From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations

Evonne Ng[1, 2]   Javier Romero[1]   Timur Bagautdinov[1]   Shaojie Bai[1]

Trevor Darrell[2]   Angjoo Kanazawa[2]   Alexander Richard[1]

[1]Codec Avatars Lab, Meta, Pittsburgh   [2]University of California, Berkeley

Input: conversational audio   **Intermediate Output:** 3D face, hand, and body motion   **Final output:** Photoreal conversational avatar
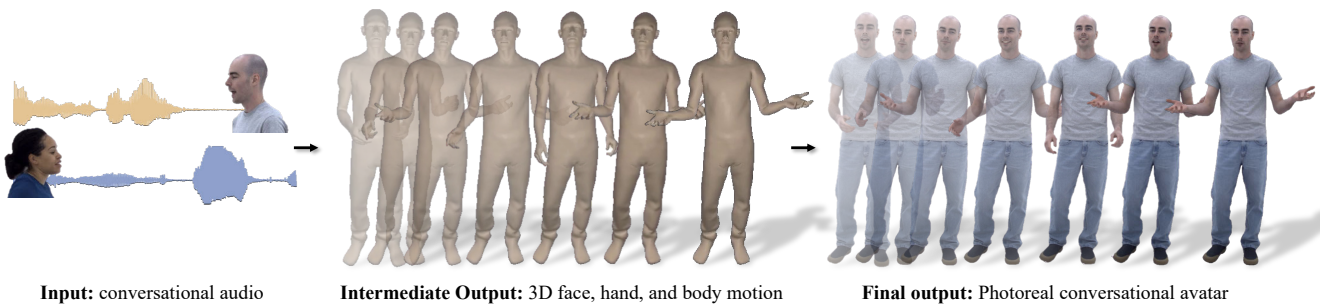
Figure 1. **Synthesizing photoreal conversational avatars.** Given the audio from a dyadic conversation, we generate realistic conversational motion for the face, body, and hands. The motion can then be rendered as a photorealistic video. Please see results video.

Abstract

We present a framework for generating full-bodied photorealistic avatars that gesture according to the conversational dynamics of a dyadic interaction. Given speech audio, we output multiple possibilities of gestural motion for an individual, including face, body, and hands. The key behind our method is in combining the benefits of sample diversity from vector quantization with the high-frequency details obtained through diffusion to generate more dynamic, expressive motion. We visualize the generated motion using highly photorealistic avatars that can express crucial nuances in gestures (e.g. sneers and smirks). To facilitate this line of research, we introduce a first-of-its-kind multi-view conversational dataset that allows for photorealistic reconstruction. Experiments show our model generates appropriate and diverse gestures, outperforming both diffusion- and VQ-only methods. Furthermore, our perceptual evaluation highlights the importance of photorealism (vs. meshes) in accurately assessing subtle motion details in conversational gestures. Code and dataset available on project page.

1. Introduction

Consider talking to your friend in a telepresent world, where they appear as the generic golden mannequin shown in Figure 1 (middle). Despite the mannequin's ability to act out rhythmic strokes of arm motion that seemingly follow your friend's voice, the interaction will inevitably feel robotic and uncanny. This uncanniness stems from the limitations imposed by non-textured meshes which mask subtle nuances like eye gaze or smirking. Photorealistic details can effectively convey these nuances, allowing us to express diverse moods during conversation. For example, a sentence spoken while avoiding eye contact differs significantly from one expressed with sustained gaze. As humans, we are especially perceptive to these micro-expressions and movements, which we use to formulate a higher-order understanding of our conversational partner's intentions, comfort, or understanding [10]. Developing conversational avatars with the level of photorealism that can capture these subtleties is therefore essential for virtual agents to meaningfully interact with humans.

Our ability to perceive these fine-grain motion patterns breaks down as we represent the motion in more abstracted forms. Chaminade *et al*. [8] demonstrates that humans have a more difficult time distinguishing real vs. fake key-framed motions (such as walking) in skeletons than in textured meshes, and even more-so in point-based representations than in skeletons. In faces, McDonnell *et al*. [26] shows that large facial motion anomalies are considerably less discernible on Toon (i.e. plain colored, comic-like) characters,

than on characters with human textures applied. Although abstract representations cannot precisely represent the level of detail needed for humans to interpret subtle conversational cues, the majority of prior works in gesture generation [2, 22, 23, 40] still assess their methods using mesh-based or skeletal representations. In this paper we advocate the importance of developing *photorealistic* conversational avatars which not only allow us to express subtle motion patterns, but also allow us to more accurately evaluate the realism of the synthesized motion.

To this end, we present a method for generating photorealistic avatars, conditioned on the speech audio of a dyadic conversation. Our approach synthesizes diverse high-frequency gestures (e.g. pointing and smirking) and expressive facial movements that are well-synchronized with speech. For the body and hands, we leverage advantages of both an autoregressive VQ-based method and a diffusion model. Our VQ transformer takes conversational audio as input and outputs a sequence of guide poses at a reduced frame rate, allowing us to sample diverse poses (e.g. pointing) while avoiding drift. We then pass both the audio and guide poses into the diffusion model, which infills intricate motion details (e.g. finger wag) at a higher fps. For the face, we use an audio conditioned diffusion model. The predicted face, body, and hand motion are then rendered with a photorealistic avatar. We demonstrate the added guide pose conditioning on the diffusion model allows us to generate more diverse and plausible conversational gestures compared to prior works. In a perceptual study, we further illustrate that evaluators can better distinguish differences between two approaches when motion is visualized with photorealistic avatars than with meshes.

