Learning to Listen: Modeling Non-Deterministic Dyadic Facial Motion

Evonne Ng[1]   Hanbyul Joo[2]   Liwen Hu[3]   Hao Li[3]   Trevor Darrell[1]

Angjoo Kanazawa[1]   Shiry Ginosar[1]

[1]UC Berkeley   [2]Seoul National University   [3]Pinscreen

Input: Speaker   **Output: Listener**

Speaker Motion & Audio   Generated Samples   Visualization
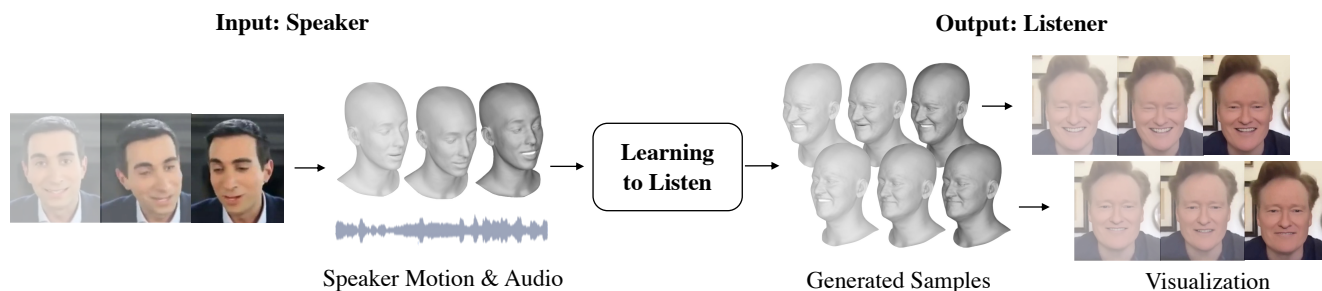
Learning
to Listen

Figure 1. **Synthesizing listeners.** Given a speaker video, we extract the audio and motion of the speaker. From these multimodal speaker inputs, our method synthesizes multiple realistic listener 3D motion sequences (top and bottom) in an autoregessive fashion. The output of our approach can be optionally rendered as photorealistic video.

Abstract

We present a framework for modeling interactional communication in dyadic conversations: given multimodal inputs of a speaker, we autoregressively output multiple possibilities of corresponding listener motion. We combine the motion and speech audio of the speaker using a motion-audio cross attention transformer. Furthermore, we enable non-deterministic prediction by learning a discrete latent representation of realistic listener motion with a novel motion-encoding VQ-VAE. Our method organically captures the multimodal and non-deterministic nature of nonverbal dyadic interactions. Moreover, it produces realistic 3D listener facial motion synchronous with the speaker (see video). We demonstrate that our method outperforms baselines qualitatively and quantitatively via a rich suite of experiments. To facilitate this line of research, we introduce a novel and large in-the-wild dataset of dyadic conversations. Code, data, and videos available at https://evonneng.github.io/learning2listen/.

1. Introduction

"Thus the body of the speaker dances in time with his speech. Further, the body of the listener dances in rhythm with that of the speaker!"

— CONDON AND OGSTON, *1966*

When we speak, it is rarely in a void — rather, there is often a listener at the other end of the conversation. As a speaker, we are acutely aware of what the listener is doing. A slight off-sync motion or a diverted look may throw us off, suggesting the listener is bored or otherwise preoccupied, leaving us feeling misunderstood [36]. Indeed, successful conversations rely on a coordinated dance between the speaker and the listener in which the two signal to each other that they are communicating with one another and not with anyone else [36]. This chameleon effect [12] of nonverbal mimicry during conversation results in smoother interactions, increases the liking between interaction partners, establishes rapport [38], and may even predict the long term outcome of psychotherapy [49]. Interestingly, nonverbal feedback from a listener, such as head movement is more central to keeping a conversation flowing than content-based replies [11]. In this work, we propose a computational framework that can similarly provide nonverbal feedback in response to a speaker in a contextual and timely manner. Such an ability

is critical for virtual agents to meaningfully interact with humans, for whom nonverbal communication is central from infancy [54].

