Co-speech Gesture Video Generation with 3D Human Meshes

Aniruddha Mahapatra[1]*, Richa Mishra[1]*, Renda Li[2,3], Ziyi Chen[4], Boyang Ding[2,3], Shoulei Wang[2,3], Jun-Yan Zhu[1], Peng Chang[4], Mei Han[4], and Jing Xiao[3]

[1] Carnegie Mellon University
[2] University of Science and Technology of China
[3] Ping An Technology
[4] PAII Inc.

Abstract. Co-speech gesture video generation is an enabling technique for many digital human applications. Substantial progress has been made in creating high-quality talking head videos. However, existing hand gesture video generation methods are primarily limited by the widely adopted 2D skeleton-based gesture representation and still struggle to generate realistic hands. We introduce an audio-driven co-speech video generation pipeline to synthesize human speech videos leveraging 3D human mesh-based representations. By adopting a 3D human mesh-based gesture representation, we present a mesh-grounded video generator that includes a mesh texture map optimization step followed by a conditional GAN network and outputs photorealistic gesture videos with realistic hands. Our experiments on the TalkSHOW dataset demonstrate the effectiveness of our method over 2D skeleton-based baselines.

1 Introduction

Co-speech video generation [10] has emerged as a challenging research topic in visual content creation, offering huge potential for a wide array of applications in human-AI interactions, especially in the new era of conversational agents [28]. Co-speech video synthesis aims to generate a photorealistic and synchronized human talking video that aligns with a given speech, particularly lip movements, facial expressions, and accompanying gestures. Recently, significant progress has been made in realistic talking head video generation, including approaches building character-specific models [12, 23] and others focusing on general models applicable for any given character [11, 31, 37, 43, 57]. Nonetheless, those approaches are mostly designed around talking face generation and do not naturally extend to the upper body, especially the highly dynamic and flexible human hands.

Previous co-speech gesture video synthesis works heavily rely on 2D skeletons as the intermediate representation for image and video synthesis networks [10, 21, 33]. One limitation of this 2D skeleton-based representation is its lack of 3D

*indicates equal contribution.



Fig. 1: Given an input audio sequence and a person's identity, our audio-driven co-speech video generation method can create videos with realistic appearances, natural facial expressions, and hand gestures. Here we present video sequences generated from input audio featuring two different speakers. Some areas may appear blurry, representing natural motion blur in the videos. For additional video results, please check our website `https://cospeech-gesture-3d.github.io/website`.

body shape modeling throughout the generation process, from motion prediction to video rendering steps. Consequently, those methods often struggle to produce natural or physically plausible gesture videos, resulting in blurred or sometimes even missing parts of arms and fingers.

Other works on speech-driven gesture generation [1, 13, 52, 54, 59] adopt 3D mesh-based representation for gestures. However, their main goal is to generate gestures in the chosen 3D parameterization space rather than generating the final videos, which is vital for downstream applications.

In this paper, we present a complete pipeline to generate a co-speech video for a target speaker with realistic facial animation and hand gesture movements. Starting with an audio input, we first predict human body shapes and movements for the face, body, and hands, using a 3D human mesh model [30]. Given the estimated 3D human body model (SMPL-X) and its gestures, we use an optimization-based approach to derive a UV texture map corresponding to this SMPL-X model. Finally, we introduce a GANs-based video generation network to transform the textured meshes into the final photorealistic video. We observe that these optimized 3D textured meshes are highly effective as an intermediate representation for co-speech video synthesis. Our code and additional results are available on our website.

2 Related work

Co-speech Gesture Video Generation remains under-explored compared to talking head generation research [11, 23, 24, 31, 37, 43, 48, 53, 57]. In early

work, Ginosar et al. [10] proposed a cross-modal translation network to generate 2D skeleton-based gestures from a given speech, and then the gestures are further synthesized into video with a standalone pose-to-image network [5]. Qian et al. [33] and Liu et al. [21] further improved the 2D skeleton-based gestures by introducing the learning of template vectors and codebooks containing both common gesture patterns and rhythmic movement details, respectively, followed by a pose-to-image generator. Although exciting results have been obtained, the 2D skeleton representation inherently lacks the 3D shape constraint and often leads to unnatural and physically inaccurate gesture configurations, degrading the quality of the generated videos. Concurrent with our work, Make-Your-Anchor [15] also explored the use of 3D representation but deployed a 2D diffusion model to perform the translation.

Co-speech Human Motion Generation. Co-speech motion generation has been widely explored over the years. While certain studies specialize in facial generation [7, 8, 18, 35], other works specifically target body gestures [1, 10, 21, 33, 52, 54, 59]. However, using a unified framework to generate holistic motions, including face, body, and hands, is still challenging. Habibie et al. [13] generates the 3D mesh of the face and the keypoints of the body and hands given an input voice, but the body mesh is not generated. The introduction of a series of human body parametric models [3, 19, 22, 30] makes it possible to generate a coherent human body 3D mesh as a whole. Recently, Yi et al. [54] enables full-body 3D mesh generation by separately employing an encoder-decoder for facial motion generation and a cross-modal VQ-VAE [41] for synthesizing body and hand motion. We build upon this work but focus on co-speech video generation.