To support our approach in modeling the intricacies of human conversation, we introduce a rich dataset of dyadic interactions captured in a multi-view system. This system allows for highly accurate body/face tracking and photorealistic 3D reconstructions of both participants simultaneously. The non-scripted, long-form conversations cover a wide range of topics and emotions. In contrast to prior full-body datasets that support skeletal [22, 23] or Toon-like visualizations [24], we reconstruct photorealistic renders of each individual in the dataset. Our data also captures the dynamics of inter-personal conversations rather than individual monologues [13, 24, 40]. We will release the dataset and renderer, and hope these will encourage the investigation of gesture generation in a photorealistic manner.

To the best of our knowledge, we are the first to investigate the generation of photorealistic face, body, *and* hand motion for interpersonal conversational gestures. Our VQ- and diffusion-based method synthesizes more realistic and diverse motion compared to prior works. Furthermore, we pose an important question on the validity of evaluating conversational motion using non-textured meshes, as
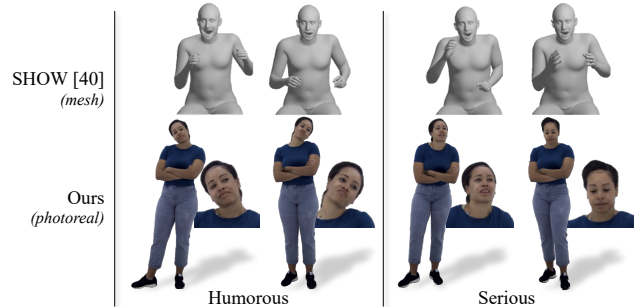

Figure 2. **Importance of photorealism** Top: Mesh annotations from prior work [40]. Bottom: Our photorealistic renderings. For the mesh, differences in laughing (top left) vs. speaking (top right) are difficult to perceive. In contrast, photorealism allows us to capture subtle details such as the smirk (bottom left) vs. grimace (bottom right), which completely changes the perception of her current mood despite similar coarse body poses.

humans may overlook or be less critical of inaccuracies in these representations. Finally, to support this investigation, we introduce a novel dataset of long-form conversations that enable renderings of photorealistic conversational avatars. Code, dataset, and renderers will all be publicly available.

2. Related Work

Interpersonal conversational dynamics. Traditionally, animating conversational avatars have involved constructing rule-based guides from lab captured motion data [6, 7, 14, 18]. These methods are often limited in variety of gestures and rely on simplifying assumptions that do not hold on in-the-wild data. As a result, there has been greater focus on using learning-based methods to predict coarse aspects of a conversation such as turn-taking [1, 22] or a single facial expression to summarize a conversation [19, 31]. While these methods focus on higher-level dynamics, our method focuses on the lower-level complexities of interactions by modeling the full range of facial expressions and body-hand motion. In contrast, Tanke *et al*. [33] predicts the full body pose, but focuses on a different task of motion forecasting, where the goal is to generate plausible future body poses for a triad given their past body motion.

More recently, there have been works on modeling cross-person interaction dynamics by predicting the listener's fine-grain 2D [12] or 3D gestural motion from the speaker's motion and audio [21, 28], text [29], or stylized emotion [45]. However, all these methods generate only the head pose and facial expression of the listener alone. On the other extreme, Lee *et al*. [23] models only the finger motion of the speaker in a dyadic conversation. In contrast, our method is the first to consider the full range of 3D face, body, and hand motion for interpersonal conversation while using a single model to handle both speaking and listening motion.
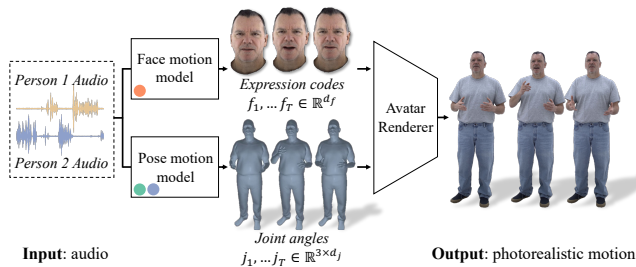
Figure 3. **Method Overview** Our method takes as input conversational audio and generates corresponding face codes and body-hand poses. The output motion is then fed into our trained avatar renderer, which generates a photorealistic video. For details on the face/pose models, please see Figure 4.

Gestural motion generation. Prior works on diffusion have explored audio to dance [36], text to motion [34], or even audio to gestures [2, 3, 41, 44]. In [2, 3], body motion of a speaker is synthesized using a diffusion model conditioned on audio or text respectively. Meanwhile, Yu *et al.* [41] focuses only on the face by using a diffusion-based method with contrastive learning to produce lip sync that is both accurate and can be disentangled from lip-irrelevant facial motion. While these methods model only the body or the face, our approach generates the full *face, body, and hands* of the conversational agent simultaneously.

SHOW [40] addresses this issue by training separate VQ's to produce face, body, and hand motion given a speaker's audio. While our approach similarly focuses on generating the full range of face, body, and hand motion for a conversational agent, our approach significantly differs in that we visualize on photorealistic avatars as opposed to mesh-based renderings. As depicted in Figure 2, their mesh can represent large arm movements that follow a rhythm, but struggles to capture crucial distinctions between a laugh and opening one's mouth to speak (top). In contrast, we are the first to employ photoreal avatars that can express subtle nuances such as a grimace vs. a smirk (bottom). We demonstrate in our analysis (Sec. 5.2) that photorealism greatly affects the evaluation paradigm for conversational agents.

We further differentiate from these prior works [2, 3, 40, 41] in that we model *interpersonal* communication dynamics of a dyadic conversation as opposed to a single speaker in a monadic setting. As a result, our method must model both listener and speaker motion, and generate motion that not only looks realistic with respect to the audio, but also reacts realistically to the other individual in conversation.