Modeling nonverbal feedback during dyadic interaction is a difficult problem, as listener responses are nondeterministic in nature. Moreover, speakers are inherently multimodal, as they communicate both verbally via speech, and nonverbally via face and body motion. Capturing interaction in its natural setting requires addressing both challenges. The task of modeling human conversations has a long history. However, unlike traditional rule-based methods [5, 10, 22, 29] or methods that rely on modeling hand-defined simple motion characteristics such as smiles [51] or head nods [22, 29], we wish to model the true complexity of the interaction. This is hard to achieve and generalize using conventional database methods that generate motion via a lookup into a database of ground truth motion [35, 52, 58]. We, therefore, learn to model these dyadic conversational dynamics implicitly in a data-driven way by directly observing human conversations in in-the-wild videos.

Given a video of a speaker, we extract their speech audio, and facial motion (Figure 1(left)). We combine information from both modalities using a motion-audio cross-attention transformer. From this multimodal speaker input, we learn to autoregressively synthesize multiple modes of motion representing different possible responses of a listener who moves synchronously with the speaker (Figure 1(right)).

Modeling the nondeterminism in listener responses is a key element in capturing conversational dynamics. Previous attempts to tackle this problem applied various techniques but fell short of achieving realistic outputs [33]. We propose to learn a realistic manifold of listener motion by quantizing the space of listener motion with a novel sequence-encoding VQ-VAE [56], which efficiently captures a wide range of motion in a discrete format that is well-suited for learning. To the best of our knowledge, we are the first to extend VQ-VAE models to the domain of motion synthesis. The learned discrete codebook of listener motion allows us to predict a multinomial distribution of future motion. From this distribution we can sample a wide range of possible modes of motion representing different perceptually-plausible listeners, capturing their inherent non-deterministic nature. Furthermore, we demonstrate our learned discrete latent codes can stay on the manifold of realistic motion ensuring no motion drift occurs even in long-horizon predictions. Meanwhile, the autoregressive nature of our method allows us to consider speaker sequences of any length.

To support our data-driven approach to modeling human conversation, data is needed in the form of video-taped dyadic interactions where both parties are ideally filmed from a head-on frontal view. This kind of data is hard to come by. While the first investigation of interactional synchrony in conversation dates back to Condon and

Ogston in 1966 [15], current studies still mostly rely on in-lab footage [13, 19, 24, 29] or small-scale motion-capture datasets [7, 33]. Notable exceptions are [17, 44], yet the footage has not been made publicly available. We collect a large-scale source of data in the form of split-screen recorded online interviews where the speaker and listener are captured in frontal view. Our dataset, which consists of 72 hours of in-the-wild conversations, enables the investigation of dyadic communication using the latest machine learning methods.

We evaluate the synthesized listener motion compared to ground truth as well as baseline methods and ablations via an extensive quantitative study. We employ a wide array of metrics to test the realism and diversity of the synthesized motion, and the synchronization of the listener's motion with that of the speaker. While measuring realism and diversity centers on the generated motion of the listener in isolation, synchrony captures aspects of the dyad as a whole. We further corroborate our qualitative findings by inviting human observers to evaluate our results. While we assess our method using the raw 3D mesh output, we additionally illustrate our results by translating the 3D output to pixels for viewing purposes only, as synthesized video provides a richer perceptual context. Under both quantitative and qualitative measures, our method significantly outperforms all baselines. Our synthesized listeners were deemed plausible by human observers when compared to ground-truth motion. This highlights our method's ability to produce realistic-looking motion that is synchronous with a given speaker.

Our main contribution is in our learning-based approach towards understanding human interactional communication in conversation. We combine multimodal speaker inputs via motion-audio cross-attention. We extend vector quantization to the domain of motion synthesis and learn a quantized space of motion in which we autoregressively predict multiple modes of perceptually realistic listener motion. To support future endeavors in this direction we *publicly release* a novel dataset of 72 hours of in-the-wild dyadic conversational videos with detailed 3D annotations capturing subtleties in expression and fine-grain head motion.