Conditonal Video Generation. Several works [5, 38, 46, 58] proposed transferring motion from a source video to animate a target image. They identify the keypoint correspondence between the source frames and the target image and compute the dense optical flow using local deformations of the paired correspondences. In our setting, however, we do not have a source driving RGB video, but rather, the audio input. Other works focus on unconditional video generation from random latents [39, 40, 42, 50], but do not support audio or motion conditions. Several recent and concurrent works explored the use of diffusion models for video generation and editing [4, 6, 9, 32, 49, 51]. Despite their versatility across diverse objects, those methods have yet to show effectiveness and efficiency in generating human videos. The most closely related methods in video generation are Vid2Vid [44] and WC-Vid2Vid [25]. These approaches synthesize RGB videos from sequences of segmentation maps or edge maps. However, these methods require training on large datasets, which differs from our problem setup, where we have only a single, relatively short video of a character.

3 Method

Our goal is to create a co-speech gesture video given aninput audio — a challenging task due to the limited data and camera viewpoints for each subject.

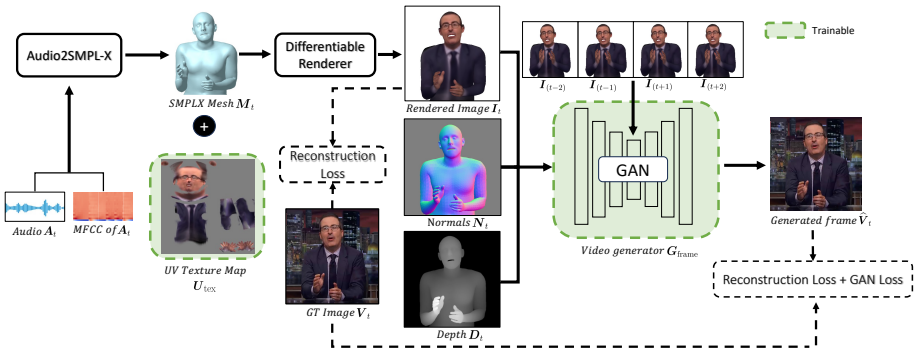


Fig. 2: Overview. Given an input audio $\boldsymbol{A}_t$, we first train a network to predict the plausible human motion of face, hands, and body, as denoted as SMPL-X mesh $\boldsymbol{M}_t$ for each frame in Section 3.1. Here t is the frame index. Subsequently, we optimize the UV texture map $\boldsymbol{U}_{\text{tex}}$ for each character using differentiable rendering so that we can reconstruct all video frames given the texture map. Finally, in Section 3.2, we design a conditional GAN-based video generator $\boldsymbol{G}_{\text{frame}}$ to synthesize the final video given the current 2D rendered image $\boldsymbol{I}_t$, the normal map $\boldsymbol{N}_t$, the depth map $\boldsymbol{D}_t$, as well as 2D rendered images from nearby frames. The video generator is trained with a combination of reconstruction loss and GAN loss.

To address this challenge, we decompose this task into three steps. First, given an audio input, we predict plausible motions of the face, hands, and body in Section 3.1. Second, we learn to recover the character's appearance given the estimated geometry and motion parameters. Given the estimated appearance, geometry, and motion, we train a conditional GAN model to synthesize the final video in Section 3.2.

3.1 Audio to Mesh

We first predict a plausible human motion sequence of face, hands, and body given the audio input based on the TALKSHOW model [55], which regresses the parameters of the SMPL-X model [30] from the audio. The parameters include shape, pose, and expression coefficients. Once we predict these parameters, we use the SMPL-X model to get the mesh corresponding to each frame. Below, we briefly describe the architecture of the TALKSHOW model [55].

Face Motion Generator. The face motion generator aims to predict facial movements and expressions, given input audio and the speaker's identity. It employs an encoder-decoder architecture, where the encoder uses the pre-trained WAV2VEC2 model [2], and the decoder consists of six layers of temporal convolutional networks (TCNs). The encoder first processes the audio sequence to extract audio embeddings. These embeddings are then concatenated with the speaker's identity and passed through the decoder to obtain the facial expres-

sions and pose parameters. The network is trained with the mean squared error (MSE) loss between the predicted and ground truth parameters.

Body and Hand Motion Generator. The body and hand generator adopts two-stage training to generate realistic, diverse, and continuous body and hand poses. Specifically, in the first stage, two independent VQ-VAE models [41] are used to learn compositional quantized motion codebooks of the body and hand, which improves the diversity of the generation. To ensure the temporal consistency of poses generated from the separate body and hand codebooks, we adopt Gated PixelCNN [27] to model the interaction conditions between body and hand movements.

Consistent with the TALKSHOW model [55], we use standard VQ-VAE loss and reconstruction loss for both training stages. Once we have trained the model, given input audio $\boldsymbol{A}$, for a particular speaker, we can obtain a sequence of SMPL-X meshes, $\boldsymbol{M}_{1:T} = \{\boldsymbol{M}_t\}_{t=1}^{T}$, featuring mouth movements and hand gestures, specific to this speaker, where $t = \{1...T\}$, and T denotes the total number of meshes in the sequence.

