t)\boldsymbol{k}_i + t\boldsymbol{k}_j)\|_2 & \text{if } 0 < t < 1, \\ \|\boldsymbol{p} - \boldsymbol{k}_j\|_2 & \text{if } t \ge 1, \end{cases} \quad \text{where} \quad t = \frac{(\boldsymbol{p} - \boldsymbol{k}_i)\cdot(\boldsymbol{k}_j - \boldsymbol{k}_i)}{\|\boldsymbol{k}_i - \boldsymbol{k}_j\|_2^2}. \quad (6)$$

Here, t denotes the normalized distance between $\boldsymbol{k}_i$ and the projection of $\boldsymbol{p}$ onto the edge.

To derive the edge map $\boldsymbol{S} \in \mathbb{R}^{H\times W}$, we take the maximum value at each pixel across all heatmaps:

$$\boldsymbol{S}(\boldsymbol{p}) = \max_{ij} \boldsymbol{S}_{ij}(\boldsymbol{p}). \quad (7)$$

We generate heatmaps at various resolutions. Inspired by SPADE Park et al. (2019), we treat these structural heatmaps as semantic guidance for image generation. A U-Net with residual blocks utilizes spatial semantic control from the edge heatmaps to refine the final video output.

Training Objective. We employ a conditional adversarial loss Mirza & Osindero (2014), along with perceptual similarity loss Johnson et al. (2016) and $L1$ loss for image refinement. The

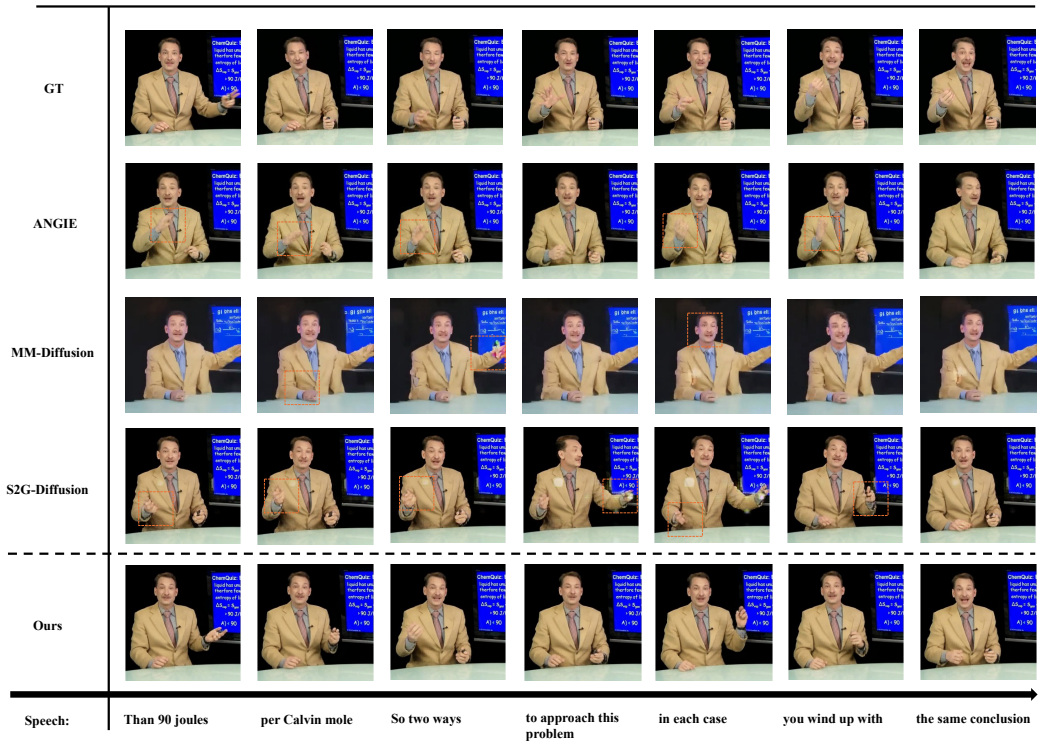


Figure 4: **Visual comparisons.** Our method generates high-quality hand and shoulder motions, and presents metaphoric gestures when saying *"90 joules,"* and *"in each case."*

discriminator utilizes the edge heatmap as a condition to compare generation against ground truth:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{I_{gt}, map}\left[\log D\left(I_{gt}, map\right)\right] + \mathbb{E}_{map}\left[\log\left(1 - D\left(map, I_{gan}\right)\right)\right], \tag{8}$$

where map denotes the edge heatmap, and $(\cdot, \cdot)$ indicates concatenation.

5 Experiments

Since our work focuses on joint gesture motion and video generation, the main experiments primarily compare our approach with existing methods that also address joint generation. Comparisons specifically for gesture motion generation and avatar video rendering are deferred to the Appendix, where they are treated as separate, disentangled modules and compared with relevant works.

5.1 Experimental Settings

Dataset and preprocessing. We utilize PATS Ginosar et al. (2019); Ahuja et al. (2020) for the experiments. It contains 84,000 clips from 25 speakers with a mean length of 10.7s, 251 hours in total. For a fair comparison, following the literature Liu et al. (2022c); He et al. (2024) and replace the missing subject, with 4 speakers are selected (*Noah*, *Kubinec*, *Oliver*, and *Seth*). All video clips are cropped with square bounding boxes, centering speaks, resized to 256×256. We defer the additional details in the Appendix. After filtering, we obtain around 1000 clips for each speaker, randomly divided into 90% for training and 10% for evaluation, 4,000 in total.

Baseline Methods. We benchmark Realistic-Gesture against several co-speech gesture video generation methods: (1) ANGIE Liu et al. (2022c), a work in co-speech gesture video synthesis; (2) MM-Diffusion Ruan et al. (2023), an audio-video generation model demonstrated on the AIST++ dataset Li et al. (2021) that produces audio-driven human motion videos; and (3) S2G-Diffusion He et al. (2024), the most recent advancement in this domain. Notably, due to MM-Diffusion's fixed generation of 34 frames, we segment the audio accordingly for each generation.

5.2 Quantitative Evaluation

Evaluation Metrics. We evaluate gesture motion and pixel-level video quality separately. For gesture motion metrics, we use **Fréchet Gesture Distance (FGD)** Yoon et al. (2020) to measure

Table 1: Quantitative results on the test set. Bold indicates the best performance. Our method performs better in terms of both gesture motions and video generation quality.