Conversational datasets. There is a growing number of large scale datasets for conversational motion [23, 24, 27, 40]. Pose parameters for the face, body and hands of a monologue speaker are released at large scale in [24, 40]. Similarly [23, 28] provide only the body and hand reconstructions. However, all these datasets release only enough

information to reconstruct coarse human meshes or textured avatars through blendshapes that lack photorealism and high-frequency details [24].

Given the popularity of the task of audio-driven lip syncing, there are many datasets with open-sourced pipelines for generating facial motion [9, 20, 32, 35, 39, 41, 43], though these approaches are limited to either 2D video or 3D mesh-based animation. Complementing such work with a focus on the face, Ginosar *et al.* [13] provides a way to render out the body and hands of a monologue speaker. To the best of our knowledge, we are the first to provide a dataset with full simultaneous reconstructions of the *face, body, and hands*, and to consider this in a *dyadic* conversational setting.

3. Photoreal full body motion synthesis

Given raw audio from a conversation between two people, we introduce a model that generates corresponding photorealistic face, body, and hand motion for one of the agents in the dyad. We represent the face as latent expression codes from the recorded multi-view data following [25], and the body pose as joint angles in a kinematic skeleton. As shown in Fig. 3, our system consists of two generative models that produce sequences of expression codes and body poses given audio from the dyadic conversation as input. Expression codes and body pose sequences can then be rendered frame-by-frame using our trained neural avatar renderer [5] which produces the full textured avatar with the face, body, and hands from a given camera view.[1]

Note that the body and face follow highly different dynamics. First, the face is strongly correlated with the input audio, particularly in terms of lip motion, while the body has a weaker correlation with speech. This leads to greater diversity in plausible body gestures for a given speech input. Second, since we represent face and body in two different spaces (learned expression codes vs. joint angles), each of them follow different temporal dynamics. We therefore model the face and body with two separate motion models. This allows the face model to spend its capacity on generating speech-consistent facial details, and the body model to focus on generating diverse yet plausible body motion.

The **face motion model** is a diffusion model conditioned on input audio and lip vertices produced by a pre-trained lip regressor (Fig. 4a). For the **body motion model**, we found that a purely diffusion-based model conditioned only on audio produces less diverse motion that appears temporally uncanny. However, the quality improves when we condition on diverse guide poses. We therefore split the body motion model into two parts: First, an autoregressive audio-conditioned transformer predicts coarse guide poses at 1fps

---

[1]Face expression codes, tracked joint angles, and the pre-trained full-body renderer are released as part of the dataset.
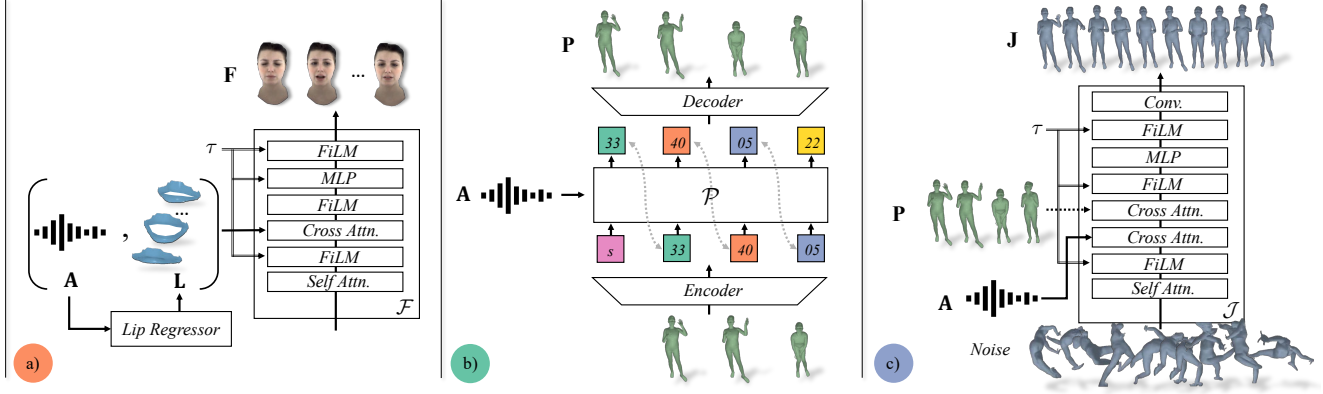
Figure 4. **Motion generation** (a) Given conversational audio **A**, we generate facial motion **F** using a diffusion network conditioned on both audio and the output of a lip regression network **L**, which predicts synced lip geometry from speech audio. (b) For the body-hand poses, we first autoregressively generate guide poses **P** at a low fps using a VQ-Transformer. (c) The pose diffusion model then uses these guide poses and audio to produce a high-frequency motion sequence **J**.

(Fig. 4b), which are then consumed by the diffusion model to in-fill fine grain and high-frequency motion (Fig. 4c). We describe the model components in detail below.

Notation. We denote the audio of the agent as $\mathbf{a}_{self}$ and the audio of the conversation partner as $\mathbf{a}_{other}$. For both audio streams, we extract Wav2Vec [4] features such that the audio input is $\mathbf{A} = (\mathbf{a}_{self}, \mathbf{a}_{other}) \in \mathbb{R}^{2 \times d_a \times T}$, with d_a denoting the feature dimension of Wav2Vec features.