3.2 Mesh-Grounded Video Generation

Given the generated mesh sequence, our objective is to generate a photorealistic video following the speaker's audio input. We leverage conditional GAN [5,17,25, 44] to generate the photorealistic video conditioned on the sequence of SMPL-X meshes $\boldsymbol{M}_{1:T}$.

Challanges. However, directly using these untextured SMPL-X meshes as conditioning for the conditional GAN model can lead to artifacts in the generated video. Notably, inconsistencies arise in the hands of the speaker between the generated video and the input mesh. This discrepancy becomes more severe when the training video only consists of a limited number of frames. We hypothesize that generating realistic hands using conditional GAN from untextured meshes is a significant challenge, primarily due to the large motion blur in the ground-truth training frames. The artifacts are predominantly visible in the hands, as they exhibit more motion than other relatively stationary body parts.

To address this issue, we propose a method that first optimizes the UV texture map from the training video for the SMPL-X meshes. Subsequently, we use these textured meshes as additional conditioning to generate the final video. Using these textured meshes helps us to ground the appearance of different body regions.

Mesh Texturing. Given a sequence of SMPL-X meshes, denoted as $\boldsymbol{M}_{1:T}$, we aim to optimize a UV texture map, $\boldsymbol{U}_{\text{tex}} \in \mathbb{R}^{N\text{x}N\text{x}3}$, shared across all the meshes and initialized with standard normal distribution $\mathcal{N}(0, \mathbb{I}^{N\text{x}N\text{x}3})$. More specifically, given a sequence of cameras, $\boldsymbol{P}_{1:T}$, where the t-th camera corresponds to the t-th mesh, we project the meshes from 3D world space to a 2D image using a differentiable rendering function $\mathcal{F}(\cdot)$ [34]. We then measure the per-pixel difference, expressed as the $\mathbf{L}2$ norm between the 2D rendered image $\boldsymbol{I}_t$ of

the mesh $\boldsymbol{M}_t$, and the ground-truth frame $\boldsymbol{V}_t$ for the t-th frame. Formally, we optimize the UV texture map according to the following objective:

$$\begin{aligned} \boldsymbol{U}_{\text{tex}} = \arg\min_{\boldsymbol{U}_{\text{tex}}} \sum_{t=1}^{T} \|\boldsymbol{I}_t - \boldsymbol{V}_t\|^2, \\ \text{where} \quad \boldsymbol{I}_t = \mathcal{F}(\boldsymbol{M}_t, \boldsymbol{P}_t, \boldsymbol{U}_{\text{tex}}). \end{aligned} \tag{1}$$

Here, the differentiable rendering function $\mathcal{F}(\cdot)$ takes the camera pose $\boldsymbol{P}_t$, the mesh $\boldsymbol{M}_t$, and the shared texture map $\boldsymbol{U}_{\text{tex}}$, and produces a 2D rendered image $\boldsymbol{I}_t$. Once we learn a UV texture map $\boldsymbol{U}_{\text{tex}}$, after many optimization iterations, we can render a sequence of textured meshes $\boldsymbol{I}_{1:T}$, from their original untextured forms $\boldsymbol{M}_{1:T}$.

As the meshes and the ground-truth exhibit small misalignment, we observe that the learned textures within the UV texture map are slightly blurry and tend to be an average representation of adjacent regions in the ground-truth images. However, this does not significantly impede our approach. We rely on the GAN model to learn the high-frequency details and adapt the mesh structure for generating photorealistic video. Notably, the textured meshes provide rough color and shape guidance for the GAN model.

Discussion. We have also experimented with more expressive neural 3D representations with view-dependent effects, such as dynamic NeRF representation (e.g., INST [60]). However, these models, while capable of producing highly detailed textures and fine-grained details for hair and face, tend to introduce severe artifacts for hands due to their complex deformations and self-occlusion. These artifacts can negatively impact the downstream video generation results. Therefore, we have chosen a less detailed but more robust 3D textured mesh representation and rely on a GANs-based video generator to add details, textures, and realistic lighting effects.

Additionally, we tried an alternate approach to learn per-vertex textures instead of a UV texture map using a similar optimization-based approach. We observed that both approaches yield very similar results.

Video Generation with Textured Meshes. As previously mentioned, we use an image-conditional GAN model [17,45] to generate a photorealistic video from a sequence of textured meshes. Specifically, we use a UNet [36] based generator $\boldsymbol{G}_{\text{frame}}$. Instead of generating all the frames simultaneously, we generate each frame $\boldsymbol{V}_t$ sequentially conditioned on the rendered image $\boldsymbol{I}_t$. We use SPADE [29] normalization to incorporate conditioning in $\boldsymbol{G}_{\text{frame}}$. To ensure temporal consistency across the generated frames, Vid2Vid [44] and WC-Vid2Vid [25] conditions the generation of frame $\widehat{\boldsymbol{V}}_t$ on preceding k generated frames. However, we observe that this can lead to error propagation in the subsequent frames. Instead, we condition the generation of frame $\widehat{\boldsymbol{V}}_t$ on a context window of $2k$ textured meshes: k prior and k subsequent frames.