Name	Gesture-Motion Evaluation					Video Quality Assessment		
	FGD ↓	FGD-o ↓	Div. ↑	BAS ↑	PCM ↑	FVD ↓	VQA_A ↑	VQA_T ↑
Ground Truth	0.0	0.0	14.01	1.00	1.00	0.00	95.69	5.33
ANGIE	67.52	57.65	6.67	0.78	0.37	526.25	88.14	4.73
MM-Diffusion	137.62	132.33	3.21	0.65	0.11	-	79.56	4.24
S2G-Diffusion	23.65	15.44	10.85	0.97	0.45	486.13	93.55	5.40
Ours	1.30	1.21	13.26	0.99	0.57	476.12	96.33	6.08

the distribution gap between real and generated gestures in feature space, **Diversity (Div.)** Lee et al. (2019) to calculate the average feature distance between generated gestures, **Beat Alignment Score (BAS)** following Li et al. (2021), and **Percent of Correct Motion parameters (PCM)**, difference of generation deviate from ground-truth following Chen et al. (2024). We extract 2D human poses for face and body using MMPose OpenMMLab (2020), which differs from S2G-Diffusion that focuses solely on body poses.

For pixel-level video quality, we assess **Fréchet Video Distance (FVD)** Unterthiner et al. (2018) for the overall quality of gesture videos, $\mathbf{VQA}_A$ for aesthetics and $\mathbf{VQA}_T$ for technical quality based on Dover Wu et al. (2023), pretrained on a large-scale dataset with labels ranked by real users.

Evaluation Results. We present quantitative evaluations in Tab. 1. Our approach significantly outperforms existing methods in gesture motion metrics, achieving an FGD of **1.30** and a Diversity score of **13.26**. For video quality assessment, we use the FVD metric to evaluate the similarity of the generated video distribution to the ground-truth videos. Our model achieves the lowest FVD among the compared methods, demonstrating superior performance. The VQA_A and VQA_T metrics measure perceived user preferences for video generation content. Notably, our approach yields a VQA_A of **96.33** and a VQA_T of **6.08**, surpassing the ground-truth videos. This success can be attributed to our structure-aware image enhancement design. In contrast, MM-Diffusion produces limited gesture patterns due to its design, which generates only a few continuous frames and struggles to learn diverse motion patterns from speech audio. ANGIE leverages MRAA Siarohin et al. (2021) for regional coarse motion patterns but lacks the precision necessary for motion control aligned with speech, resulting in low diversity and beat alignment. S2G-Diffusion performs better than ANGIE but still fails to generate fine-grained gesture patterns, as it relies on image optical flows without adequately focusing on the nuances of human facial and body movements.

5.3 Qualitative Evaluation

Table 2: Subjective evaluation results are shown as Mean Opinion Scores (MOS).

Methods	MOS_1	MOS_2	MOS_3	MOS_4
	User Study			
GT	4.7	4.7	4.7	4.65
MM-Diffusion	1.35	1.65	1.4	1.55
ANGIE	1.95	3.25	1.9	2.25
S2G-Diffusion	3.0	**3.6**	3.15	3.0
Ours	**3.35**	3.05	**3.35**	**3.25**

Evaluation Results. We provide qualitative evaluations of video generation in Fig. 4. MM-Diffusion generates unrealistic shoulder movements. ANGIE produces misaligned gesture motions with the accompanying speech. Although S2G-Diffusion shows improvement over ANGIE, it struggles with local regions, such as the hands, due to its reliance on unsupervised keypoints for global transformations, which neglects local deformations. In contrast, our method demonstrates high-quality video generation, particularly in the facial and body areas. The alignment between gesture and speech is notably enhanced through our speech-content-aware gesture latent representation. For example, when the actor says *"90 joules,"* he points to the screen, and he emphasizes phrases like *"so two ways"* and *"in each case"* by raising his hands as metaphoric gestures. This coordination exemplifies our approach's capability to produce contextually relevant and expressive gestures.

User Study. We conducted a user study to evaluate the visual quality of our method. We sampled 80 videos from each method and ground-truth, and invited 20 participants to conduct Mean Opinion Scores (MOS) evaluations. The rating ranges from 1 (poorest) to 5 (highest). Participants rated the videos on: (1) MOS_1: *"How **realistic** does the video appear?"*, (2) MOS_2: *"How **diverse** does the gesture pattern present?"*, (3) MOS_3: *"Are speech and gesture **synchronized** in this video?"* and (4) MOS_4: *"What is your **overall evaluation** of the video"*. The videos were presented in random order to capture participants' initial impressions. As shown in Tab. 2, our method outperformed others

Table 3: Ablations of our method. We exam the keypoint design, gesture representation, gesture generator architecture, training & inference strategy and image-refinement. Bold indicates the best performance.

Kp Repr.	FVD↓	LPIPS↓	PSNR↑
Unsup-kp	387.05	0.05	27.41
2D-pose	272.18	0.05	27.26
+ flex kp	377.14	0.06	25.36
full-model	225.77	0.04	27.17

(a) Configurations for keypoint design.

G-Repr.	FGD↓	Div.↑	PCM↑
baseline	262.675	18.142	0.279
+RVQ	34.940	6.713	0.327
+ separate	21.473	10.536	0.412
+ distill	1.303	13.260	0.582

(b) Gesture motion representations.

G-Gen.	FGD↓	Div.↑	PCM↑
w/o res	3.372	11.359	0.513
concat	3.415	11.314	0.514
w/o align	8.382	11.452	0.373
full-model	1.303	13.260	0.582

(c) Generator archiecture design.

Refine	VQA$_A$↑	VQA$_T$↑	FVD↓
w/o refine	91.248	5.381	492.341
+ UNet	93.958	5.479	484.323
+ skeleton	95.902	5.479	475.636
+ heatmap	96.326	6.081	476.120

(d) Image-refinement strategies.

M-Ratio	FGD↓	Div.↑	PCM↑
Uni 0-1	3.348	14.312	0.513
Uni .3-1	3.232	12.58	0.512
Uni .5-1	1.303	13.260	0.582
Uni .7-1	1.790	13.49	0.572

(e) mask-ratio during training.

iter.	FGD↓	Div.↑	PCM↑
5	1.303	13.260	0.582
10	1.642	13.40	0.575
15	1.828	13.49	0.573
20	1.881	13.49	0.572

(f) Mask decoding steps.

