Democratizing High-Fidelity Co-Speech Gesture Video Generation

Xu Yang[1*‡] Shaoli Huang[2*] Shenbo Xie[1*] Xuelin Chen[2] Yifei Liu[1] Changxing Ding[1†]
[1]South China University of Technology [2]Tencent AI Lab
ftyang_xu@mail.scut.edu.cn shaol.huang@gmail.com eeshbx34@mail.scut.edu.cn
xuelin.chen.3d@gmail.com ft_lyf@mail.scut.edu.cn chxding@scut.edu.cn

Figure 1. Examples of co-speech gesture videos created by our framework. These examples demonstrate the advantages of our framework regarding visual quality and synchronization while being adaptable to various speakers and contexts. To address privacy concerns, all reference images of speakers are synthesized using AI.

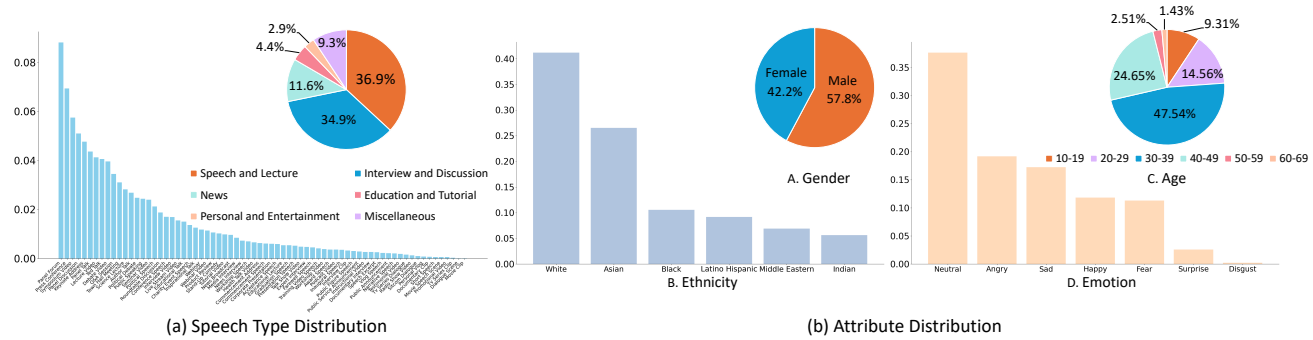
Abstract

Co-speech gesture video generation aims to synthesize realistic, audio-aligned videos of speakers, complete with synchronized facial expressions and body gestures. This task presents challenges due to the significant one-to-many mapping between audio and visual content, further complicated by the scarcity of large-scale public datasets and high computational demands. We propose a lightweight framework that utilizes 2D full-body skeletons as an efficient auxiliary condition to bridge audio signals with visual outputs. Our approach introduces a diffusion model conditioned on fine-grained audio segments and a skeleton extracted from the speaker's reference image, predicting skeletal motions through skeleton-audio feature fusion to ensure strict audio coordination and body shape consistency. The generated skeletons are then fed into an off-the-shelf human video generation model with the speaker's reference image to synthesize high-fidelity videos. To democratize research, we present CSG-405—the first public dataset with 405 hours of high-resolution videos across 71 speech types, annotated with 2D skeletons and diverse speaker demographics. Experiments show that our method exceeds state-of-the-art approaches in visual quality and synchronization while generalizing across speakers and contexts. Code, models, and CSG-405 are publicly released at https://mpi-lab.github.io/Democratizing-CSG/

*Equal Contribution.
†Corresponding Author.
‡Part of his work was done during an internship at Tencent AI Lab.

Table 1. Comparison in statistics between our CSG-405 database and existing public ones for co-speech gesture video generation.

Dataset	#Clips	#Hours	Video Resolution	#Speakers	#Speech Types	Motion Annotation Face	Body	Hands
PATS [1]	4,800	13.1	256×256	4	2	✗	✓	✓
TED-talks [6]	1,322	3.1	384×384	391	1	✗	✗	✗
CSG-405	147,550	405	512×512	17787	71	✓	✓	✓



(a) Speech Type Distribution (b) Attribute Distribution

Figure 2. More details of CSG-405. (a) The proportion of clips for each speech type. (b) Attribute distribution in gender, ethnicity, age, and emotion.

1. Introduction

Co-speech gesture video generation aims to synthesize a realistic video of a person speaking, and the obtained video has to be aligned with both the given speech audio and the reference image of the speaker. It has broad applications, including human-machine interaction [19] and digital entertainment [3]. Notably, this task presents unique challenges compared to talking face generation [28–30, 41] and full-body human video generation [16, 20, 22]. Compared to the former, it must produce videos that include both facial expressions and hand gestures. Compared to the latter, it must generate audio-aligned talking humans rather than general human movements like dancing.

To synthesize a co-speech gesture video, early methods [3, 18, 19, 24] for this task typically involve warping the reference image of the speaker based on the audio condition. However, the warping operation tends to result in artifacts when encountering significant pose variations. Recent works [23, 24, 42] utilize diffusion models [12, 17] to achieve warping-free synthesis. Moreover, to facilitate the coordination between visual content and the audio, they impose auxiliary control conditions, e.g., 3D human motions [24] and hand skeletons extracted from a reference video [42], on the diffusion model.

Despite these efforts, high-fidelity co-speech gesture video generation is still challenging due to the substantial one-to-many mapping between the audio condition and visual contents [23, 24]. Recent efforts to handle this problem have mainly been made by industry [23, 24, 42]. Their common ground is collecting large-scale private training data, e.g., 2,200 hours of video data in [24] and 200 hours in [23], and train their models using high-end GPU devices to bridge this mapping. In comparison, existing publicly available datasets for this task are small in size and lack diversity in speech scenarios. For example, existing works [3, 18, 19] only adopt the 13 hours of video data selected from the PATS dataset [1] and 3 hours of video data from TED talks in the TED-talks dataset [6] for experiments. This situation hampers democratization in the research of high-fidelity co-speech gesture video generation.