$$\widehat{\boldsymbol{C}}_t = \{\boldsymbol{I}_{t-k}, \boldsymbol{I}_{t-k+1}, ..., \boldsymbol{I}_{t+k-1}, \boldsymbol{I}_{t+k}\} \tag{2}$$



Fig. 3: Example sequences generated by our method. Here we show two video sequences generated by our method. For each sequence, we visualize both the textured meshes and the final image results. Our method is able to synthesize realistic video results with natural facial expressions and hand gesture movements. Please refer to our *website* for more video results.

$$\widehat{\boldsymbol{V}}_t = \boldsymbol{G}_{\text{frame}}(\boldsymbol{I}_t, \widehat{\boldsymbol{C}}_t) \tag{3}$$

We find that this strategy ensures temporal consistency in the generated photorealistic video, V, without encountering the problem of error propagation. Similar to Mensha et al. [26], we also condition the generation of $\widehat{\boldsymbol{V}}_t$ on the depth $\boldsymbol{D}_t$ and the rendered normal map $\boldsymbol{N}_t$ obtained from $\boldsymbol{M}_t$.

$$\widehat{\boldsymbol{V}}_t = \boldsymbol{G}_{\text{frame}}(\boldsymbol{I}_t, \boldsymbol{D}_t.\boldsymbol{N}_t, \widehat{\boldsymbol{C}}_t). \tag{4}$$

Our training objective involves $L1$ reconstruction loss $\mathcal{L}_{\text{L1}}$, conditional GAN loss $\mathcal{L}_{\text{GAN}}$ [17], discriminator feature matching loss [45]$\mathcal{L}_{\text{FM}}$, and VGG-based perceptual loss [56] λ_{VGG}.

$$\mathcal{L} = \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{VGG}}\mathcal{L}_{\text{VGG}} + \mathcal{L}_{\text{GAN}} + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}}, \tag{5}$$

where $\lambda_{\text{L1}}, \lambda_{\text{FM}}$, and λ_{VGG} are scalar coefficients of $L1$ reconstruction loss, feature matching loss [45], and VGG-based perceptual loss [56], respectively. For more details, please refer to the Section 4.

4 Experiments

Datasets. For our work, we have used the dataset of speakers provided by Talk-SHOW [54]. The dataset contains the identities of 4 different speakers, namely,

"Oliver", "Chemistry", "Conan", and "Seth". We have taken the longest video out of a set of videos for each of the four speakers. This is because, except for Oliver, most of the videos for the other speakers contain less than 10K frames in total, which, in our observation, were insufficient to train the GAN model properly. Each video is at 30 fps and 720 x 1280 resolution. We take a square crop of the video to include the upper body of the speaker, including the hands, as we are mainly focused on generating a photorealistic talking video of a speaker with character-specific hand gestures. The cropped video frames are resized to 512x512 resolution. We also determine perspective cameras for each of the cropped videos for the four different speakers. This is essential as the meshes provided by the TalkSHOW [54] dataset contain full-body meshes. Hence, to use cropped videos, we need to align the projection of the meshes with our cropped videos. We split each video into training and testing splits such that the first 80% of the frames are used for training and the remaining for test. To train the frame generator model, $\boldsymbol{G}_{\text{frame}}$, we also generate rendered textured meshes using Section 3.2 and their depth and normal maps.

Implementation Details. Following the settings in [54], the clip length and total epoch in the three training stages are 88 and 100, respectively. The batch size when training the face, VQ-VAEs, and body and hand generator are 1, 128, and 128, and the learning rates are set to constant 1×10^{-3}, 1×10^{-4} and 1×10^{-4} respectively. When training the body and hand generator, the weights of the two VQ-VAEs of the last epoch are loaded and frozen.

For mesh texturing, we optimize the parameters of a UV texture map as described in Section 3.2. The UV texture map is of shape NxNx3, where $N = 256$ is in our case. This is to make the optimization process faster. We also tried with $N = 512$ but did not notice visible improvements. We perform the optimization using $\mathbf{L}2$ loss of the mesh rendering w.r.t the corresponding ground-truth training video frames. Meshes are rendered at 256x256, and ground-truth video frames are resized to 256x256. We also tried rendering the meshes and optimizing at 512x512 but found the learning process very slow. Once the UV maps are learned, we render the textured meshes for the video frames at 512x512 resolution for training and evaluation.

For training the video frame generation model, $\boldsymbol{G}_{\text{frame}}$, we use a learning rate of 2×10^{-3} and use the TTUR method to update the learning rate [14] and train for 50 epochs. We train the model with a batch size of 16. The training and inference are set at 512x512 resolution. We emperically set $\lambda_{\text{L1}} = 10.0, \lambda_{\text{FM}} = 10.0$, and $\lambda_{\text{VGG}} = 1.0$. We set the value of $k = 2$ in Section 3.2, i.e., we use two previous and two subsequent textured meshes as context to generate frame at time t. We train our model for 50K iterations on a single RTX6000 GPU (16GB VRAM) with a batch size of 8. Each training iteration takes 0.54s. Inference requires 9GB VRAM and takes 0.013s per frame.