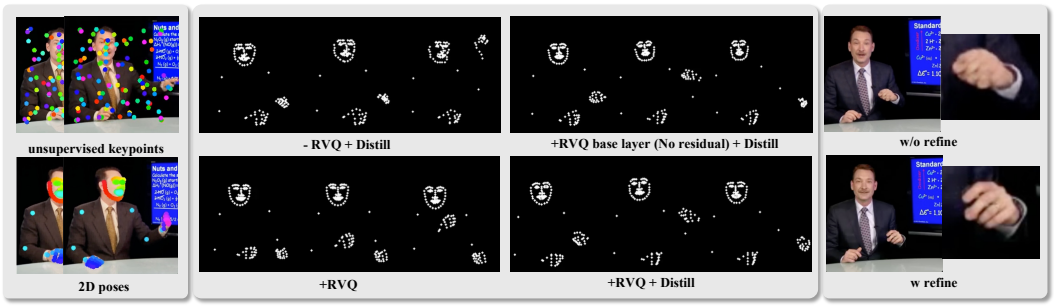


Figure 5: **Ablation visualizations.** Left: motion by unsupervised keypoints or 2d poses; Middle: RVQ-based gesture representation and generation; Right: image-refinement helps hand generation.

across realness, synchronization, and overall quality, but lower in diversity than S2G-Diffusion. We defer additional details of the user study in the appendix.

5.4 Ablation Study

In this section, we present ablation study of keypoint design for image warping, gesture pattern representation exploraton, gesture generator architecture design, and varios comparisons of image-refinement. We defer additional experiments in the Appendix.

Motion Keypoint Design We evaluate three keypoint representations for image-warping: (1) unsupervised keypoints for global optical-flow transformation (as in ANGIE and S2G-Diffusion), (2) 2D human poses, and (3) 2D human poses augmented with flexible learnable points. Each design is assessed using TPS Zhao & Zhang (2022) transformation, with self-reconstruction based on these keypoints for evaluation. We compare the full-model reconstruction with refinement against the first three designs without refinement. As shown in Tab. 3a, learnable keypoints lead to a significant decrease in FVD, highlighting their inadequacy for motion control. The 2D landmark-based keypoints yield slightly lower SSIM scores, likely due to their limited capacity to represent global transformations. The inclusion of flexible keypoints does not enhance the image-warping outcomes. Consequently, we opt to utilize 2D pose landmarks exclusively for our study.

Motion Representation. We evaluate several configurations: (1) baseline: no motion representation, relying solely on the generator to synthesize raw 2D landmarks; (2) + RVQ: utilizing Residual VQ (RVQ) to encode joint face-body keypoints; (3) + separate motion: employing two RVQs for independent face and body motions; (4) + distill: learning joint embeddings for speech and gesture in both face and body motions, followed by distillation for RVQ tokenization. We discover RVQ significantly improve the precise pose location while distillation leads to natural movements.

Generator Design. We explore various designs for the gesture generator: (1) w/o res: no residual gesture decoder; (2) concat: instead of using cross-attention for audio control, we concatenate

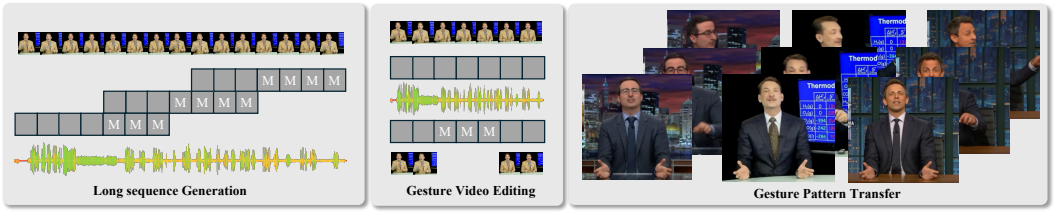


Figure 6: Our model supports multiple video gesture generation end editing applications.

the audio features with gesture latent features element-wise during generation; (3) w/o align: the audio encoder is randomly initialized rather than initialized from face and body contrastive learning. Our findings indicate that the Residual Gesture Generator significantly enhances finger motion generation. The cross-attention design outperforms element-wise concatenation, while the pre-alignment of the audio encoder notably improves FGD. We attribute this improvement to the shared similarities in the codebook and audio encoders during contrastive alignment.

Image Refinement. We examine various network designs for motion generation, specifically: (1) w/o refine: no image refinement, relying solely on image warping; (2) + UNet: employing a standard UNet; (3) + pose skeleton: integrating connected skeleton maps as in the diffusion ReferenceNet Hu et al. (2023); (4) + edge heatmap: substituting the previous design with our learnable edge heatmap. Our experiments reveal that the edge heatmap outperforms skeleton maps, likely due to the learnable thickness of connections, which provides better semantic-aware generation guidance.

Training and Inference Strategy. We evaluate the mask ratio during training and the number of inference steps during decoding. As shown in Tab. 3f, our model requires only 5 inference steps, in contrast to over 50 or 100 steps in diffusion-based models. Furthermore, a uniform masking ratio between 0.5 and 1 during training yields optimal performance.

5.5 Application

Long Sequence Generation. Shown on the left of Fig. 6, to generate long sequences, we start with the initial frame and the corresponding target audio, which we segment into smaller windows. After generating the first segment, we use the last few frames of the generated output as the new starting frame conditions for the next segment of audio, allowing for a iterative outpainting.

Video Gesture Editing. For gesture editing and inpainting, we first extract the keypoints from a given video sequence and tokenize the face and body movements into motion tokens. Thanks to the model's bidirectional decoding capability, we insert [MASK] tokens wherever edits are needed. Trained on temporal masking, the model generate coherent gestures in the masked areas. By incorporating different speech input and speaker embeddings, we can create new gesture patterns and re-render the video based on the edited latent motion tokens.

Gesture Pattern Transfer. Given the design of our framework, with different identity embedding, the model can generate different gesture patterns given the input identity embedding control given the same audio. Please see the demo videos in our Appendix for more details.