Our key insight here is that the sequence of expressive 2D full-body skeletons is an essential and economical auxiliary condition to relieve the above one-to-many mapping problem. On the one hand, the 2D skeleton is much more concise than visual data, and its synthesis is computationally efficient. On the other hand, the 2D full-body skeleton has been a popular auxiliary condition for existing diffusion-based human video generation models [16, 21, 42, 49], meaning we can easily enhance their ability to generate co-speech gesture videos with improved skeleton conditions. However, high-fidelity audio-to-skeleton prediction is non-trivial, as any artifact, e.g., misalignment between the audio and lip movements, significantly harms the quality of obtained speech videos. To handle this problem, we adopt two strategies. First, we condition this prediction with the skeleton extracted from the reference image of the target speaker, which enables the generated skeletons to conform to the speaker's body shape. Second, we concatenate the embeddings of skeletons and those of audio segments along the feature dimension as the input of the diffusion model, which enforces strict coordination compared with other conditioning strategies, e.g.,

cross-attention [13]. Finally, we feed the obtained skeleton sequence, along with the reference image of the speaker, into an off-the-shelf human video generation model to synthesize co-speech gesture videos.

Moreover, we contribute a large-scale dataset named Co-Speech Gesture (CSG-405) that is publicly available[1]. As shown in Table 1, CSG-405 contains 147,550 clips, covering 71 common speech types, totaling 405 hours. Specifically, each original video has passed a multi-stage filtering process that includes skeleton-detection quality and audio-lip sync assessment. These videos are then segmented into clips of 5-15 seconds, followed by cropping and resizing to a resolution of 512×512 pixels. Compared to existing public datasets [1, 6], our dataset features a large number of speakers, diverse speech scenarios, rich 2D skeleton annotations, and high image resolution. As shown in Figure 2, it includes speakers of different genders, ages, and ethnicities. It also covers formal (e.g., lectures, speeches, and debates) and casual (e.g., chats and Vlogs) scenarios. We adopt this dataset to train our audio-to-skeleton prediction model.

Our contributions can be summarized as follows. First, we propose a co-speech gesture video generation framework based on a lightweight audio-to-skeleton prediction model [50] and an off-the-shelf human video generation model. Second, to the best of our knowledge, we build the first publicly available large-scale dataset for co-speech gesture video generation. Third, extensive experiments demonstrate that our approach generates high-fidelity co-speech gesture videos across various speakers and scenarios, reliably promoting the performance of both general human video generation [21, 49] and co-speech gesture video generation models [42]. We hope our work can expedite the research on high-fidelity co-speech gesture video generation.

2. Related Works

Co-Speech Gesture Video Generation. Existing works for co-speech gesture video generation can be categorized into the warping-based [3, 18, 19] and warping-free [23, 24] methods. The warping-based methods first predict motion features according to the given audio and the speaker's reference image. These features are then utilized to warp the reference image into co-speech gesture video frames. According to the property of motion features, the warping-based methods can be further divided into implicit and explicit approaches. The implicit ones [3] employ an unsupervised method [7] to model the latent motion features, while the explicit ones [18, 19] usually adopt thin-plate spline (TPS) keypoints [4, 8] to represent the motion features. The motion features are typically first transformed into optical flows [11], by which the reference image can be warped into video frames. The downside of the warping operation is that it is prone to produce artifacts when encountering large pose variations.

Several recent works [23, 24, 42] discard the warping operation. They are typically diffusion model-based, and their models are trained using private large-scale datasets. Besides the speaker's reference image and the given audio, they generally adopt auxiliary conditions to achieve more fine-grained control. For example, Corona et al. [24] predicted the sequence of 3D human motions [31] from audio as the auxiliary condition. Meng et al. [42] adopted the sequence of hand skeletons extracted from another video. However, these auxiliary conditions may not be precise. Specifically, predicting the 3D motion itself is a challenging problem, while the sequence of hand skeletons extracted from another video can hardly match the target speaker's audio and body shape.

Besides, existing public datasets for co-speech gesture video generation are usually small, with low image resolutions. For example, existing works [3, 18, 19] only adopt the 1,200 clips of only four speakers and 1,322 cilps of TED talk shows selected from the PATS [1] and TED-talks [6] datasets, respectively. Their image resolutions are 256×256 and 384×384 pixels, respectively. These datasets' limited scale and diversity have significantly affected the fidelity of co-speech gesture video generation.

In this paper, we democratize high-fidelity co-speech gesture video generation by an expressive and reliable audio-to-skeleton prediction model and making use of off-the-shelf powerful human video generation models. We also contribute a large-scale dataset that surpasses existing public ones in scale, diversity, and quality.

Human Video Generation. Human video generation [2, 16, 20, 28, 29] aims to create realistic human videos from a given human image and other conditions such as audio and pose sequences. It includes many popular subtasks, e.g., full-body human video generation [2, 16, 20] and talking face [28–30, 37, 38] generation. The former [16, 20, 21] usually adopts a sequence of skeleton [9, 10] as condition to produce full-body human videos. For example, Li et al. proposed the Animate Anyone method [16] that is based on the powerful Stable Diffusion model [17]. It also introduces a reference net to Stable Diffusion, which further improves the appearance consistency between the obtained video frames and the reference human image. To promote model performance across different body shapes, recent works incorporate human body information, e.g., densepose [20, 36], 3D meshes [51, 52], depth maps [22], and normal maps [22, 31, 32], as auxiliary conditions to the diffusion model. Besides, Zhang et al. [21] proposed to incorporate keypoint detection confidence into the pose guidance, which enhances the model's robustness to keypoint

[1]We release all links, tools, codes, and skeleton data for downloading and automatically preprocessing the involved raw speech videos.