Baselines. To verify the importance of using 3D human body representation for high-quality speaker video generation, we compare our method against a baseline that uses a 2D keypoint as the human body representation, a widely utilized method for human video generation. Our baseline utilization draws inspiration

from prior works such as Ginosar et al. [10], which employs 2D keypoint as the representation of body gestures, and Lu et al. [23], which utilizes 2D facial keypoint to generate talking head video. In our baseline, we extract the joints from the corresponding SMPL-X mesh for the upper body, the face and the hands. This ensures that the temporal consistency for our baseline is the same as our method. In the context of the video frame generation model, we maintain the same training setup as our proposed approach. Although Ginosar et al. [10], Lui et al. [21], and Qian et al. [33] also generate co-speech gesture video, we could not compare them directly. The code is not publicly available for Lui et al. [21]. Additionally, Ginosar et al. [10] and Qian et al. [33], provide parts of their code for generating speaker-specific 2D skeleton keypoints from audio input but not for generating the RGB final video.

Evaluation. We evaluate the quality of video generated on the test sequence for each speaker in the dataset. Under this setting, we already have the ground-truth (untextured) SMPL-X meshes and corresponding ground-truth RGB video frames. We first texturize the (untextured) SMPL-X meshes using the learned UV texture map and then generate the photorealistic video with the GAN model. Since we already know the corresponding ground-truth RGB frames, we can use pixel-aligned metrics like PSNR, SSIM, and LPIPS [56] to compare the quality of the generated video. Note that in our work, we mainly focus on the photorealistic generation of the speaker and not the background. Thus we use modified PSNR and LPIPS that only take into account the speaker pixels, ignoring the background. In addition, not considering the background for evaluation is important because the ground-truth frames are clipped from a very long video, causing the background shifts over time due to subtle camera movement. This results in shifting and flickering of the background in the generated videos, which undermines the importance of evaluating the key component of our method. The mask of the foreground speaker is obtained using a Video Matting method [20] applied on the ground-truth sequence. The implementation of masked PSNR and LPIPS is taken from pix2latent [16].

User Study. In addition to PSNR, SSIM, and LPIPS [56], to further conduct a user study to to assess the quality of our final method, where the video is generated directly from input audio against the baseline method, our method (w/o textured meshes), and our method (w/o depth and normal maps). Since we do not have the corresponding ground-truth video for end-to-end evaluation, we only rely on user study to verify the effectiveness of our full method. We perform a paired comparison, where the Amazon MTurkers are asked to select which video is more realistic for the same input audio.

Details. For the user study, each video was assigned to 50 MTurkers. The videos from part of the test sequence are approximately 30 seconds long, whereas the videos generated from input audio are 8 seconds long on average. In their evaluation, we requested the participants to (1) ignore the shifts in the background video because of the aforementioned reason and (2) pay more attention to the quality of generated arms and hands.

Table 1: Baseline and ablation comparison. Here, we compare our full method (ours) with two ablated methods: ours (w/o textured mesh), ours (w/o depth and normal), and a 2D keypoint-based baseline. We conduct quantitative comparisons regarding video quality, regarding SSIM [47], PSNR, and LPIPS [56]. We observe that our method performs better than the baseline in all the metrics and that the majority of the performance gain in our method comes from the use of textured meshes.

Speaker	Method	SSIM (↑)	PSNR (↑)	LPIPS (↓)
Oliver	2D-keypoint Baseline	0.759	15.951	0.482
	Ours (w/o textured mesh)	0.783	16.233	0.411
	Ours (w/o depth and normal)	0.753	16.214	**0.407**
	Ours	**0.786**	**16.403**	0.408
Chemistry	2D-keypoint Baseline	0.759	14.862	0.411
	Ours (w/o textured mesh)	0.761	14.962	0.397
	Ours (w/o depth and normal)	0.776	14.969	**0.394**
	Ours	**0.777**	**15.021**	**0.394**
Conan	2D-keypoint Baseline	0.866	20.002	0.41
	Ours (w/o textured mesh)	0.89	20.112	0.41
	Ours (w/o depth and normal)	0.89	20.127	**0.398**
	Ours	**0.891**	**21.125**	**0.398**
Seth	2D-keypoint Baseline	0.787	14.589	0.399
	Ours (w/o textured mesh)	0.795	14.933	0.328
	Ours (w/o depth and normal)	0.798	14.862	0.334
	Ours	**0.801**	**15.044**	**0.312**

4.1 Mesh to Video

As mentioned in Section Section 4, we partition the entire video into training and test data. In this context, we evaluate the quality of video generation conditioned on the ground-truth untextured mesh from the test segments within each of the four speaker's video sequences.

Quantitative Comparison. Since the corresponding ground-truth frame sequences are available, we can assess the similarity and fidelity of the generated video in comparison to the ground-truth video. Table 1 presents a comparison between the frames generated by our method and a baseline method utilizing 2D keypoint as input conditioning (described in Section 4). It is noted that the metrics are computed only w.r.t the foreground speaker in the generated frames, and not the background. We observe that videos generated by our method consistently outperform the ones generated by the 2D keypoint-based baseline method on all the metrics and for all four speaker identities, demonstrating that our method generates videos that are more pixel-aligned with the ground-truth video sequence.