6 Conclusion

We present **Realistic-Gesture**, a framework for generating realistic co-speech gesture videos. To ensure the gestures cohere well with speech, we propose speech-content aware gesture motion representation though knowledge distillation from the gesture-speech aligned features obtained through contrastive learning. Our masked gesture motion generator enables the creation and editing of long, high-quality gesture motion sequences. Our pixel-level refinement module further improves the transformation of inferred gesture motions into realistic animations for large-scale body motion. We believe this work will encourage further exploration of the relationship between gesture patterns and speech context for more compelling gesture video generations in the future.

Realistic Gesture: Co-Speech Gesture Video Generation Through Context-Aware Gesture Representation

Supplementary Material

A Overview

The supplementary material is organized into the following sections:

- Section B: Dataset Details and Preprocessing
- Section C: Additional Implementation Details
- Section D: Speech-Gesture Alignment
- Section E: Additional Experiments
- Section F: Time and Resource Consumption
- Section G: User Study Details
- Section H TPS-based Image Warping
- Section I Ethical Ethical Considerations
- Section J: Limitations

For more visualization, please see the additional demo videos.

B Dataset Details and Preprocessing

B.1 Preprocessing

We found that many videos used in ANGIE Liu et al. (2022c) and S2G-Diffusion He et al. (2024), particularly for the subject *Jon*, are no longer available. To address this, we replaced *Jon* with *Noah*. We utilized the PATS Ginosar et al. (2019) metadata to download videos from YouTube and preprocess them. After filtering, we obtained 1080 videos for *Oliver*, 1080 for *Kubinec*, 1080 for *Seth*, and 988 for *Noah*. For the testing dataset, we collected 120 videos for *Oliver*, 120 for *Kubinec*, 120 for *Seth*, and 94 for *Noah*.

During the dataset preprocessing, while for image-generation we use the whole video preprocessed as above, for for the speech-gesture alignment and gesture pattern generation modules, we further preprocess the data by slicing them into smaller chunks following S2G-Diffusion He et al. (2024). Specifically, based on the source training dataset, the keypoint sequences and audio sequences are clipped to 80 frames (3.2s) with stride 10 (0.4s) for training. We obtain 85971 overlapping training examples and 8867 testing examples for gesture pattern modeling.

B.2 Feature Representation

Gesture Keypoints. We utilize RTMPose Jiang et al. (2023) from MMPose OpenMMLab (2020) for whole-human-body keypoint identification. The keypoint definition is based on by 133 CoCo human pose estimation. Due to the PATS Ginosar et al. (2019) only contains the upper body, we select 68 face landmarks for face motion modeling, 3 for left shoulder, 3 for right shoulder, 21 for left hand and 21 for right hand separately, which results in flattened face feature with dim of 136 and body feature with dim of 96.

Audio Features. The audio features are pre-extracted WavLM features (dim of 1024) with additional low-level mel-spectrum and beat information with dimension of 34. We concatenate them channel-wise as the speech feature.

B.3 Dataset License.

The video data within PATS dataset include personal identity information, and we strictly adhere to the data usage license "CC BY - NC - ND 4.0 International," which permits non-commercial use.

C Additional Implementation Details

We jointly train the framework on four speakers. The following sections provide the technical details for each module's training.

Optimizer Settings. All modules utilize the Adam Optimizer Kingma (2014) during training, with a learning rate of 1×10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.999$.

Speech-Gesture Alignment. For aligning speech with facial and bodily gestures, we implement two standard transformer blocks for encoding each modality. The latent dimension is configured to 384, accompanied by a feedforward size of 1024. We calculate the mean features for both modalities and project them using a two-layer MLP in a contrastive learning framework, with a temperature parameter set to 0.7.

Residual Vector Quantization (RVQ) Tokenization. We employ four layers of codebooks for residual vector quantization Lee et al. (2022) for both face and body modalities, each comprising 512 codes. To address potential collapse issues during training, we implement codebook resets. The RVQ encoder and decoder are built with two layers of convolutional blocks and a latent dimension of 512. We avoid temporal down-sampling to ensure the latent features maintain the same temporal length as the original input sequences. During RVQ training, we set $\alpha = 1$ and $\beta = 0.5$ to balance gesture reconstruction with speech-context distillation.

Mask Gesture Generator. The generator takes sequences of discrete tokens for both face and body, derived from the RVQ codebook. This module includes two layers of audio encoders for face and body, initialized based on the Speech-Gesture Alignment. The latent dimension is again set to 384, with a feedforward dimension of 1024, and it features eight layers for both modalities. A two-layer MLP is utilized to project the latent space to the codebook dimension, and cross-entropy is employed for model training. We calculate reconstruction and acceleration loss by feeding the predicted tokens into the RVQ decoder. A reconstruction loss of 50 is maintained during training, and the mask ratio is uniformly varied between 0.5 and 1.0. For inference, a cosine schedule is adopted for decoding. The Mask Gesture Generator is trained over 1000 epochs, taking approximately 1.5 days to complete.

Residual Gesture Generator. The Residual Gesture Generator is designed similarly to the Mask Gesture Generator but utilizes only six layers for the generator. It features four embedding and classification layers corresponding to the RVQ tokenization scheme for residual layers. This module is trained for an additional 500 epochs, requiring about 0.5 days to finalize.

Image Warping. For pixel-level motion generation, we utilize Thin Plate Splines (TPS) Zhao & Zhang (2022). Our framework tracks 116 keypoints (68 for the face and 48 for the body). The number of TPS transformations K is set to 29, with each transformation utilizing $N = 4$ paired keypoints. In accordance with TPS methodologies, both the dense motion network and occlusion-aware generators leverage 2D convolutions to produce 64×64 weight maps for optical flow generation, along with four occlusion masks at various resolutions (32, 64, 128, and 256) to facilitate image frame synthesis.

Image-refinement. We use the UNet similar to S2G-Diffusion He et al. (2024) to restore missing details, further improve the hand and shoulder areas. We keep the training loss to be the same except the added conditional adversarial loss based on edge heatmap. For the network design difference, we add the multi-level edge heatmap as additional control for different resolutions (32, 64 and 128). Each correponds to a SPADE Park et al. (2019) block to inject the semantic control into the current generation.

D Speech-Gesture Alignment

To validate the effectiveness of Speech-Gesture Alignment, inspired by TMR Petrovich et al. (2023) we propose the following speech2gesture and gesture2speech retrieval as the evaluation benchmark.

Table 4: **Speech-to-Gesture Motion retrieval benchmark on PATS:** We establish two evaluation settings as described in Section D.