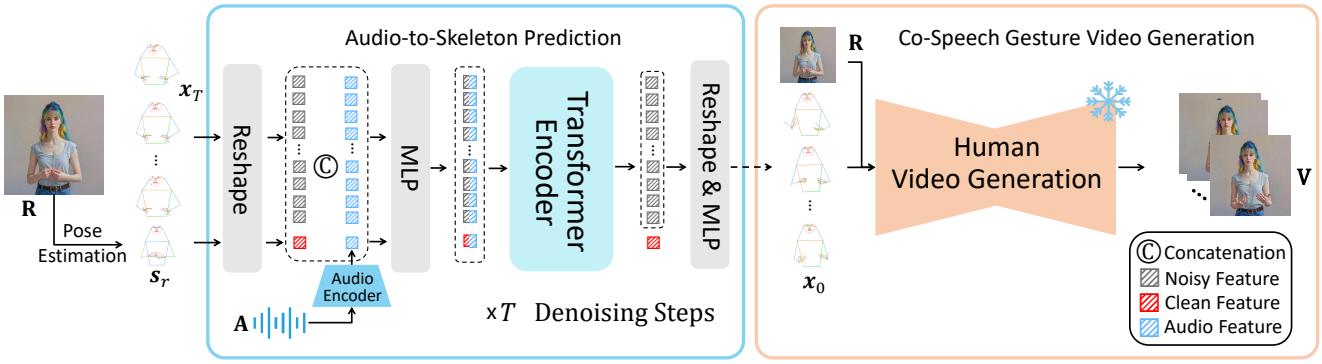


Figure 3. Overview of our co-speech gesture video generation framework. We concatenate the 2D skeleton of the reference image $\mathbf{R}$ with the noisy skeleton sequence $\boldsymbol{x}_T$ along the frame dimension, providing the body shape cue of the speaker. We then concatenate the embeddings of skeletons and those of audio segments along the feature dimension as the input of the diffusion model, enforcing strict temporal synchronization. Finally, we employ one off-the-shelf human video generation model to produce the co-speech gesture video $\mathbf{V}$ with the synthesized skeleton sequence as an auxiliary condition.

detection errors. Finally, Tu et al. [49] proposed to utilize a face encoder to extract facial features from the reference image, which are then fed into the video generation model to enhance the fidelity of synthesized human videos.

Talking face generation [28–30, 37, 38, 41] aims to create videos with natural facial expressions from a speech audio and a reference facial image. Existing methods can be categorized into one- and two-stage approaches. The one-stage methods [29] adopt a diffusion model [17] to directly synthesize talking face videos from the audio. In contrast, the two-stage methods [28, 30, 38] first generate intermediate face representations, e.g., 3D Morphable Model (3DMM) coefficients [39], and then feed them into a video generation model. Compared to this task, co-speech gesture video generation involves body parts below the shoulders.

3. Methods

We first describe our co-speech gesture video generation framework in Section 3.1 and Figure 3. Then, we introduce the way to construct our large-scale CSG-405 database in Section 3.2.

3.1. Co-Speech Gesture Video Generation

Preliminaries. We denote the audio as $\mathbf{A}$, the reference image of the target speaker as $\mathbf{R}$, and the skeleton extracted from $\mathbf{R}$ as $\boldsymbol{s}_r \in \mathbb{R}^{1\times 2K}$, where K is the number of 2D body keypoints in the skeleton. Given a random pair of $\mathbf{A}$ and $\boldsymbol{s}_r$, our audio-to-skeleton prediction model G_s produces a sequence of high-fidelity and expressive skeletons $\mathbf{S}_g \in \mathbb{R}^{F\times 2K}$, where F means the frame number. We feed $\mathbf{S}_g$ and $\mathbf{R}$ into an off-the-shelf human video generation model G_v, which produces the final co-speech gesture video $\mathbf{V}$. The whole framework can be formulated as: $\mathbf{S}_g = G_s(\boldsymbol{s}_r, \mathbf{A})$ and $\mathbf{V} = G_v(\mathbf{S}_g, \mathbf{R})$.

Audio-to-Skeleton Prediction. We build G_s as an audio-conditioned diffusion model, which synthesizes $\mathbf{S}_g$ by progressive denoising from random Gaussian noise. During training, keypoint coordinates in each ground-truth skeleton sequence is represented as $\boldsymbol{x}_0 \in \mathbb{R}^{F\times 2K}$. After T diffusion steps, $\boldsymbol{x}_0$ becomes an isotropic Gaussian noise $\boldsymbol{x}_T$. Then, we optimize parameters of G_s by denoising $\boldsymbol{x}_T$ according to the time step t and audio condition $\mathbf{A}$. Following [5], we require G_s to predict $\boldsymbol{x}_0$ instead of the added noise in each denoising step. The loss function can be formulated as:

$$L = \mathbb{E}_{\boldsymbol{x}_0, t}\left[\|\boldsymbol{x}_0 - G_s(\boldsymbol{x}_t, t, \mathbf{A})\|_2^2\right],$$

where $\boldsymbol{x}_t$ denotes the noisy skeleton sequence in step t.

As shown in Figure 3, we implement the above model according to the Diffusion Transformer (DiT) architecture [25]. Moreover, as any artifact in the predicted skeleton harms the subsequent video generation process, we propose the following two strategies to achieve high-fidelity audio-to-skeleton prediction.