2. Related Work

We discuss related works concerned with conversational agents and motion synthesis. For a review of interactional motion in human communication, see supplementary.

Interactional Motion in Conversational Agents. Prior works on conversational avatars manually incorporated different aspects of interactional motion [5,10,22,29,53]. These approaches designed rule-based methods to generate agents that can interact via appropriate facial gestures [22, 29, 53], speech [10], or a combination of modalities [5]. All these methods use lab-recorded motion capture sequences. These either limit the variety of captured gestures, or rely on simplifying assumptions for motion generation which do not

hold for in-the-wild data.

Prior data-driven methods predict the 2D motion of one person in a conversation as a function of the other's motion [17, 44]. These require a pre-defined dictionary achieved by clustering motion frequencies or 2D facial keypoints from the training set. Rather, we reason in 3D and learn a discretized latent space that captures the manifold of facial motion. Other methods using 3D investigate interactional dynamics while focusing on full 3D body motion and turn taking [2, 34]. Others tackle the problem of facial gestures in conversation by simplifying the task to predicting head nods [2], estimating head pose [23], or generating a single image of a facial expression that summarizes the entire speaker sequence [30, 44]. In contrast, our method captures the natural complexity of interactions by considering the full range of facial expressions and head rotations.

Recent methods began generating 3D facial motion with additional inputs from the listener such as text [14] or speech [32, 33]. Most similar to our approach is that of Jonel *et al.* [33], who propose a Glow-based method [25, 37]. However, their method takes as input the full temporal context of listener audio and is reported to perform better without any audio input. In contrast, our method does not use any listener audio as additional input. Additionally, we quantitatively demonstrate that *each of the input modalities is essential* to its performance.

Conditional Motion Synthesis. Gestural motion synthesis has previously relied on convolutional auto-encoders to learn a representation of human motion [17, 20, 33, 34, 43]. Some methods incorporated an adversarial loss [20, 43] or experimented with flow models [33] and other sampling-based methods [17] to generate more diverse and realistic motion. Recent works demonstrated the success of using transformers in generating diverse motion with long-range dependencies [9, 39, 40, 48]. These generate possible motion segments conditioned on action [48], 3D human motion trajectory in a scene conditioned on a goal [9], or dance motion from audio [39, 40]. Similarly, we employ a transformer-based predictor for conditional motion synthesis. Additionally, to the best of our knowledge, we are the *first to demonstrate the benefits of using vector quantization (VQ-VAE [56]) to achieve improved motion synthesis results*. In essence, rather than relying on the addition of Perlin noise [47] for improved realism, we learn the fine details of realistic motion in a data-driven way.

3. Method

Our goal is to model the conversational dynamics between a speaker and a listener. To test whether our model captures the subtleties of face-to-face communication, we synthesize the interactional motion responses of the listener, which are known to be essential to the flow of conversation [12, 36, 38]. We define the following task: *given the 3D facial motion*
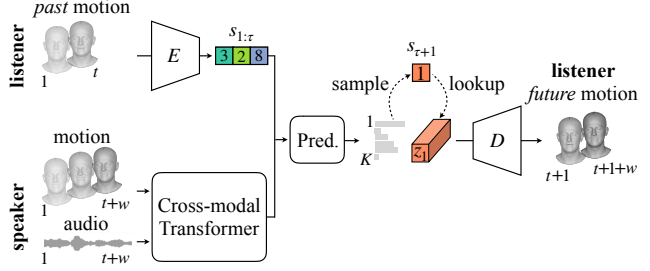

Figure 2. **Overview**: We predict a distribution over future listener motion conditioned on multimodal inputs from a speaker. We use cross-modal attention to fuse the speaker audio and motion input, and a novel sequence-encoding VQ-VAE to discretize past listener motion. Our autoregressive Predictor outputs a distribution over the K discrete codebook indices, from which we sample a code for the next timestep. We obtain the continuous future listener motion by decoding the sampled codebook index.

and audio of the speaker, we autoregressively predict the corresponding facial motion of the listener.