Setting	Speech-Face retrieval					Face-Speech retrieval				
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑
(a) All	0.181	0.350	0.485	0.722	1.343	0.226	0.361	0.429	0.677	1.207
(a) w/o mask	0.142	0.326	0.388	0.656	1.112	0.158	0.299	0.343	0.612	1.026
(b) Small batches	26.230	45.318	59.330	77.019	89.858	24.977	44.822	59.894	77.775	90.264
(b) w/o mask	25.373	44.221	60.432	78.141	88.232	24.534	44.532	59.121	74.232	87.675
Setting	Speech-Body retrieval					Body-Speech retrieval				
	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑
(a) All	0.102	0.237	0.327	0.587	1.230	0.158	0.271	0.406	0.654	1.320
(a) w/o mask	0.112	0.143	0.303	0.494	1.023	0.144	0.253	0.384	0.599	1.187
(b) Small batches	25.542	43.660	57.954	77.471	90.309	24.052	43.874	58.495	76.986	89.745
(b) w/o mask	23.437	40.653	54.332	74.983	88.273	22.454	40.235	56.383	74.436	88.675

Table 5: **Gesture Generation Comparison on BEAT-X.** We report FGD $\times 10^{-1}$, BC $\times 10^{-1}$, Diversity, MSE $\times 10^{-8}$, and LVD $\times 10^{-5}$. Realistic-Gesture improves FGD and diversity compared with existing methods.

	FGD ↓	BC ↑	Diversity ↑	MSE ↓	LVD ↓
Rhythmic GesticulatorAo et al. (2022)	6.453	6.558	9.132	-	-
TalkSHOWYi et al. (2023)	6.209	6.947	13.47	7.791	7.771
EMAGE Liu et al. (2023)	5.512	**7.724**	13.06	7.680	7.556
Ours (w/o Distillation)	7.479	7.395	12.12	7.656	7.671
Ours	**4.650**	7.370	**13.55**	**7.343**	**7.432**

Evaluation settings. The retrieval perforamnce is measures under recall at various ranks, R@1, R@2, etc.Ṙecall at rank k indicates the percentage of times the correct label is among the top k results; therefore higher is better. We define two settings, by changing the evaluation set. Note that, for this retrival, we are not based on the full sequence test dataset but the sliced clips, with each lasting for 3.2 seconds and 80 frames. The size of testing dataset is 8867.

(a) **All** test set samples for face and body motions are used as a first setting. This set is problematic because the speech and gesture motion should not be of one-to-one mapping relationship.

(c) **Small batch** size of 32 speech-gesture pairs are randomly picked, reporting average performance.

Given this evaluation definition, we evaluate the speech-gesture alignment in Tab. 4. Based on the retrieval evaluation, we discover the gesture patterns and speech context are very hard to have precise one-to-one mapping relationship as shown by the significantly low performance of retrieval. Due to global contrastive alignment cannot guarantee the global alignment, without applying mask reduces the retrial accuracy for both face and body. Based on setting (c), within a small batch size of 32, the model achieves significantly higher performance, indicating the alignment pre-training does provide the model with the discrimination over different speech context and the motion. For each setting, we construct an ablation without applying temporal masking. The results demonstrate that temporal masking can increase the robustness of retrieval.

E Additional Experiments

In the main paper, we have shown our method achieves promising joint gesture motion and video generation. To understand the disentangled gesture and video avatar generation separately, we further conduct Gesture Generation and Video Avatar Animation experiments separately to compare our method with the corresponding representative works for each domain.

E.1 GESTURE GENERATION

Experiment Settings We select BEAT-X Liu et al. (2023) as the dataset for additional gesture generation comparison. For consistency, we will exclude the image-to-animation component from our method and extend gesture representation from 2D to 3D poses. (with SMPL-X expressions for face gestures, as in the existing literature) We compare the gesture generation module of our work with representative state-of-the-art methods in co-speech gesture generation Ao et al. (2022); Yi et al. (2023); Liu et al. (2023). We further design a baseline without using contextual distillation.

Experiment Results As shown in Tab. 5, our method significantly improve the SMPL-X based co-speech gesture generation with lower FGD and higher diversity. Specifically, Our methods have present smoother gesture motion patterns compared with existing works. It demonstrates the effectiveness of contextual distillation for the motion representation learning in our framework. We defer the video comparisons in the Appendix videos for reference.

Long Sequence generation To understand the capability of our framework for long sequence generation, we conduct an ablation study for both PATS and BEAT-X dataset. For BEAT-X, we cut the testing audios into segments of 256 (about 8.53 seconds) for short sequence evaluation and use raw testing audios for long sequence evaluation in Tab. 5. Shown in Tab. 6, it is interesting for PATS dataset, long-sequence generation as an application in the main paper presents quality lower than normal settings. However, for BEAT-X dataset, the generation quality is not affected much. We attribute this difference caused by the dataset difference. Because PATS dataset consists training video lengths with a average of less than 10 seconds, the model presents less diverse gesture patterns. However, in BEAT-X, most of gesture video sequences are over 30 or 1 minutes, our method further benefits from this long sequence learning precess and presents higher qualities.

Table 6: Long Sequence Generation Quality.

Dataset	Setting	FGD	Diversity	BAS
PATS	$\leq$10s	1.303	13.260	0.996
	$>$10s	2.356	11.956	0.994
BEAT-X	$\leq$10s	4.747	13.14	7.323
	$>$10s	4.650	13.55	7.370