We denote a sequence of T face expression codes as $\mathbf{F} = (f_1, \ldots, f_T)$, where each $f_t \in \mathbb{R}^{256}$ represents a face expression for frame t. A body motion sequence of T frames is represented by $\mathbf{J} = (j_1, \ldots, j_T)$, where $j_t \in \mathbb{R}^{d_j \times 3}$ is a vector containing three rotation angles for each of the d_j body joints that define a pose at frame t. We follow the forward kinematic representation [5], where the body and hand pose of a given person can be constructed from the relative rotations of each joint with respect to its parent joint.

3.1. Face Motion Diffusion Model

To generate facial motion from audio input, we construct an audio-conditioned diffusion model. We follow the DDPM [15] definition of diffusion. The forward noising process is defined as:

$$q(\mathbf{F}^{(\tau)}|\mathbf{F}^{(\tau-1)}) \sim \mathcal{N}\big(\sqrt{\alpha_\tau}\mathbf{F}^{(\tau-1)}, (1-\alpha_\tau)\mathbf{I}\big), \quad (1)$$

where $\mathbf{F}^{(0)}$ approximates the clean (noise-free) sequence of face expression codes $\mathbf{F}$, $\tau \in [1, \ldots, \dot{T}]$ denotes the forward diffusion step, and $\alpha_\tau \in (0, 1)$ follows a monotonically decreasing noise schedule such that as τ approaches $\dot{T}$, we can sample $\mathbf{F}^{(\dot{T})} \sim \mathcal{N}(0, \mathbf{I})$.

To reverse the noising process, we follow [15, 30] and define a model to denoise $\mathbf{F}^{(0)}$ from the noisy $\mathbf{F}^{(\tau)}$. The next step $\mathbf{F}^{(\tau-1)}$ of the reverse process can then be obtained by applying the forward process to the predicted $\mathbf{F}^{(0)}$. We

predict $\mathbf{F}^{(0)}$ with a neural network $\mathcal{F}$:

$$\mathbf{F}^{(0)} \approx \mathcal{F}(\mathbf{F}^{(\tau)}; \tau, \mathbf{A}, \mathbf{L}), \quad (2)$$

where **A** are the input audio features and $\mathbf{L} = (l_1, \ldots, l_T)$ is the output of a pre-trained audio-to-lip regressor following [9], but limited to lip vertices instead of full face meshes. We train the lip-regressor on 30h of in-house 3D mesh data. Each $l_t \in \mathbb{R}^{d_l \times 3}$ is a predicted set of d_l lip vertices at frame t given audio **A**. Tab. 2 shows, conditioning on both the lip regressor output and audio significantly improves lip sync quality over conditioning on audio alone.

The diffusion model is trained with the simplified ELBO objective [15],

$$\mathcal{L}_{simple} = \mathbb{E}_{\tau, \mathbf{F}}\big[\mathbf{F} - \mathcal{F}(\mathbf{F}^{(\tau)}; \tau, \mathbf{A}, \mathbf{L})\big]. \quad (3)$$

We train our model for classifier-free guidance [16] by randomly replacing either conditioning with $\mathbf{A} = \emptyset$ and $\mathbf{L} = \emptyset$ during training with low probabilities. To incorporate the audio and lip vertex information, we use a cross attention layer. Timestep information is incorporated with a feature-wise linear modulation (FiLM) layer, see Fig. 4a.

3.2. Body Motion Model

To generate body motion, we extend the conditional diffusion model by introducing guide poses sampled at 1fps as additional conditioning. This allows us to model more expressive motion. Similar to the face model that did not generate accurate lip motion when conditioned on audio alone, we found that the body model generates less plausible motion with limited diversity when conditioned on audio only.

More formally, to generate a full body motion sequence at 30fps, we train the body diffusion model with guide poses $\mathbf{P} = \{j_{k \cdot 30} | 1 \le k \le T/30\}$ taken at 1fps. These guide poses are obtained by subsampling the original 30 fps
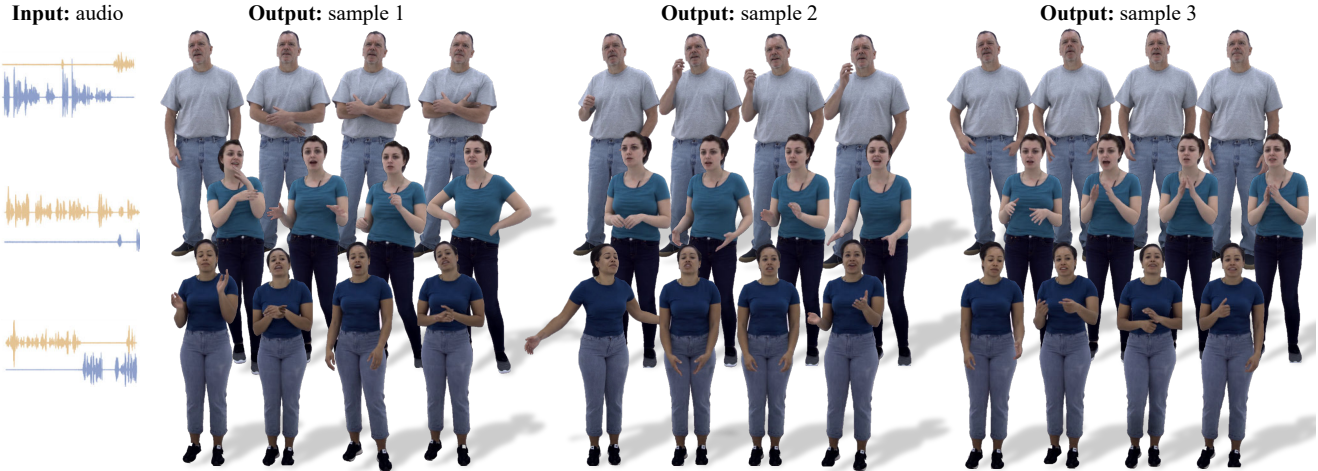
Figure 5. **Diversity of guide pose rollouts** Given the input audio for the conversation (predicted person's audio in gold), the transformer $\mathcal{P}$ generates diverse samples of guide pose sequences with variations in listening reactions (top), speech gestures (middle), and interjections (bottom). Sampling from a rich codebook of learned poses, $\mathcal{P}$ can produce "extreme" poses e.g. pointing, itching, clapping, etc. with high diversity across different samples. These diverse poses are then used to condition the body diffusion model $\mathcal{J}$.