First, we enable model prediction to accommodate the body shape of the target speaker. To achieve this goal, we concatenate $\boldsymbol{x}_t$ with $\boldsymbol{s}_r$ along the frame dimension as the input of the diffusion model for each denoising step t. By modeling their temporal dependencies with the self-attention layers, G_s effectively leverages cues from $\boldsymbol{s}_r$ to generate $\mathbf{S}_g$ that conforms to the speaker's body shape.

Second, we enforce strict coordination between $\mathbf{S}_g$ and $\mathbf{A}$. Existing approaches to co-speech gesture generation usually impose the audio condition via cross-attention [3, 19, 27]. In this scheme, the correspondence between each gesture and audio segment is obscure, harming the obtained gestures' fidelity. To handle this problem, we extract a sequence of more fine-grained segment-level audio features $\hat{\mathbf{A}} \in \mathbb{R}^{F\times 768}$ from $\mathbf{A}$ using the wav2vec 2.0 model [47], where the segments number is the same as the video frame number. Then, we concatenate $\boldsymbol{x}_t$ with $\hat{\mathbf{A}}$ along the feature

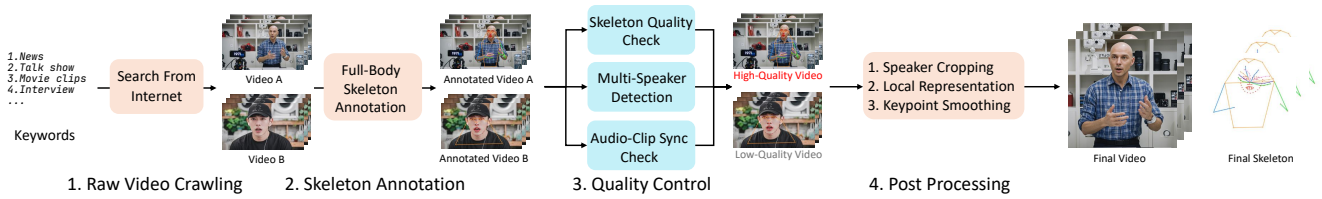


Figure 4. An overview of our data collection pipeline.

dimension in each denoising step. To maintain dimension consistency between input embeddings, we concatenate a zero vector to $\mathbf{s}_r$. Finally, we utilize a linear projection layer to project the concatenated features to a dimension suitable to the DiT layers. Compared to cross-attention [13], our strategy ensures one-to-one correspondence between each gesture and audio segment, which promotes the fidelity and expressiveness of generated skeletons.

Finally, we train G_s with the classifier-free guidance strategy [48]. Specifically, we replace the audio condition $\mathbf{A}$ with $\varnothing$ using a probability of 10% during training, which results in an unconditional generation. When $\varnothing$ is adopted, we concatenate a set of learnable vectors that are randomly initialized to $\boldsymbol{x}_t$. During inference, we make a trade-off between generation diversity and fidelity as follows:

$$G'_s(\boldsymbol{x}_t, t, \mathbf{A}) = G_s(\boldsymbol{x}_t, t, \varnothing) + \alpha \cdot (G_s(\boldsymbol{x}_t, t, \mathbf{A}) - G_s(\boldsymbol{x}_t, t, \varnothing)),$$

where α denotes the classifier-free guidance scale.

Co-Speech Gesture Video Generation. We feed both the predicted skeleton sequence $\mathbf{S}_g$ and the reference image $\mathbf{R}$ into an off-the-shelf human video generation model G_v, which generates the final co-speech gesture video $\mathbf{V}$. Since $\mathbf{S}_g$ well aligns with the audio condition $\mathbf{A}$ and the reference image $\mathbf{R}$, G_v can produce more realistic human videos compared to those using skeletons extracted from other reference videos [21, 42, 49].

3.2. Dataset Construction

Dataset Description To further democratize the research on high-fidelity co-speech gesture video generation, we build a large-scale database named CSG-405 and make it publicly available. This dataset contains 147,550 clips that cover 71 common types of speeches, with a total duration of 405 hours. Each clip is of 5-15 seconds and adjusted to 25FPS. The mean duration of each clip is 9.88 seconds. We also provide high-quality skeleton annotation for each video frame. Each skeleton contains 133 keypoints on the face, hands, and other body parts. Besides, all audio signals are sampled at a rate of 16K.

As shown in Table 1 and Figure 2, our database features a large number of speakers. They are of different genders, ethnicities, ages, and emotions. Similar to [33, 34], the number of speaker identities and person attributes is estimated using a facial image understanding model [35].

Data Collection Pipeline. We illustrate our data collection pipeline in Figure 4. It incorporates four stages including raw video crawling, skeleton annotation, quality control, and post-processing.

First, we utilize GPT-4o [43] to enumerate speech types. We search for videos of each speech type according to keyword from Youtube and download them with a high-resolution of 1280×720 pixels. Figure 2 (a) summarizes the proportion of obtained videos for each speech type, which indicates that our database covers diverse real-life speech types.

Second, we annotate a full-body skeleton for each video frame. Specifically, we leverage the powerful DWPose model [10] to annotate 133 body keypoints and also obtain a confidence score for each keypoint. The obtained keypoints follow the COCO-Whole-Body format [44]. As shot transitions may occur within a video, we detect sudden changes in camera motion or image color. According to the detected changes, we segment each video into clips with a length of 5-15 seconds.

Third, we carefully check the quality of skeletons, video, and audio. For the skeleton, we remove clips with missing upper body keypoints, those captured from the side or back viewpoints, those containing very small human figures, and those with basically static poses. If multiple persons appear in a video, we process each person's skeleton independently. For the video and audio, we utilize PyAnnote [45] to detect and remove clips where multiple persons speak simultaneously. We also adopt SyncNet [46] to remove clips where the lip movements misalign with the audio.