E.2 VIDEO AVATAR ANIMATION

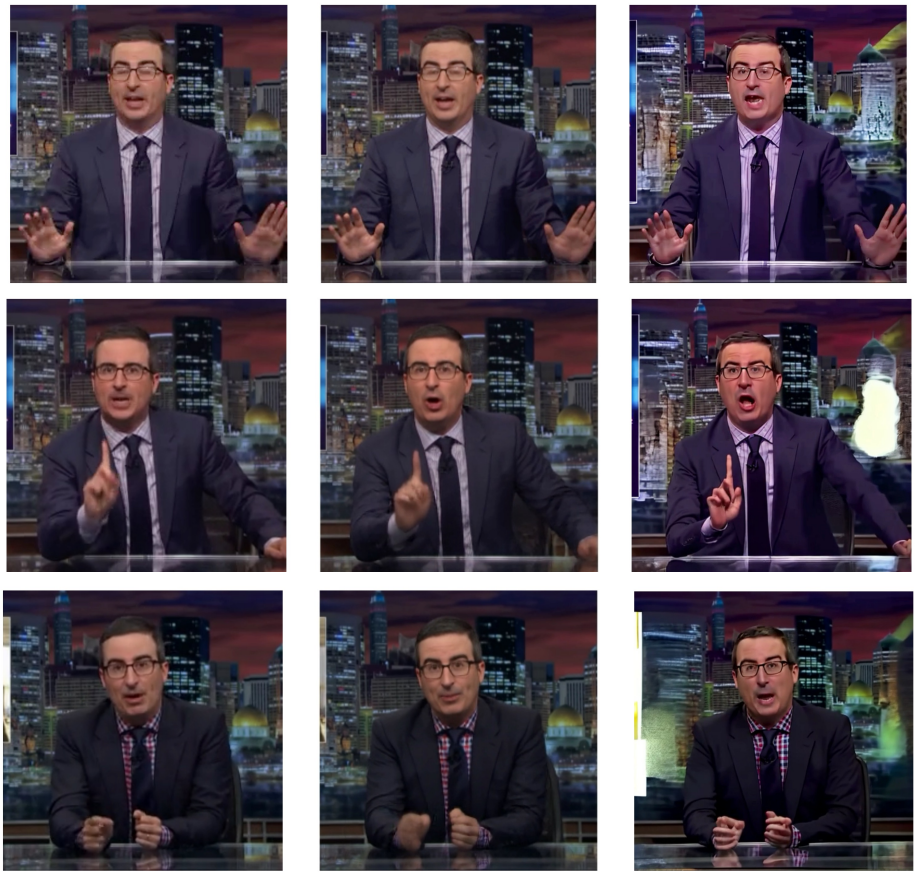
Experiment Settings. We select PATS dataset as in main paper for avatar rendering comparison. We processed the videos into 512x512 for Diffusion-bassed model AnimateAnyone Hu et al. (2023). We extract the 2D poses by MMPOse OpenMMLab (2020) for pose guidance for the Diffusion Model, and maintain all the training details as in AnimateAnyone for consistency.

Experiment Results. We compare the gesture generation module of our work with representative AnimateAnyone Hu et al. (2023). As shown in Fig. 7, though AnimateAnyone achieves better video generation quality for hand structure of the speaker centering in the video, it fails to maintain the speaker identity, making the avatar less similar to the source image compared with our method. In addition, due to the entanglement of camera motions and speaker gesture motions within the dataset, AnimateAnyone fails to separate two types of motions from the source training video, thus leading to significant background changes over time and dynamic inconsistency. Unlike completely relying on human skeletons as conditions in AnimateAnyone, our method benefits from Warping-based method, which has the capability of resolving the background motions in addition to the speaker motion. We defer visual comparisons in the Appendix videos.

F TIME AND RESOURCE CONSUMPTION

In Tab. 7, we present a comparison of training and inference times against existing baseline methods. For audio-gesture generation, our model's training time is comparable, albeit slightly slower, than that of ANGIE Liu et al. (2022c) and S2G-Diffusion He et al. (2024), primarily due to the inclusion of additional modules. However, it is considerably faster than MM-Diffusion Ruan et al. (2023). Notably, our method excels in inference speed, outperforming all other baselines.

While the training of image-warping and image refinement requires a lot of time, our method leads to a substantial reduction in overall time and resource usage compared to MM-Diffusion and other stable-diffusion-based video generation approaches. Furthermore, the generative masking paradigm



Ground-Truth Ours AnimateAnyone

Figure 7: **Comparison of Video Avatar Animation** Though presented with worse hand structure reconstruction, we achieve better identity preserving and significantly better background motion.

we employ significantly cuts down inference times when compared to diffusion-based models like S2G-Diffusion or the autoregressive generations in ANGIE.

We further compared image-warping based method computation requirements with Stable Diffusion-based models like AnimateAnyone Hu et al. (2023) in Tab. 8.

Table 7: **Time consumption comparison** of training (1 NVIDIA A100 GPU) and inference (1 NVIDIA GeForce RTX A6000 GPU).

Name	Training	Training Breakdown	Inference (video of ∼10 sec)
ANGIE	∼5d	Motion Repr. ∼3d + Quantize ∼0.2d + Gesture GPT ∼1.8d	∼30 sec
MM-Diffusion	∼14d	Generation ∼9d + Super-Resolution ∼5d	∼600 sec
S2G-Diffusion	∼5d	Motion Decouple ∼3d + Motion Diffusion ∼1.5d + Refine ∼0.5d	∼35 sec
Ours	∼6d	Quantize ∼0.2d + Mask-Gen ∼1.5d + Res-Gen ∼0.5d + Img-warp & Refine ∼3.5d	∼3 sec

Table 8: **Resource consumption comparison** with Stable-Diffusion-based Image-Animation models (1 NVIDIA A100 GPU), * means our re-implementation on PATS dataset.

Methods	Training↓	Batch Size	Resolution	Memory↓	Training Task	Inference↑
AnimateAnyone*	10 days	4	512	44 GB	Pose-2-Img	-
AnimateAnyone*	5 days	4	512	36GB	Img-2-Vid	15s
Ours	2.5 days	64	256	64 GB	Img-Warp	$\leq$1s
Ours	1 day	64	256	48GB	Img-Refine	$\leq$ 1s
Ours	3.5 days	32	512	60GB	Img-Warp	$\leq$1s
Ours	1 day	32	512	40GB	Img-Refine	$\leq$1s

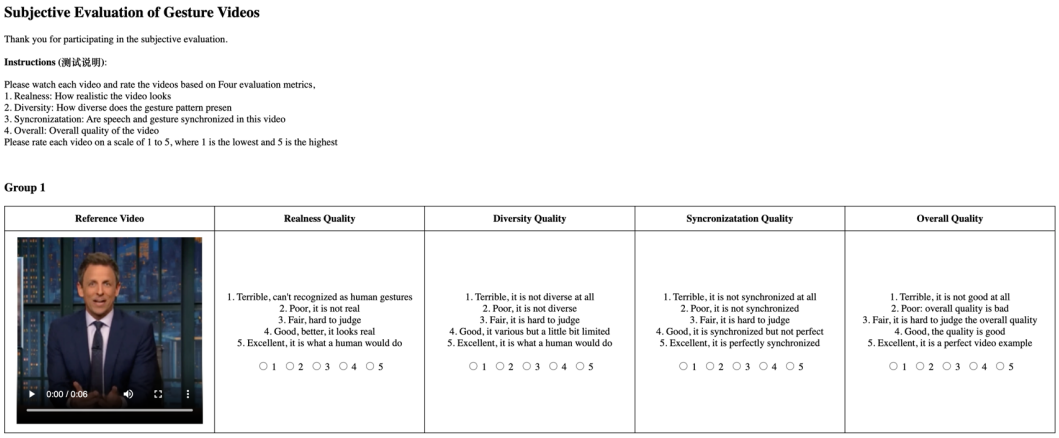


Figure 8: Screenshot of user study website.