To represent the ongoing flow of conversation, we define a transformer-based predictor, $\mathcal{P}$, that learns to model temporally long-range patterns in the input sequence (Sec. 3.4). The predictor takes two inputs: one corresponding to the speaker and the other to the listener (Figure 2). To model the speaker's audio and facial motion, we introduce a motion-audio cross-modal transformer that learns to fuse the two modalities (Sec. 3.3). To represent the manifold of realistic listener facial motion, we extend VQ-VAE [56] to the domain of motion synthesis and learn a codebook of a discrete latent space (Sec. 3.2). This discrete representation enables us to predict a multinomial distribution over the next timestep of motion. Thus, the output of the autoregressive predictor is a distribution over possible synchronous and realistic listener responses, from which we can sample multiple trajectories.

3.1. Problem Definition

Let $\mathbf{F} = \{\mathbf{f}_i\}_{i=1}^{T}$ be a temporal sequence of facial motions $\mathbf{f}_i$. We use $\mathbf{F}^S$ and $\mathbf{F}^L$ to denote the motion of the speaker and listener respectively. For each timestep $t \in [1, T]$, we take as input a speaker's facial motion $\mathbf{F}_{1:t}^S = (\mathbf{f}_1^S, \cdots, \mathbf{f}_t^S)$ and their corresponding speaker audio sequence $\mathbf{A}_{1:t}^S$, along with any previously predicted past listener motion $\hat{\mathbf{F}}_{1:t-1}^L$, if available. Our predictor, $\mathcal{P}$, then autoregressively predicts the corresponding listener facial motion one time-step at a time:

$$\hat{\mathbf{f}}_t^L = \mathcal{P}(\mathbf{F}_{1:t}^S, \mathbf{A}_{1:t}^S, \hat{\mathbf{F}}_{1:t-1}^L), \qquad (1)$$

where $\mathcal{P}$ learns to model the distribution over the next timestep of listener motion

$$p(\hat{\mathbf{f}}_t^L | \mathbf{F}_{1:t}^S, \mathbf{A}_{1:t}^S, \mathbf{F}_{1:t-1}^L). \qquad (2)$$

To obtain speaker-only audio, we filter out all listener audio back-channels using sound source separation [45]. To

represent the motion, we estimate the 3D facial expressions and orientations from video frames of human conversations using a 3D Morphable Face Model (3DMM) [4, 8, 41, 46]. 3DMMs are parametric facial models that allow us to directly regress disentangled coefficients corresponding to facial expression, head orientation, and identity-specific shape from a single image [60]. This process results in facial expression coefficients $\beta_t \in \mathbb{R}^{d_m}$, where d_m is the dimension of the expression coefficient, a normalized 3D head pose $R_t \in SO(3)$, and shape coefficents that we discard to obtain an identity-agnostic representation. Our facial representation at time t, $\mathbf{f}_t \in \mathbb{R}^{d_m+3}$, is a concatenation of expression and orientation (in Euler angles):

$$\mathbf{f}_t = [\beta_t, R_t]. \tag{3}$$

We normalize facial orientation by computing the mean frontal face direction per video (*i.e.*, orientation at rest pose) and align all head poses in the sequence with respect to this rest pose. This allows us to achieve a camera-view agnostic representation. In contrast to the 2D representations used in some prior works [17, 44], our 3D representation is invariant to changes in facial shape, scale, and camera pose, allowing us to generalize across new faces and camera viewpoints.

3.2. Quantized Listener Motion Codebooks

We extend the use of VQ-VAE [56] to produce multiple realistic modes of different listener responses. VQ-VAE was originally proposed as a method to learn a quantized codebook of image elements from which images could be synthesized autoregressively. Convolutional architectures were used both for learning the codebook and for recombining the discrete elements into images [56]. While the synthesis step was later replaced by transformer architectures that can learn long-range connections [16], image-generation approaches employ a convolutional encoder-decoder pair. This is well-suited for images but not for temporal sequences where convolving over the temporal domain may lose high-frequency information. We design a novel sequence-encoding VQ-VAE where we utilize transformers for the encoder-decoder pair. To the best of our knowledge, we are the first to apply a VQ-VAE to the domain of motion generation.