Finally, we crop each speaker with margins from video frames and resize all cropped regions to 512×512 pixels. We also convert keypoint coordinates into local motion representations [26]. Specifically, we adopt the nose tip and wrists as the root nodes for keypoints on the face and hands, respectively. For keypoints on the other body parts, we utilize the neck keypoint as the root node. Then, we compute the coordinate of each keypoint relative to its corresponding root node. This strategy disentangles holistic motion with subtle facial and hand movements, facilitating our model to synthesize high-fidelity and expressive skeletons. Besides, we apply temporal smoothing on all keypoints except for those on the mouth to reduce irregular jittering. As the mouth usually moves drastically during speech, we keep keypoint coordinates on the mouth unchanged.

4. Experiments

Datasets and Metrics. We conduct experiments on PATS [1], TED-talks [6], and our collected CSG-405 datasets. We randomly select 133 clips from our CSG-405 database for testing and divide the remaining data into 90% for training and 10% for validation. The training, validation and testing data are about 364, 41, and 0.5 hours, respectively. After training on CSG-405, the obtained models are directly evaluated on PATS and TED-talks without further fine-tuning. Following [3, 19], we employ the 474 clips of four speakers (Jon, Kubinec, Oliver, and Seth) from the PATS [1] database for testing. The mean length of these clips is 9.8 seconds and the resolution of video frames is unified to 256×256 pixels. The testing set of the TED-talks [6] database include 122 clips. The resolution of video frames is unified to 384×384 pixels and the frame number ranges from 64 to 1,024 in each clip.

We perform evaluations using both the paired and unpaired configurations. In the former setting, the audio and reference image are from the same video, while the latter composes audio-image pairs from different videos. Existing methods [23, 24, 42] typically overlook the latter setting, which is evidently more practical in real-world applications. For the paired setting, we adopt Structural Similarity (SSIM) [53] and Peak Signal to Noise Ratio (PSNR) [54] to evaluate the visual quality of the generated videos. For the unpaired setting, we employ Fréchet Inception Distance (FID) [40] and Fréchet Video Distance (FVD) [15] to evaluate the general video quality. Additionally, we employ SyncNet [46] to calculate Sync-C and Sync-D, which assess audio-lip synchronization accuracy. We adopt the cosine similarity (CSIM) to evaluate the identity consistency.

Implementation Details. We employ four V100 GPUs with a memory size of 32GB to train our audio-to-skeleton prediction model. We set the batch size to 128 and adopt the Adam optimizer [14] with a learning rate of 5e-5 to train this model for 2,000,000 steps. With the obtained skeleton sequences, we employ one general human video generation model named StableAnimator [49] and one co-speech gesture video generation model named EchoMimicV2 [42] for final video generation, respectively. They are denoted as "Ours (StableAnimator)" and "Ours (EchoMimicV2)" in Figure 5, Figure 7, and Table 2. It is worth noting that EchoMimicV2 adopts the hand skeletons rather than full-body skeletons as auxiliary control condition, so we extract the hand skeletons from our produced full-body skeletons for "Ours (EchoMimicV2)". For evaluations on our CSG-405 database, we further employ one general human video generation method named MimicMotion [21] to evaluate the generalizability of our audio-to-skeleton prediction model, which is denoted as "Ours (MimicMotion)".

4.1. Qualitative Comparisons

We perform qualitative comparisons between our approaches and state-of-the-art methods in Figure 5, Figure 7 and Figure 6. The methods in comparison include co-speech gesture generation methods such as S2G-MDDiffusion [19], DiffTED [18], EchoMimicV2 [42], VLOGGER [24] and CyberHost [23], as well as general full-body human video generation methods like MimicMotion [21] and StableAnimator [49].

First, it is shown that the warping-based methods [18, 19] tend to produce blurry body parts, e.g., hands, when encountering large pose changes, as highlighted in Figure 7. In contrast, our method is warping-free and able to produce clearer body parts. Second, Figure 5 and Figure 7 show that methods that require driven skeleton condition, e.g., MimicMotion [21], StableAnimator [49], and EchoMimicV2 [42], may produce body shapes that contradict with those of the reference images in the unpaired setting. This is because the driven skeleton condition they adopt is extracted from another video. In comparison, our method significantly reduces this error by generating skeletons according to the body shape revealed in the reference image.

Since the VLOGGER [24] and CyberHost [23] models are not publicly available, we perform qualitative comparisons with them using demo videos from their respective project homepages. As shown in Figure 6, our approach surpasses both VLOGGER and CyberHost in the naturalness and diversity of synthesized gestures.

4.2. Quantitative Comparisons

We further present quantitative comparisons in Table 2. Our methods achieve the best overall performance across the PATS, TED-talks, and CSG-405 datasets in the unpaired setting.

As presented in the table, existing human video generation methods, e.g., StableAnimator [49], demonstrate strong performance in the paired setting, suggesting their strong potential for co-speech gesture video generation. However, these approaches exhibit significant degradation under unpaired conditions, which is primarily due to the misalignment between the driven skeleton condition and the reference image. Our methods address this limitation by adaptive skeleton generation according to the body shape of the reference image, achieving state-of-the-art results in unpaired scenarios. In the paired setting, StableAnimator achieves the best SSIM and PSNR metrics. This is because they directly utilize ground-truth skeletons as a condition, while we generate skeletons with randomness from audio. It is also worth noting that EchoMimicV2 achieves suboptimal performance in paired settings. This is because it tends to generate videos with the speaker facing forward and positioned strictly centrally, which can be misaligned with the original co-speech gesture video. We also conduct experi-

Table 2. Quantitative comparisons between our method and state-of-the-art methods on PATS [1], TED-talks [6], and our CSG-405 datasets. Ours† (StableAnimator) means that we fine-tune StableAnimator on our CSG-405 database.