ground truth body pose sequence **J**. The body motion diffusion model $\mathcal{J}$ is then the same network as the face motion diffusion model $\mathcal{F}$, but is conditioned on the subsampled guide poses, i.e. $\mathbf{J}^{(0)} \approx \mathcal{J}(\mathbf{J}^{(\tau)}; \tau, \mathbf{A}, \mathbf{P})$. The guide poses are incorporated using an additional cross attention layer (see Fig. 4c). At inference time, however, ground truth guide poses are not available and need to be generated.

Guide pose generation. To generate guide-poses at inference time, we train an autoregressive transformer to output coarse keyframes at 1fps that adhere to the conversational dynamics. As autoregressive transformers typically operate on discrete tokens [11, 28, 40], we first quantize the 1 fps guide pose sequence using a residual VQ-VAE [38]. Residual VQ-VAEs are similar to vanilla VQ-VAEs [37], but they recursively quantize the residuals of the previous quantization step instead of stopping after a single quantization step. This leads to higher reconstruction quality [38, 42].

Let $\mathbf{Z} = (z_1, \ldots, z_K)$ be the resulting quantized embedding of the K-length guide pose sequence $\mathbf{P}$, where $z_k \in \{1, \ldots, C\}^N$, C is codebook size, and N is the number of residual quantization steps. We flatten this $K \times N$-dimensional quantized embedding $\mathbf{Z}$ to obtain $\hat{\mathbf{Z}} = (\hat{z}_1, \ldots, \hat{z}_{K \cdot N})$. We predict $\hat{\mathbf{Z}}$ with an audio-conditioned transformer $\mathcal{P}$, which outputs a categorical distribution over the next token given prior predictions and audio,

$$p(\hat{z}_k | \hat{z}_{1:k-1}, \mathbf{A}) = \mathcal{P}(\hat{z}_{1:k-1}; \mathbf{A}). \quad (4)$$

We train the transformer using a simple cross entropy loss on the task of next-token prediction with teacher forcing:

$$\mathcal{L}_{\mathcal{P}} = - \sum_{k \in \{1, \ldots, K \cdot N\}} \log \Pr\left[\mathcal{P}(z_{1:k-1}, \mathbf{A}) = z_k\right]. \quad (5)$$

At test time, we use nucleus sampling [17] to predict the sequence of motion tokens. We can easily control the level of variability seen across samples by increasing or decreasing the cumulative probability.

The guide-pose transformer is illustrated in Fig. 4b. For further architecture details on the residual VQ-VAE and the transformer architecture refer to supplemental.

3.3. Photorealistic Avatar Rendering

Given both the generated facial expression sequence **F** and the generated body pose sequence **J**, the full photorealistic avatar can be rendered as illustrated in Fig. 3. Following [5], we use a learning-based method to build our drivable avatars. The model takes as input one frame of facial expression f_t, one frame of body pose j_t, and a view direction. It then outputs a registered geometry and view-dependent texture, which is used to synthesize images via rasterization. The model is a conditional variational autoencoder (cVAE) consisting of an encoder and decoder, both parameterized as convolutional neural networks. The cVAE is trained end-to-end in a supervised manner to reconstruct images of a subject captured in a multi-view capture setup. We train a personalized avatar renderer for each subject in our dataset. For details, please refer to [5].

4. Photorealistic conversational dataset

While there are a plethora of datasets on dyadic interactions [23, 28], all such datasets are only limited to upper body or facial motion. Most related is Joo et al. [22], which introduces a small-scale dataset of triadic interactions as a subset of the Panoptic Studio dataset. The data includes 3D skeletal reconstructions of face, body and hands as well as

Figure 6. **Results** Our method produces gestural motion that is synchronous with the conversational audio. During periods where the person is listening (top), our model correctly produces still motion, seemingly as if the avatar is paying attention. In contrast, during periods of talking (bottom), the model produces diverse gestures that move synchronously with the audio.

audio and multi-view raw video footage. The limited ($\approx$ 3 hours) and specific data (only focused on haggling), makes it difficult to learn diverse motion distributions.

Inspired by this work, we introduce a medium-scale dataset capturing dyadic conversations between pairs of individuals totaling to 8 hours of video data from 4 participants, each engaging in 2 hours of paired conversational data. To ensure diversity of expressions and gestures, we prompt the actors with a diversity of situations such as selling, interviews, difficult scenarios, and everyday discourse.

Most notably, to the best of our knowledge, we are the first to provide a dataset accompanied with fully photorealistic renderings of the conversational agents. Rather than generating and evaluating motion via 3D meshes, our multiview dataset allows us to reconstruct the full face, body, and hands in a photorealistic manner. Visualizing via these renderings allow us to be more perceptive to fine-grain details in motion that are often missed when rendered via coarse 3D meshes. Our evaluations confirm the importance of evaluating gestural motion using photo-real avatars.