G User Study Details

For user study, we recruited 20 participants with good English proficiency. To conduct the user study, we randomly select 80 videos from ground-truth, MM-Diffusion Ruan et al. (2023), ANGIE Liu et al. (2022c), S2G-Diffusion He et al. (2024). Each user works on 20 videos, with 4 videos from each of the aforementioned methods. The users are not informed of the source of the video for fair evaluations. A visualization of the user study is shown in Fig. 8.

H TPS-based Image-Warping

In this paper, we utilize Thin Plate Splines (TPS) Zhao & Zhang (2022) to model deformations based on human poses for image-warping. Here, we provide additional details on this approach.

The TPS transformation accepts N pairs of corresponding keypoints $(p_i^{\mathbf{D}}, p_i^{\mathbf{S}})$ for $i = 1, 2, \ldots, N$ (referred to as control points) from a driving image $\mathbf{D}$ and a source image $\mathbf{S}$. It outputs a pixel coordinate mapping $\mathcal{T}_{tps}(\cdot)$, which represents the backward optical flow from $\mathbf{D}$ to $\mathbf{S}$. This transformation is founded on the principle that 2D warping can be effectively modeled through a thin plate deformation mechanism. The TPS transformation seeks to minimize the energy associated with bending this thin plate while ensuring that the deformation aligns accurately with the control points. The mathematical formulation is as follows:

$$\begin{aligned} \min \iint_{\mathbb{R}^2} \left(\left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial y^2} \right)^2 \right) dxdy, \\ \text{s.t.} \quad \mathcal{T}_{tps}(p_i^{\mathbf{D}}) = p_i^{\mathbf{S}}, \quad i = 1, 2, \ldots, N, \end{aligned} \tag{9}$$

where $p_i^{\mathbf{D}}$ and $p_i^{\mathbf{S}}$ denote the i^{th} keypoints in $\mathbf{D}$ and $\mathbf{S}$ respectively. As shown in Zhao & Zhang (2022), it can be demonstrated that the TPS interpolating function satisfies Eq. (9):

$$\mathcal{T}_{tps}(p) = A \begin{bmatrix} p \\ 1 \end{bmatrix} + \sum_{i=1}^{N} w_i U \left(\left\| p_i^{\mathbf{D}} - p \right\|_2 \right), \tag{10}$$

where $p = (x, y)^\top$ represents the coordinates in $\mathbf{D}$, and $p_i^{\mathbf{D}}$ is the i^{th} keypoint in $\mathbf{D}$. The function $U(r) = r^2 \log r^2$ serves as a radial basis function. Notably, $U(r)$ is the fundamental solution to the biharmonic equation Selvadurai & Selvadurai (2000), defined by:

$$\Delta^2 U = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^2 U \propto \delta_{(0,0)}, \tag{11}$$

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

$$\delta_{(0,0)} = \begin{cases} \infty, & \text{if } (x, y) = (0, 0) \\ 0, & \text{otherwise} \end{cases}, \quad \text{and} \iint_{\mathbb{R}^2} \delta_{(0,0)}(x, y)\, dxdy = 1, \tag{12}$$

indicating that $\delta_{(0,0)}$ is zero everywhere except at the origin, where it integrates to one.

We denote the i^{th} keypoint in image $\mathbf{X}$ (either $\mathbf{D}$ or $\mathbf{S}$) as $p_i^{\mathbf{X}} = (x_i^{\mathbf{X}}, y_i^{\mathbf{X}})^\top$, and we define:

$$r_{ij} = \left\| p_i^{\mathbf{D}} - p_j^{\mathbf{D}} \right\|, \quad i, j = 1, 2, \ldots, N.$$

Next, we construct the following matrices:

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

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

We can then determine the affine parameters $A \in \mathcal{R}^{2\times 3}$ and the TPS weights $w_i \in \mathcal{R}^{2\times 1}$ by solving the following equation:

$$[w_1, w_2, \cdots, w_N, A]^\top = L^{-1}Y. \tag{13}$$

In Eq. (10), the first term $A \begin{bmatrix} p \\ 1 \end{bmatrix}$ represents an affine transformation that aligns the paired control points $(p_i^{\mathbf{D}}, p_i^{\mathbf{S}})$ in linear space. The second term $\sum_{i=1}^{N} w_i U\left(\left\| p_i^{\mathbf{D}} - p \right\|_2\right)$ accounts for nonlinear distortions that enable the thin plate to be elevated or depressed. By combining both linear and nonlinear transformations, the TPS framework facilitates precise deformations, which are essential for accurately capturing motion while preserving critical appearance details within our framework.

I Ethical Considerations

While this work is centered on generating co-speech gesture videos, it also raises important ethical concerns due to its potential for photo-realistic rendering. This capability could be misused to fabricate videos of public figures making statements or attending events that never took place. Such risks are part of a broader issue within the realm of AI-generated photo-realistic humans, where phenomena like deepfakes and animated representations pose significant ethical challenges.

Although it is difficult to eliminate the potential for misuse entirely, our research offers a valuable technical analysis of gesture video synthesis. This contribution is intended to enhance understanding of the technology's capabilities and limitations, particularly concerning details such as facial nuances and temporal coherence.

In addition, we emphasize the importance of responsible use. We recommend implementing practices such as watermarking generated videos and utilizing synthetic avatar detection tools for photo-realistic images. These measures are vital in mitigating the risks associated with the misuse of this technology and ensuring ethical standards are upheld.

J Limitations

While our method have achieved significant improvements over existing baselines, there are still two limitations of the current work.