Qualitative Comparison. In Figure 4, we compare the quality of generated video using our method against video generated using a 2D keypoint-based baseline on all actors, Notably, even for actors with a considerably small training set, the videos generated by our method (Figure 4, 4-th row), employing rendering of textured meshes as conditioning, yields consistent hand regions. In contrast, the video generated by the 2D keypoint-based baseline displays significant artifacts within the hand region. The generated hands appear mostly indistinct and disappear in many frames. We hypothesize that discrepancy may arise from the grounding provided by the textured mesh, which ensures the permanence of the hands, while the 2D keypoint cannot provide a strong prior for the video generator, $\boldsymbol{G}_{\text{frame}}$.

4.2 Audio to Video

In this section, we evaluate our end-to-end method, i.e., the visual quality of the generated video given input audio.

Qualitative Comparison. We show videos generated using our method from input audio during inference, featuring two different actors, "Oliver" (Figure 3, top row) and "Chemistry" (Figure 3, bottom row), alongside their respective generated textured meshes. These two actors serve as very distinct settings, as "Oliver" contains a relatively large number of training frames (25K), whereas "Chemistry" is much more challenging (7K) frames. Although the meshes generated by Section 3.1 are in very novel poses compared to that seen by the frame generator during training, for both of the actors, the frame generator can generate a high-fidelity video of the actors following the structure in the input textured mesh. This further highlights the robustness of our method, leveraging 3D (textured) intermediate representation to accommodate novel poses and gestures during inference. We also conduct a user study in Table 2. We see that MTurkers prefer our video by a very large margin compared to the ones generated by the 2D-keypoint baseline method. Please check the video results on the *website*.

4.3 Albation Study

In the following section, we study the importance of using textured meshes (compared to untextured meshes) and using depth and normal maps as additional conditioning for video generation. We evaluate these variants on four videos using various automated metrics, including SSIM [47], PSNR, and LPIPS [56] in Table 1 and also conduct user preference study (in Table 2). Please check the video results in the *website*.

Mesh Texturing. From Table 1, we see that not using textured meshes hurts the performance of the generator the most in terms of PSNR, SSIM, and LPIPS [56]. From Table 2, we see that MTurkers prefer our video by a very large margin compared to our method using untextured meshes. Additionally, we include a visual comparison of results generated by our method with and without using textured

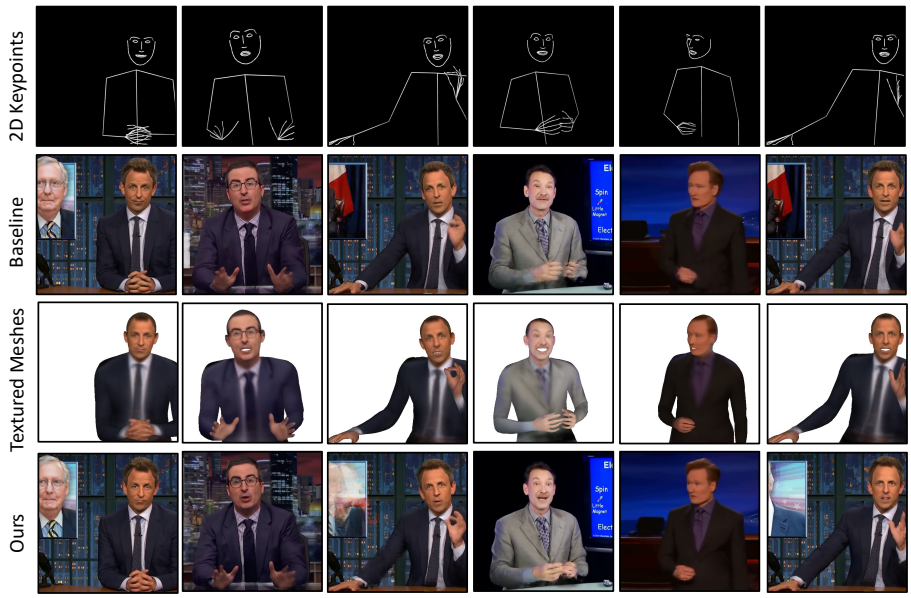


Fig. 4: Qualititative comparison of our method with the 2D keypoint baseline. Our method produces more realistic hands and gesture motion compared to the baseline that relies on 2D keypoint representation. Please refer to the *website* for more detailed comparisons.

Table 2: User Study. We perform a user study to evaluate the quality of the generated videos, comparing our method with baseline and ablations for four speakers. The videos are generated directly from an audio input. The numbers show the percentage of users who prefer videos generated by our method compared to ablations. Users prefer our method to the 2D-keypoint baseline, ours (w/o textured meshes), and ours (w/o depth and normal maps).

Method	Speaker			
	Oliver	Chemistry	Conan	Seth
2D-keypoint Baseline	70%	96%	84%	96%
Ours (w/o textured mesh)	68%	88%	86%	68%
Ours (w/o depth and normal)	54%	42%	58%	66%

meshes. Figure 5 (left) shows the loss of high-frequency details and blurry artifacts around the hands and the face region (highlighted by the red bounding box).