Dataset	Method	Paired		Unpaired				
		SSIM↑	PSNR↑	CSIM↑	FID↓	FVD↓	Sync-C↑	Sync-D↓
PATS	S2G-MDDiffusion [19]	0.59	16.84	**0.77**	85.23	989.98	3.94	10.47
	EchoMimicV2 [42]	0.57	16.35	0.73	97.86	1243.48	5.44	9.01
	StableAnimator [49]	**0.75**	**22.93**	0.63	91.83	1131.73	4.05	9.88
	Ours (EchoMimicV2)	0.56	15.70	0.73	89.36	1113.50	5.41	**9.00**
	Ours (StableAnimator)	0.56	16.35	0.76	73.56	928.26	5.22	9.72
	Ours† (StableAnimator)	0.60	16.68	**0.77**	**70.38**	**877.07**	**5.49**	9.36
TED-talks	DiffTED [18]	0.71	18.53	0.75	83.80	1138.29	0.74	13.05
	EchoMimicV2 [42]	0.66	18.09	0.71	86.88	943.08	4.58	9.65
	StableAnimator [49]	**0.81**	**25.50**	0.55	88.29	890.48	4.10	10.20
	Ours (EchoMimicV2)	0.65	17.67	0.74	77.46	785.47	4.88	9.41
	Ours (StableAnimator)	0.61	17.18	0.75	**68.84**	670.21	5.44	9.21
	Ours† (StableAnimator)	0.63	17.30	**0.76**	70.05	**650.20**	**5.59**	**9.15**
CSG-405	S2G-MDDiffusion [19]	0.55	14.02	0.39	216.80	2538.70	2.48	11.54
	DiffTED [18]	0.67	15.98	0.61	138.82	1852.10	0.66	13.06
	MimicMotion [21]	0.75	20.47	0.60	115.65	1751.33	3.26	11.12
	EchoMimicV2 [42]	0.71	17.88	**0.82**	90.01	1444.00	6.53	8.09
	StableAnimator [49]	**0.83**	**24.27**	0.67	94.94	1532.74	5.13	9.50
	Ours (MimicMotion)	0.59	14.98	0.76	73.77	1264.22	4.43	10.36
	Ours (EchoMimicV2)	0.69	17.02	**0.82**	70.03	1111.61	**6.59**	**8.04**
	Ours (StableAnimator)	0.65	16.26	0.81	66.47	988.43	5.82	9.11
	Ours† (StableAnimator)	0.67	16.62	**0.82**	**62.93**	**984.13**	6.28	8.68

Figure 5. Qualitative comparisons between our methods and state-of-the-art methods on our CSG-405 database.

Table 3. Ablation study on each key component of our model on CSG-405 dataset.

Method	SSIM↑	PSNR↑	CSIM↑	FID↓	FVD↓	Sync-C↑	Sync-D↓
w/o Reference Skeleton	0.57	13.75	0.57	127.40	1704.49	4.88	9.94
Cross-Attention Conditioning	0.63	15.71	0.76	78.10	1023.52	0.77	13.64
w/o Local Representation	0.62	15.02	0.71	86.69	1073.47	0.49	13.59
Ours	**0.67**	**16.62**	**0.82**	**62.93**	**984.13**	**6.28**	**8.68**

ments using the StableAnimator finetuned on our CSG-405 database and observed improved performance. This indicates that our database can further enhance the performance of existing human video generation models for co-speech gesture video generation.

4.3. Ablation Study

Finally, we conduct ablation study on our CSG-405 database to justify each key design in our audio-to-skeleton prediction model. We adopt the finetuned StableAnimator as the video generator and summarize experimental results in Table 3.

Effectiveness of the Reference Skeleton. In this experiment, we evaluate whether it is beneficial to adopt the reference skeleton $\boldsymbol{s}_r$ as the input of our audio-to-skeleton prediction model. This experiment is denoted as "w/o Reference Skeleton" in Table 3. It is shown that its performance is significantly lower than our audio-to-skeleton prediction model. This is reasonable as it loses the prior about the target speaker's body shape reflected in the reference image. Therefore, it tends to produce skeletons with incorrect body shapes, resulting in unrealistic co-speech gesture videos.

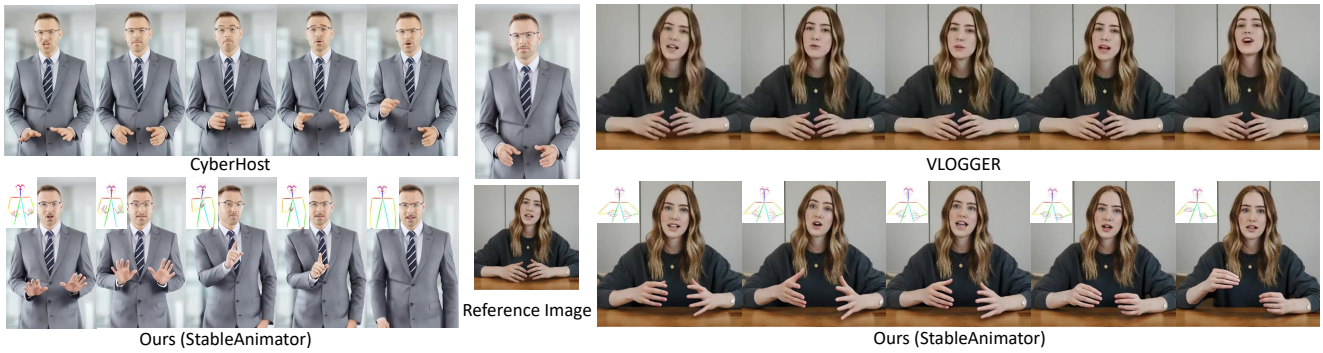


Figure 6. Qualitative comparisons between CyberHost [23], VLOGGER [24], and our approach.