First, the generation quality still exhibit blurries and flickering issues. The intricate structure of hand hinders the generator in understanding the complex motions. In addition, PATS dataset is sourced from in-the-wild videos of low quality. Most frames extracted from videos demonstrate blurry hands, limiting the network learning. Thus, it is important to collect the high-quality gesture video dataset with clearer hands to further enhance the generation quality.

Second, when modeling the whole upper-body, it is hard to achieve synchronized lip movements aligned with the audio. Even though we explicit separate the face motion and body motion to deal with this problem, there is no regularization on lip movement. We would like to defer this problem to the future works that models disentangled and fine-grained motions for each face and body region.

REFERENCES

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*, pp. 1884–1895, 2020.

Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6):1–19, 2022.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Judee K Burgoon, Thomas Birk, and Michael Pfau. Nonverbal Behaviors, Persuasion, and Credibility. *Human communication research*, 17(1):140–169, 1990.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody Dance Now. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-Image Generation Via Masked Generative Transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. DiffSHEG: A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Expression and Gesture Generation, 2024. URL https://arxiv.org/abs/2401.04747.

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.

Jan P De Ruiter, Adrian Bangerter, and Paula Dings. The Interplay Between Gesture and Speech in the Production of Referring Expressions: Investigating the Tradeoff Hypothesis. *Topics in cognitive science*, 4(2):232–248, 2012.

Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. Diffusion-based co-speech gesture generation using joint text and audio representation. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, ICMI '23. ACM, October 2023. doi: 10.1145/3577190.3616117. URL http://dx.doi.org/10.1145/3577190.3616117.

Jacob Devlin. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning Individual Styles of Conversational Gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2019.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *Proceedings of the International Conference on Machine Learning*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Xingzhe He, Bastian Wandt, and Helge Rhodin. AutoLink: Self-Supervised Learning of Human Skeletons and Object Outlines by Linking Keypoints, 2023. URL https://arxiv.org/abs/2205.10636.

Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2263–2273, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Machine Learning*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. *arXiv preprint arXiv:2311.17117*, 2023.

Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization, 2017. URL https://arxiv.org/abs/1703.06868.

Ziyao Huang, Fan Tang, Yong Zhang, Xiaodong Cun, Juan Cao, Jintao Li, and Tong-Yee Lee. Make-your-anchor: A diffusion-based 2d avatar generation framework. *arXiv preprint arXiv:2403.16510*, 2024.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose, 2023. URL https://arxiv.org/abs/2303.07399.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution, 2016. URL https://arxiv.org/abs/1603.08155.

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.

Diederik P Kingma. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive Image Generation Using Residual Quantization, 2022. URL https://arxiv.org/abs/2203.01941.

Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to Music. *Proceedings of the Neural Information Processing Systems Conference*, 32, 2019.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13401–13412, 2021.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. MAGE: Masked Generative Encoder to Unify Representation Learning and Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023.

Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. DisCo: Disentangled Implicit Content and Rhythm Learning for Diverse Co-Speech Gestures Synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3764–3773, 2022a.

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. *arXiv preprint arXiv:2203.05297*, 2022b.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Naoya Iwamoto, Bo Zheng, and Michael J Black. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Masked Audio Gesture Modeling. *arXiv preprint arXiv:2401.00374*, 2023.

Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-Driven Co-Speech Gesture Video Generation. *Proceedings of the Neural Information Processing Systems Conference*, 35:21386–21399, 2022c.

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 Conference on Computer Vision and Pattern Recognition*, pp. 10462–10472, 2022d.

Xiaofeng Mao, Zhengkai Jiang, Qilin Wang, Chencan Fu, Jiangning Zhang, Jiafu Wu, Yabiao Wang, Chengjie Wang, Wei Li, and Mingmin Chi. Mdt-a2g: Exploring masked diffusion transformers for co-speech gesture generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 3266–3274. ACM, October 2024. doi: 10.1145/3664647.3680684. URL http://dx.doi.org/10.1145/3664647.3680684.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, volume 8, 2015.

Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets, 2014. URL https://arxiv.org/abs/1411.1784.

OpenMMLab. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose, 2020.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Mathis Petrovich, Michael J. Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis, 2023. URL https://arxiv.org/abs/2305.00976.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. MMM: Generative Masked Motion Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1546–1555, 2024.

Lawrence Rabiner and Ronald Schafer. *Theory and Applications of Digital Speech Processing*. Prentice Hall Press, 2010.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Ludan 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 Conference on Computer Vision and Pattern Recognition*, 2023.

APS Selvadurai and APS Selvadurai. The Biharmonic Equation. *Partial Differential Equations in Mechanics 2: The Biharmonic Equation, Poisson's Equation*, pp. 1–502, 2000.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First Order Motion Model for Image Animation. In *Proceedings of the Neural Information Processing Systems Conference*, 2019.

Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion Representations for Articulated Animation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

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 Conference on Computer Vision and Pattern Recognition*, pp. 11050–11059, 2022.

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.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. EMO: Emote Portrait Alive-Generating Expressive Portrait Videos with Audio2Video Diffusion Model Under Weak Conditions. *arXiv preprint arXiv:2402.17485*, 2024.

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.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning, 2018. URL https://arxiv.org/abs/1711.00937.

Qingfu Wan, Wei Zhang, and Xiangyang Xue. DeepSkeleton: Skeleton Map for 3D Human Pose Regression, 2017. URL https://arxiv.org/abs/1711.10796.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural Codec Language Models Are Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2301.02111*, 2023a.

Congyi Wang. T2m-hifigpt: Generating high quality human motion from textual descriptions with residual discrete representations, 2023. URL https://arxiv.org/abs/2312.10628.

Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. DisCo: Disentangled Control for Referring Human Dance Generation in Real World. *arXiv preprint arXiv:2307.00040*, 2023b.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *CVPR*, 2018.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023.

Zunnan Xu, Yachao Zhang, Sicheng Yang, Ronghui Li, and Xiu Li. Chain of generation: Multi-modal gesture synthesis via cascaded conditional control, 2023. URL https://arxiv.org/abs/2312.15900.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating Holistic 3D Human Motion from Speech. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

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*, 39(6), 2020.

Jian Zhao and Hui Zhang. Thin-Plate Spline Motion Model for Image Animation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3657–3666, 2022.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. *arXiv preprint arXiv:2403.14781*, 2024.