To create the photorealistic renderings, we captured both individuals simultaneously in multi-view capture domes. One person stood in a full-body dome while the other sat in a head-only dome. During the conversations, both viewed screens of the other person in real-time. We can then reconstruct high fidelity renderings of the face only for one individual [25], and the face, body, and hands for the other [5]. To train our method, we use the ground truth from the fullbody capture to supervise our method. We will publicly release audio, video, precomputed joint angles, face expression codes, and trained personalized avatar renderers.

5. Experiments

We evaluate the ability of our model to effectively generate realistic conversational motion. We quantitatively measure the realism and diversity of our results against tracked ground truth data ($\mathbf{F}, \mathbf{J}$). We also perform a perceptual eval-

uation to corroborate the quantitative results and to measure appropriateness of our generated gestures in the given conversational setting. Our results demonstrate evaluators are more perceptive to subtle gestures when rendered on photorealistic avatars than on 3D meshes.

5.1. Experimental Setup

Evaluation Metrics. Following a combination of prior works [2, 3, 28], we use a composition of metrics to measure the realism and diversity of generated motion.

- FD_g: "geometric" realism measured by distribution distance between generated and ground truth *static* poses. We directly calculated the Frechet distance (FD) in the expression $\mathbb{R}^{d_f}$ and pose space $\mathbb{R}^{d_j \times 3}$.
- FD_k: "kinetic" motion realism. Similar to above but distributions calculated on the velocities of motion sequences $\delta \mathbf{P}$. Computed in expression $\mathbb{R}^{T \times d_f}$ and pose space $\mathbb{R}^{T \times d_j \times 3}$.
- Div_g: "geometric" pose diversity. We randomly sample 30 expression, pose pairs within a motion sequence and compute average L2 distances between pairs to measure diversity of *static* expressions/poses in the set.
- Div_k: Temporal variance across a sequence of expressions/poses. Measures amount of motion in a sequence.
- Div_{sample}: Diversity across different samples. We group samples generated by the same audio and calculate variance across the samples.

Together, these metrics measure both the realism and diversity of the generated gestures in conversation.

Baselines and ablations. We compare to the following:
- **Random**: Random motion sequences from the train set.
- **KNN**: A segment-search method commonly used for synthesis. Given input audio, we find its nearest neighbor from the training set and use its corresponding motion segment as the prediction. We use audio features from Wav2Vec [4] to encode the audio.

	$\text{FD}_g \downarrow$	$\text{FD}_k \downarrow$	$\text{Div}_g \uparrow$	$\text{Div}_k \uparrow$	$\text{Div}_{sample} \uparrow$
GT			3.09	2.50	
Random	$9.37_{1.4}$	$1.44_{0.04}$	$3.10_{0.09}$	$2.49_{0.4}$	$3.97_{0.8}$
KNN	$8.44_{1.6}$	$0.62_{0.09}$	$2.13_{0.05}$	$1.21_{0.3}$	$1.96_{0.3}$
SHOW [40]	$4.97_{0.7}$	$2.60_{0.10}$	$2.10_{0.09}$	$0.77_{0.1}$	$2.82_{0.2}$
LDA [2]	$5.08_{0.2}$	$1.04_{0.07}$	$2.45_{0.06}$	$1.88_{0.3}$	$2.68_{0.4}$
Ours Uncond	$8.45_{1.3}$	$1.53_{0.08}$	$2.74_{0.07}$	$2.06_{0.4}$	$2.94_{0.3}$
Ours w/o **P**	$5.08_{0.4}$	$1.13_{0.09}$	$2.47_{0.06}$	$1.67_{0.3}$	$2.06_{0.4}$
Ours w/o **A**	$3.94_{0.1}$	$0.98_{0.10}$	$2.69_{0.08}$	$2.16_{0.4}$	$2.71_{0.3}$
Ours	$2.94_{0.2}$	$0.96_{0.07}$	$2.98_{0.07}$	$2.36_{0.4}$	$3.58_{0.5}$

Table 1. **Baselines and ablations** vs. ground truth poses (GT).
$\downarrow$ indicates lower is better. We average across all subjects in the dataset. We sample 5 sequences for Div_{sample} and average across all samples for each metric. Standard deviation as subscript (μ_σ).

	Horizontal L2 Error $\downarrow$	Vertical L2 Error $\downarrow$	Mesh L2 $\downarrow$
SHOW [40]	2.76	2.15	2.25
Ours w/o **L**	2.62	2.43	2.24
Ours	2.29	1.89	1.76

Table 2. **Lip reconstructions** The vertical (horizontal) distance is the distance between top and bottom (left and right) keypoints along the y (x) axis. The errors shown are L2 differences between ground truth and generated distances. Mesh L2 is the error in generated vs. GT mesh vertices on the lip region. Errors in mm^2.

- **SHOW** [40]: VQ-VAE based method that uses a transformer to autoregressively output motion conditioned on the audio of a speaker. They have separate models for face, body, and hands. Given [40] is trained on monologues, we retrain their model for our domain.
- **LDA** [2]: Audio to motion diffusion model trained in a monologue setting. We re-train to adapt to our domain.
- **Ours Uncond**: (ablation) unconditional motion generation without audio or guide pose conditioning.
- **Ours w/o P**: (ablation) audio conditioned motion diffusion without guide pose conditioning. Similar to LDA [2].
- **Ours w/o A**: (ablation) guide pose conditioned motion diffusion model but without audio conditioning. Similar to diffusion infilling approaches.