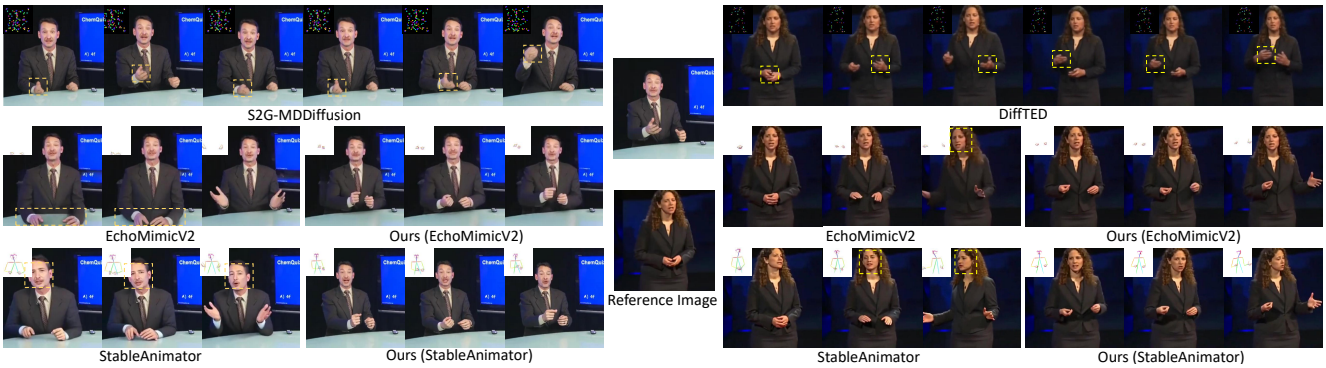


Figure 7. Qualitative comparisons between our methods and state-of-the-art methods on PATS [1] and TED-talks [6] databases.

Effectiveness of Our Audio Conditioning. As described in Section 3.1, we concatenate the embeddings of the skeleton and audio segments along the feature dimension as the input of the diffusion model. In this experiment, we justify that this audio conditioning strategy outperforms the popular cross-attention method [13]. As shown in Table 3, our audio conditioning mechanism achieves better Sync-C and Sync-D metrics, indicating better performance in synchronization between the audio condition and synthesized gestures. This is because compared with the common cross-attention strategy, our approach enforces fine-grained correspondence between the audio and skeleton segments.

Effectiveness of Local Motion Representation. As introduced in Section 3.2, we adopt the local motion representation [26] for keypoints on the face and hands in our model. As revealed in Table 3, the local motion representation leads to more realistic videos with better audio-gesture synchronization. The reason is that it disentangles the global skeleton motion with the subtle facial and hands movements, enabling our model focus more on generating expressive facial expressions and hand movements.

5. Conclusion

In this paper, we present a novel co-speech gesture video generation framework based on audio-conditioned skeleton prediction. Our method bridges the audio-visual alignment challenge by integrating reference-aware skeleton generation with feature-level audio-skeleton fusion. Furthermore, we introduce CSG-405, the first large-scale public dataset with diverse speech scenarios and high-quality annotations, which addresses the data scarcity problem in existing research. Experiments demonstrate that our approach significantly outperforms current state-of-the-art methods in both video quality and audio-visual synchronization. By releasing both the technical framework and dataset, we aim to democratize high-fidelity co-speech gesture synthesis.

Broader Impacts. This work democratizes high-fidelity co-speech gesture video generation by proposing a audio-to-skeleton prediction model and a large-scale dataset. While our work advances co-speech gesture video synthesis for human-centered applications, we advocate for the development of ethical frameworks (e.g., synthetic content detection tools) alongside generative technologies to ensure their responsible deployment and societal well-being.

Acknowledgement. This work was supported by the 2024 Tencent AI Lab Rhino-Bird Focused Research Program, the National Natural Science Foundation of China under Grant 62476099 and 62076101, Guangdong Basic and Applied Basic Research Foundation under Grant 2024B1515020082 and 2023A1515010007, the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004, and the TCL Young Scholars Program.

References

[1] C. Ahuja, D. Lee, Y. Nakano, L. Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *ECCV*, 2020. 2, 3, 6, 7, 8

[2] J. Karras, A. Holynski, T. Wang, I. Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *CVPR*, 2023. 3

[3] X. Liu, Q. Wu, H. Zhou, Y. Du, W. Wu, D. Lin, Z. Liu. Audio-Driven Co-Speech Gesture Video Generation. In *NeurIPS*, 2022. 2, 3, 4, 6

[4] F. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. In *TPAMI*, 1989. 3

[5] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, A. Bermano. Human motion diffusion model. In *ICLR*, 2023. 4

[6] Y. Yoon, W. Ko, M. Jang, J. Lee, J. Kim, G. Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *ICRA*, 2019. 2, 3, 6, 7, 8

[7] A. Siarohin, O. Woodford, J. Ren, M. Chai, S. Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 3

[8] J. Zhao, H. Zhang. Thin-plate spline motion model for image animation. In *CVPR*, 2022. 3

[9] Z. Cao, T. Simon, S. Wei, Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[10] Z. Yang, A. Zeng, C. Yuan, Y. Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 2023. 3, 5