Depth and Normal Maps. As shown in Figure 5 and Table 1, the performance of our method also decreases slightly when we remove the depth and normal maps from the conditional inputs of the GAN-based generator. From Table 2, we draw a similar conclusion that MTurkers prefer videos generated by our method

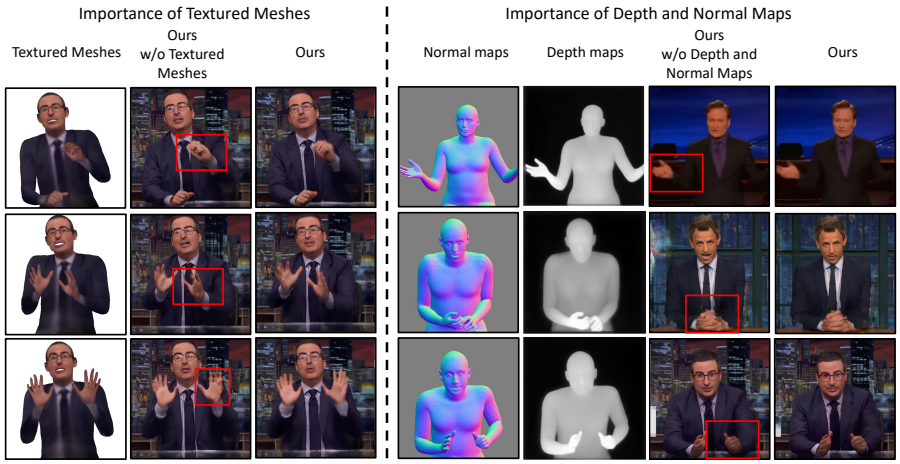


Fig. 5: Ablation study (qualitative comparison). We present an ablation study with qualitative comparisons between our full method and two ablated versions: ours w/o textured mesh (left) and ours w/o depth and normal (right). We observe that our method consistently produces more visually appealing results, with particular improvements noticeable in the hand regions (highlighted in the red bounding boxes).

in most cases, but not by a very large margin compared to the aforementioned ablation study. As highlighted in Figure 5 (right) red bounding box, ours (w/o depth and normal maps) synthesizes more blurry hand gestures. The blurriness may not be reflected significantly in automated metrics. However, it will reduce the video quality by a large margin.

In essence, from both Table 1 and Table 2, we observe that using textured meshes leads to significant improvement in the quality of generated videos, whereas using depth and normal maps as extra conditioning marginally benefits the video generation quality.

5 Discussion and Limitations

In summary, we have introduced a complete system designed to generate co-speech gesture videos from audio inputs. By employing explicit 3D representations of the human body, face, and hands, we have effectively bridged the gap between audio input and video output. We have demonstrated better results on the TalkSHOW dataset compared to a baseline method using 2D skeleton-based representation for gestures in terms of both quantitative metrics and human preference.

Limitations. Our current method still has several limitations. First, We cannot handle videos with dynamic backgrounds with moving objects, as we only model

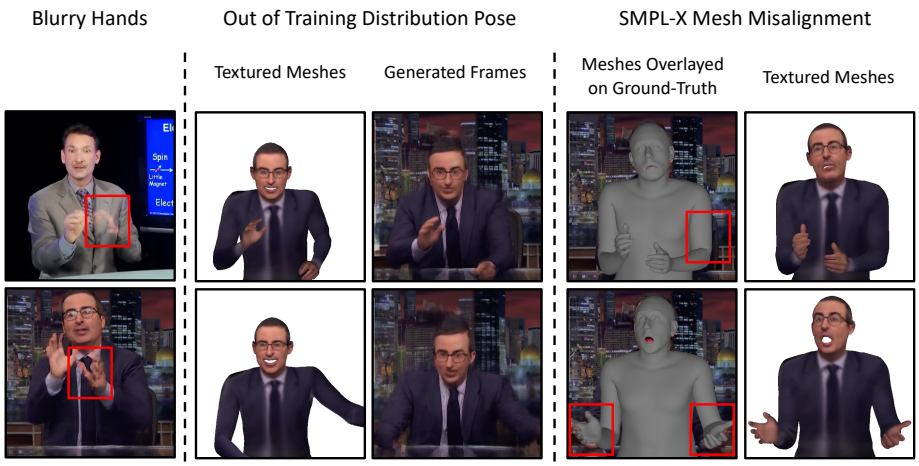


Fig. 6: Limitations. Some of the limitations of our method are that: (left) Our method sometimes synthesizes blurry textures in the arms and hands in the generated frames (highlighted in red). (middle) Our method struggles to generate realistic videos when the mesh pose at inference differs significantly from the speaker's training pose. In this example, the mesh (pose and camera) is taken from the speaker of 'Chemistry,' and the final frames are generated from the model trained on 'Oliver' sequences. (right) Our method does not further optimize the mesh parameters obtained from TalkSHOW training data, leading to misalignment between the ground truth textured mesh and the RGB image (highlighted in red), mainly caused by slight errors in tracking. This also leads to learning blurry textures for the mesh using optimization, especially in the hands region.

the 3D geometry and motion of the foreground humans. How to model the dynamics and geometry of more objects will be an interesting future direction.

