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. (2022) 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.

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

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 1: **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 2: **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.Recall 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. 1. 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 contextural distillation.

Experiment Results As shown in Tab. 2, 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. Shown in Tab. 3, 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

Table 3: Long Sequence Generation Quality.

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

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.

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. 1, 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. 4, 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. (2022) 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 we employ significantly cuts down inference times when compared to diffusion-based models like S2G-Diffusion or the autoregressive generations in ANGIE.

4

| Ground-Truth | Ours | AnimateAnyone |

Figure 1: **Comparison of Video Avatar Animation** Though presented with worse hand structure reconstruction, we achieve better identity preserving and significantly better background motion.

We further compared image-warping based method computation requirements with Stable Diffusion-based models like AnimateAnyone Hu et al. (2023) in Tab. 5.

Table 4: **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 5: **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	≤1s
Ours	1 day	64	256	48GB	Img-Refine	≤ 1s
Ours	3.5 days	32	512	60GB	Img-Warp	≤1s
Ours	1 day	32	512	40GB	Img-Refine	≤1s

Subjective Evaluation of Gesture Videos

Thank you for participating in the subjective evaluation.

Instructions (测试说明):

Please watch each video and rate the videos based on Four evaluation metrics,
1. Realness: How realistic the video looks
2. Diversity: How diverse does the gesture pattern presen
3. Syncronizatation: Are speech and gesture synchronized in this video
4. Overall: Overall quality of the video
Please rate each video on a scale of 1 to 5, where 1 is the lowest and 5 is the highest

Group 1

Reference Video	Realness Quality	Diversity Quality	Syncronizatation Quality	Overall Quality
	1. Terrible, can't recognize as human gestures 2. Poor, it is not real 3. Fair, hard to judge 4. Good, better, it looks real 5. Excellent, it is what a human would do ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	1. Terrible, it is not diverse at all 2. Poor, it is not diverse 3. Fair, it is hard to judge 4. Good, it various but a little bit limited 5. Excellent, it is what a human would do ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	1. Terrible, it is not synchronized at all 2. Poor, it is not synchronized 3. Fair, it is hard to judge 4. Good, it is synchronized but not perfect 5. Excellent, it is perfectly synchronized ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	1. Terrible, it is not good at all 2. Poor: overall quality is bad 3. Fair, it is hard to judge the overall quality 4. Good, the quality is good 5. Excellent, it is a perfect video example ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 2: 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. (2022), 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. 2.

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:

$$\min \iint_{\mathbb{R}^2} \left(\left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial y^2} \right)^2 \right) dxdy, \tag{1}$$

$$\text{s.t.} \quad \mathcal{T}_{tps}(p_i^{\mathbf{D}}) = p_i^{\mathbf{S}}, \quad i = 1, 2, \ldots, N,$$

where $p_i^{\mathbf{D}}$ and $p_i^{\mathbf{S}}$ denote the ith 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. (1):

$$\mathcal{T}_{tps}(p) = A \begin{bmatrix} p \\ 1 \end{bmatrix} + \sum_{i=1}^{N} w_i U \left(\left\| p_i^{\mathbf{D}} - p \right\|_2 \right), \tag{2}$$

where $p = (x, y)^\top$ represents the coordinates in $\mathbf{D}$, and $p_i^{\mathbf{D}}$ is the ith 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:

$$\Delta^2 U = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^2 U \propto \delta_{(0,0)}, \tag{3}$$

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

$$\delta_{(0,0)} = \begin{cases} \infty, & \text{if } (x, y) = (0, 0) \\ 0, & \text{otherwise} \end{cases}, \quad \text{and} \quad \iint_{\mathbb{R}^2} \delta_{(0,0)}(x, y) \, dxdy = 1, \tag{4}$$

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

We denote the ith 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:

$$[w_1, w_2, \cdots, w_N, A]^{\top} = L^{-1} Y. \tag{5}$$

In Eq. (2), 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

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.

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.

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.

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.

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`.

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`.

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, 2022.

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.

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`.

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.

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.

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.