[11] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 3

[12] J. Ho, A. Jain, P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 5, 8

[14] D. Kingma, J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[15] T. Unterthiner, S. VanSteenkiste, K. Kurach, R. Marinier, M. Michalski, S. Gelly. Fvs: a new metric for video generation. In *ICLR*, 2019. 6

[16] L. Hu, X. Gao, P. Zhang, K. Sun, B. Zhang, L. Bo. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. In *CVPR*, 2024. 2, 3

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4

[18] S. Hogue, C. Zhang, H. Daruger, Y. Tian, X. Guo. DiffTED: One-shot Audio-driven TED Talk Video Generation with Diffusion-based Co-speech Gestures. In *CVPR Workshop*, 2024. 2, 3, 6, 7

[19] X. He, Q. Huang, Z. Zhang, Z. Lin, Z. Wu, S. Yang, M. Li, Z. Chen, S. Xu, X. Wu. Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model. In *CVPR*, 2024. 2, 3, 4, 6, 7

[20] Z. Xu, J. Zhang, J. Liew, H. Yan, J. Liu, C. Zhang, J. Feng, M. Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, 2024. 2, 3

[21] Y. Zhang, J. Gu, L. Wang, H. Wang, J. Cheng, Y. Zhu, F. Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *ICML*, 2024. 2, 3, 5, 6, 7

[22] S. Zhu, J. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, S. Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, 2024. 2, 3

[23] G. Lin, J. Jiang, C. Liang, T. Zhong, J. Yang, Y. Zheng. CyberHost: Taming Audio-driven Avatar Diffusion Model with Region Codebook Attention. In *ICLR*, 2025. 2, 3, 6, 8

[24] E. Corona, A. Zanfir, E. Bazavan, N. Kolotouros, T. Alldieck, C. Sminchisescu. VLOGGER: Multimodal diffusion for embodied avatar synthesis. In arXiv:2403.08764, 2024. 2, 3, 6, 8

[25] W. Peebles, S. Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4

[26] S. Qian, Z. Tu, Y. Zhi, W. Liu, S. Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *ICCV*, 2021. 5, 8

[27] Y. Liu, Q. Cao, Y. Wen, H. Jiang, C. Ding. Towards variable and coordinated holistic co-speech motion generation. In *CVPR*, 2024. 4

[28] Z. Chen, J. Cao, Z. Chen, Y. Li, C. Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *AAAI*, 2025. 2, 3, 4

[29] L. Tian, Q. Wang, B. Zhang, L. Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. In *ECCV*, 2024. 3, 4

[30] Y. Ma, H. Liu, H. Wang, H. Pan, Y. He, J. Yuan, A. Zeng, C. Cai, H. Shum, W. Liu, Q. Chen. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. In *SIGGRAPH Asia*, 2024. 2, 3, 4

[31] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. Black. SMPL: A skinned multi-person linear model. In *TOG*, 2015. 3

[32] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, M. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 3

[33] T. Karras, T. Aila, S. Laine, J. Lehtinen. Progressive

growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 5

[34] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, C. Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 5

[35] S. Serengil, A. Ozpinar. A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. In *Bilisim Teknolojileri Dergisi*, 2024. 5

[36] R. Güler, N. Neverova, I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 3

[37] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, F. Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, 2023. 3, 4

[38] H. Wei, Z. Yang, Z. Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. In arXiv:2403.17694, 2024. 3, 4

[39] B. Egger, W. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, T. Vetter. 3d morphable face models—past, present, and future. In *TOG*, 2020. 4

[40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[41] H. Zhang, Z. Liang, R. Fu, B. Liu, Z. Wen, X. Liu, C. Li, Y. Liang. Efficient Long-duration Talking Video Synthesis with Linear Diffusion Transformer under Multimodal Guidance. In https://arxiv.org/abs/2411.16748, 2024. 2, 4

[42] R. Meng, X. Zhang, Y. Li, C. Ma. EchoMimicV2: Towards Striking, Simplified, and Semi-Body Human Animation. In *CVPR*, 2025. 2, 3, 5, 6, 7

[43] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, Gpt-4 technical report. In arXiv:2303.08774, 2023. 5

[44] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, P. Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 5

[45] A. Plaquet, H. Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *INTERSPEECH*, 2023. 5

[46] J. hung, A. Zisserman. Out of time: automated lip sync in the wild. In *ACCV Workshop*, 2016. 5, 6

[47] A. Baevski, Y. Zhou, A. Mohamed, M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020. 4

[48] J. Ho, T. Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 5

[49] S. Tu, Z. Xing, X. Han, Z. Cheng, Q. Dai, C. Luo, Z. Wu. StableAnimator: High-Quality Identity-Preserving Human Image Animation. In *CVPR*, 2025. 2, 3, 4, 5, 6, 7

[50] Y. Guo, M. Tan, Z. Deng, J. Wang, Q. Chen, J. Cao, Y. Xu, J. Chen. Towards Lightweight Super-Resolution With Dual Regression Learning. In *TPAMI*, 2024. 3

[51] J. Guan, Q. Yang, K. Wang, H. Zhou, S. He, Z Xu, H. Feng, E. Ding, J. Wang, H. Xie, Y. Zhao, Z. Liu. TALK-Act: Enhance Textural-Awareness for 2D Speaking Avatar Reenactment with Diffusion Model. In *SIGGRAPH Asia*, 2024. 3

[52] Z. Huang, F. Tang, Y. Zhang, X. Cun, J. Cao, J. Li, T. Lee. Make-your-anchor: A diffusion-based 2d avatar generation framework. In *CVPR*, 2024. 3

[53] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli. Image quality assessment: from error visibility to structural similarity. In *TIP*, 2004. 6

[54] A. Horé, D. Ziou. Image quality metrics: PSNR vs. SSIM. In *ICPR*, 2010. 6