5.2. Results

Through quantitative evaluations, we show that our proposed method outputs realistic motion more diverse than competing baselines. In our Mechanical Turk A/B evaluations, we demonstrate our method generates compelling and plausible gestures, consistently outperforming our strongest baseline. Additionally, the A/B tests highlight that photorealism effectively captures subtle nuances in gestures that are challenging to discern from 3D meshes. Yet these details significantly effect the evaluation of conversational motion.
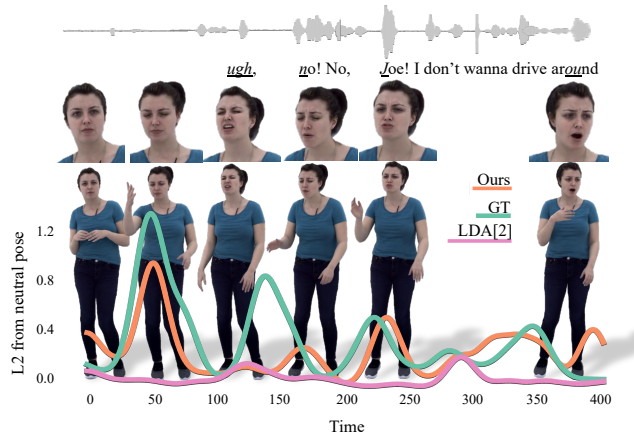

Figure 7. **Motion correlation with audio** Given audio (top), we plot the L2 distance of each pose to the mean neutral pose across 400 frames. Ours (rendered avatars, orange line) closely matches the peaks corresponding to large motion also seen in ground truth (e.g. a flick of the hand preempting the "ugh"). LDA [2] (pink) misses these peaky motions. Also note how our method generates highly expressive facial motion matching the speech.

Quantitative Results. Table 1 shows that compared to prior works, our method achieves the lowest FD scores while generating motion with highest diversity. While **Random** has good diversity that matches that of **GT**, the random segments do not appropriately match the corresponding conversational dynamics, resulting in higher FD_g. A slight improvement to **Random** is **KNN**, conventionally used for motion synthesis. While **KNN** performs better in terms of realism, matching the "kinetic" distribution of the ground truth sequences better than **Ours**, the diversity across and within samples is significantly lower, also indicated by the higher "geometric" FD. In Fig. 5, we demonstrate the diversity of guide poses our method generates. Sampling via the VQ-based transformer $\mathcal{P}$ allows us to produce significantly different styles of poses conditioned on the same audio input. The diffusion model then learns to produce dynamic motion (Fig. 6), where the motion faithfully follows the conversational audio.

Our method outperforms both a VQ-only approach **SHOW** [40] and a diffusion-only approach **LDA** [2], achieving better realism and diversity across samples. Within sequences, our method generates more motion, resulting in a higher Div_k. Fig. 7 highlights this, demonstrating **LDA** [2] produces dampened motion with less variation. In contrast, our method synthesizes variations in motion that closely match ground-truth.

Our ablations justify our design decisions. Applying our method without any conditioning (**Ours Uncond**) performs notably worse, with a realism and variance similar to that of **Random**. This suggests that while the motion generated does not match the given conversation sequence, it is
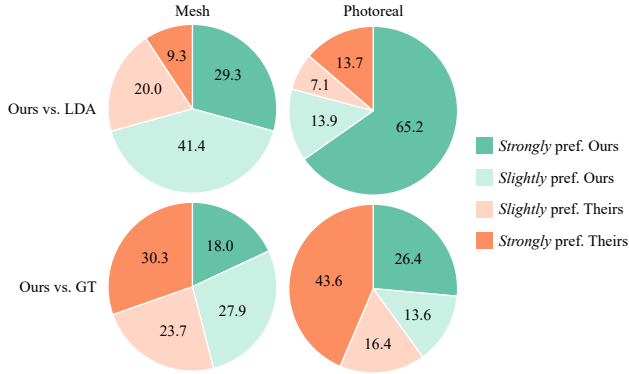
Figure 8. **Perceptual evaluation** on Ours vs. ground truth or Ours vs. our strongest baseline LDA [2]. We compare using mesh vs. photorealistic visualizations. Ours outperforms LDA [2] in both mesh and photoreal settings (top). Further, we note people are able to distinguish GT more often in the photoreal setting than with meshes (bottom). The results suggest that evaluating with photorealistic avatars leads to more accurate evaluations.

similar to real motion in the dataset. Audio only conditioning (**Ours w/o P**) improves over unconditional generation and performs similarly to **LDA** [2], an audio to motion diffusion-based method. The lower diversity in both the static poses and across a temporal sequence results in higher FD scores. When adding only the guide pose conditioning (**Ours w/o A**), both the diversities and FD scores improve significantly. This suggests that the coarse-to-fine paradigm, introduced through the predicted guide poses, helps to add diversity to the diffusion results. It also suggests that the coarse guide poses produced by the transformer $\mathcal{P}$ follow a trajectory that faithfully matches the dynamics of the conversational audio. The FD scores and diversities further improve when adding both the audio and guide pose conditioning in the body motion model $\mathcal{J}$.

Furthermore, we analyze the accuracy of our method in generating lip motion. In Table 2, we calculate the vertical and horizontal distances between two pairs of keypoints representing the top/bottom and left/right corners of the mouth, respectively. The vertical distance measures errors in mouth opening while the horizontal distance measures mouth expressions, *e.g.* a smile shifts the positions of the left/right mouth corner and increases the horizontal distance. We compare these distances against ground truth and compute the L2 error. Our approach (**Ours** in Table 2) substantially outperforms an ablation without the pretrained lip regressor (**Ours w/o L** in Table 2) and the baseline **SHOW** [40]. Qualitatively, the pretraining of the lip regressor not only improves lip syncing, but also prevents the mouth from randomly opening and closing while not talking. This results in better overall lip reconstructions, with lower errors on the face mesh vertices (*Mesh L2*).