Second, our method occasionally struggles to synthesize realistic textures for regions occluded by the arms and hands across most frames. Integrating a better inpainting method could further enhance our approach.

Third, we need to perform per-character texture optimization and generator training, which can be time-consuming. Developing encoder-based methods for both appearance and motion modeling will be a meaningful next step.

Fourth, as shown in Figure 6 (middle), our method also has difficulty generating photorealistic videos in cases where the speaker's pose differs dramatically from that in the training data.

Finally, our method also works best when trained on a relatively large training video sequence of an actor, and sometimes, it struggles to generate proper hand structures, as seen in Figure 6 (right), especially in cases with less training data.

Acknowledgments. We thank Kangle Deng, Yufei Ye, and Shubham Tulsiani for their helpful discussion. The project is partly supported by Ping An Research.

References

1. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. ACM Trans. Graph. (2023). https://doi.org/10.1145/3592097
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. NeurIPS **33** (2020)
3. Boukhayma, A., Bem, R.d., Torr, P.H.: 3d hand shape and pose from images in the wild. In: CVPR (2019)
4. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: ICCV (2023)
5. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: IEEE International Conference on Computer Vision (ICCV) (2019)
6. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models (2023)
7. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: CVPR (2019)
8. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: CVPR (2022)
9. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arxiv:2307.10373 (2023)
10. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2019)
11. Guan, J., Zhang, Z., Zhou, H., HU, T., Wang, K., He, D., Feng, H., Liu, J., Ding, E., Liu, Z., Wang, J.: Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
12. Guo, Y., Chen, K., Liang, S., Liu, Y., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
13. Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.P., Pons-Moll, G., Elgharib, M., Theobalt, C.: Learning speech-driven 3d conversational gestures from video. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. pp. 101–108 (2021)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017)
15. Huang, Z., Tang, F., Zhang, Y., Cun, X., Cao, J., Li, J., Lee, T.Y.: Make-your-anchor: A diffusion-based 2d avatar generation framework. In: CVPR (2024)
16. Huh, M., Zhang, R., Zhu, J.Y., Paris, S., Hertzmann, A.: Transforming and projecting images to class-conditional generative networks. In: ECCV (2020)
17. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
18. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
19. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
20. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: WACV (2022)

21. Liu, X., Wu, Q., Zhou, H., Du, Y., Wu, W., Lin, D., Liu, Z.: Audio-driven co-speech gesture video generation. NeurIPS (2022)
22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
23. Lu, Y., Chai, J., Cao, X.: Live Speech Portraits: Real-time photorealistic talking-head animation. ACM Transactions on Graphics **40**(6) (December 2021). https://doi.org/10.1145/3478513.3480484
24. Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., Yu, X.: Styletalk: One-shot talking head generation with controllable speaking styles. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1896–1904 (2023)
25. Mallya, A., Wang, T.C., Sapra, K., Liu, M.Y.: World-consistent video-to-video synthesis. In: ECCV. Springer (2020)
26. Mensah, D., Kim, N.H., Aittala, M., Laine, S., Lehtinen, J.: A hybrid generator architecture for controllable face synthesis. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–10 (2023)
27. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. NeurIPS **29** (2016)
28. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. NeurIPS **35** (2022)
29. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)
30. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
31. Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: ACM MM (2020)
32. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv:2303.09535 (2023)
33. Qian, S., Tu, Z., Zhi, Y., Liu, W., Gao, S.: Speech drives templates: Co-speech gesture synthesis with learned templates. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (2021)
34. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d (2020)
35. Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
37. Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., Lu, J.: Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In: CVPR (2023)
38. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: NeurIPS (December 2019)
39. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: CVPR (2022)
40. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: CVPR (2018)
41. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. NeurIPS **30** (2017)

42. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NeurIPS (2016)
43. Wang, J., Qian, X., Zhang, M., Tan, R.T., Li, H.: Seeing what you said: Talking face generation guided by a lip reading expert. In: CVPR (2023)
44. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: NeurIPS (2018)
45. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)
46. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: CVPR (2021)
47. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
48. Wu, H., Jia, J., Wang, H., Dou, Y., Duan, C., Deng, Q.: Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In: ACM MM (2021)
49. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565 (2022)
50. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
51. Yang, S., Zhou, Y., Liu, Z., , Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. In: ACM SIGGRAPH Asia Conference Proceedings (2023)
52. Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Cheng, M., Xiao, L.: Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. In: IJCAI (8 2023)
53. Yao, X., Fried, O., Fatahalian, K., Agrawala, M.: Iterative text-based editing of talking-heads using neural retargeting. ACM Transactions on Graphics (TOG) **40**(3), 1–14 (2021)
54. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: CVPR (2023)
55. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: CVPR (2023)
56. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
57. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. arXiv preprint arXiv:2211.12194 (2022)
58. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: CVPR (2022)
59. Zhu, L., Liu, X., Liu, X., Qian, R., Liu, Z., Yu, L.: Taming diffusion models for audio-driven co-speech gesture generation. In: CVPR (2023)
60. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: CVPR (2023)