Perceptual Evaluation. Given the challenge of quantifying the coherence of gestures in conversation, we primarily evaluate this aspect through a perceptual evaluation. We conducted two variations of A/B tests on Amazon Mechanical Turk. In the first, evaluators viewed motion rendered on a generic non-textured mesh. In the second, they viewed videos of the motion on photorealistic avatars.

In both cases, evaluators watched a series of video pairs. In each pair, one video was from our model and the other was from either our strongest baseline LDA [2] or ground truth. Evaluators were then asked to identify the motion that looked more plausible given the conversational audio. They were also asked to indicate how confident they were in their answer by selecting "slightly prefer" vs. "strongly prefer".

As shown in Fig. 8, ours significantly outperforms against our strongest baseline LDA [40], with about 70% of evaluators preferring our method in both the mesh and photoreal settings. Interestingly, evaluators shifted from *slightly* to *strongly* preferring ours when visualized in a photorealistic manner (top). This trend is also reflected in ours vs. ground truth (bottom). While ours performs competitively vs. ground truth in mesh renderings, it lags in the photoreal domain with 43% of evaluators *strongly* preferring ground truth over ours. Since meshes often obscure subtle motion details, it is difficult to accurately evaluate the nuances in gestures leading evaluators to be more forgiving of "incorrect" motions. Our results suggest that photorealism is essential to accurately evaluating conversational motion.

6. Conclusion

In this work, we explored generating conversational gestures conditioned on audio for fully embodied photorealistic avatars. To this end, we combine the benefits of vector quantization with diffusion to generate more expressive and diverse motion. We train on a novel multi-view, long-form conversational dataset that allows for photorealistic reconstructions. Our method produces diverse face, body, and hand motion that accurately matches the conversational dynamics. The results also underscore the significance of photorealism in evaluating fine-grain conversational motion.

Limitations and ethical considerations. While our model produces realistic motion, it operates on short-range audio. It thus fails to generate gestures requiring long-range language understanding (*e.g.* counting), which we leave for future work. Further, our work is limited to photorealistic generation of four subjects in our dataset. This limitation addresses ethical concerns since only consenting participants can be rendered, as opposed to arbitrary non-consenting humans. In releasing a dataset with full participant consent, we hope to provide researchers the opportunity to explore photorealistic motion synthesis in an ethical setting.

References

[1] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International Conference on Multimodal Interaction*, pages 74–84, 2019. 2

[2] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 2, 3, 6, 7, 8

[3] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *arXiv preprint arXiv:2303.14613*, 2023. 3, 6

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 4, 6

[5] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Trans. Graph.*, 40(4), 2021. 3, 4, 5, 6

[6] Dan Bohus and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, New York, NY, USA, 2010. Association for Computing Machinery. 2

[7] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '94, New York, NY, USA, 1994. Association for Computing Machinery. 2

[8] Thierry Chaminade, Jessica Hodgins, and Mitsuo Kawato. Anthropomorphism influences perception of computer-animated characters' actions. *Social cognitive and affective neuroscience*, 2(3):206–216, 2007. 1

[9] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 3, 4

[10] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969. 1

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 5

[12] Will Feng, Anitha Kannan, Georgia Gkioxari, and Larry Zitnick. Learn2smile: Learning non-verbal interaction through observation. *IROS*, 2017. 2

[13] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 2, 3

[14] Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J van der Werf, and Louis-Philippe Morency. Virtual rapport. In *International Workshop on Intelligent Virtual Agents*, pages 14–27. Springer, 2006. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 4

[17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 5

[18] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Virtual rapport 2.0. In *International workshop on intelligent virtual agents*, pages 68–79. Springer, 2011. 2

[19] Yuchi Huang and Saad M Khan. Dyadgan: Generating facial expressions in dyadic interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–18, 2017. 2

[20] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021. 3

[21] Patrik Jonell, Taras Kucherenko, Erik Ekstedt, and Jonas Beskow. Learning non-verbal behavior for a social robot from youtube videos. In *ICDL-EpiRob Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions, Oslo, Norway, August 19, 2019*, 2019. 2

[22] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10873–10883, 2019. 2, 5

[23] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 763–772, 2019. 2, 3, 5

[24] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *arXiv preprint arXiv:2203.05297*, 2022. 2, 3

[25] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Trans. on Graphics*, 37(4), 2018. 3, 6

[26] Rachel McDonnell, Martin Breidt, and Heinrich H Bülthoff. Render me real? investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)*, 31(4):1–11, 2012. 1

[27] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11865–11874, 2021. 3

[28] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022. 2, 3, 5, 6

[29] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[30] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171, 2021. 4

[31] Behnaz Nojavanasghari, Yuchi Huang, and Saad Khan. Interactive generative adversarial networks for facial expression generation in dyadic interactions. *arXiv preprint arXiv:1801.09092*, 2018. 2

[32] Alexander Richard, Michael Zollhoefer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[33] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9601–9611, 2023. 2

[34] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[35] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 716–731. Springer, 2020. 3

[36] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 3

[37] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 5

[38] A Vasuki and PT Vanathi. A review of vector quantization techniques. *IEEE Potentials*, 25(4):39–47, 2006. 5

[39] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128:1398–1413, 2020. 3

[40] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 2, 3, 5, 7, 8

[41] Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. Talking head generation with probabilistic audio-to-visual diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7645–7655, 2023. 3

[42] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. 5

[43] Jiangning Zhang, Liang Liu, Zhucun Xue, and Yong Liu. Apb2face: Audio-guided face reenactment with auxiliary pose and blink signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4402–4406. IEEE, 2020. 3

[44] Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20807–20817, October 2023. 3

[45] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: A benchmark dataset and baseline. In *ECCV*, 2022. 2