The advantages of this method are three-fold: (1) it allows us to predict a multinomial distribution over future motion from which we can sample many possible output modes, (2) using the learned discrete latent codes allows us to stay on the manifold of realistic motion ensuring no drift occurs (a problem for methods that directly regress continuous outputs [3]), and (3) it produces realistic motion that captures high-frequency movements.

Specifically, we train a VQ-VAE transformer encoder E and decoder D. To handle the temporal nature of the input, we learn to model longer listener motion sequences in terms of shorter temporal components. Rather than considering static expressions/rotations independently, the latent embedding covers multiple frames, allowing it to learn motion dynamics. The latent embedding represents motion segments of temporal window size $w \ll T$ from a discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K}$, where $z_k \in \mathbb{R}^{d_z}$, that we jointly learn with E and D. $\mathcal{Z}$ maps each of the K codebook entries to a discrete code element of dimension d_z. As shown in Figure 3, we can then approximate any raw listener motion segment $x = \mathbf{F}_{1:T}^{L} \in \mathbb{R}^{T \times (d_m+3)}$ of length T in three steps. First, we encode the sequence $\hat{z} = E(x) \in \mathbb{R}^{\tau \times d_z}$, where $\tau = \frac{T}{w}$ is the length of the patch-wise encoded sequence. Second, we obtain the quantized sequence, z_q, via an element-wise quantization function $\mathbf{q}(\cdot)$ that maps each element of the encoded sequence $\hat{z}$ to its closest codebook entry:

$$z_q = \mathbf{q}(\hat{z}) := \left(\arg\min_{z_k \in \mathcal{Z}} \|\hat{z}_t - z_k\| \right) \in \mathbb{R}^{\tau \times d_z}. \tag{4}$$

Finally, the reconstruction $\hat{x} \approx x$ is given by:

$$\hat{x} = D(z_q) = D(\mathbf{q}(E(x))). \tag{5}$$

We train E, D and the codebook with the loss function [56],

$$\begin{aligned} \mathcal{L}_{\text{VQ}}(E, D, \mathcal{Z}) = &\|x - \hat{x}\| \\ &+ \|\text{sg}[E(x)] - z_q\| \\ &+ \|\text{sg}[z_q] - E(x)\|, \end{aligned} \tag{6}$$

where $\|x - \hat{x}\|^2$ is a reconstruction loss, $\text{sg}[\cdot]$ is a stop-gradient operation, and $\|\text{sg}[z_q] - E(x)\|_2^2$ is a "commitment loss" [56]. After learning the codebook of listener motion, we use the pretrained encoder to quantize the listener motion input to the predictor (Figure 2).

3.3. Cross-Modal Attention for Speaker Input

From the speaker, we take as input both audio $a = \mathbf{A}_{1:t+w}^{S}$, and facial motion $m = \mathbf{F}_{1:t+w}^{S}$. Here, w is the amount of additional future context we see from the speaker. This context acts as a feedback delay that is beneficial in improving learned synchrony for robotics [57]. In contrast to the listener motion, we do not quantize the speaker inputs. While we experimented with both options, we found that speaker motion quantization did not improve results, and quantizing the audio deteriorated the results significantly. We conclude that while quantization is beneficial for predicted motion, for the quality of results as well as sampling capabilities, it is not advantageous for input modalities.

We learn to fuse the audio and motion modalities together using cross-modal attention. Cross-model attention of text and audio [1] or language and vision [31, 42, 55] has been shown to outperform early or late fusion. We extend its use to successfully fuse information from *motion* and audio, a task that proved difficult to previous approaches [33]. We additionally experimented with a naive method of concatenating audio and motion, but this resulted in empirically
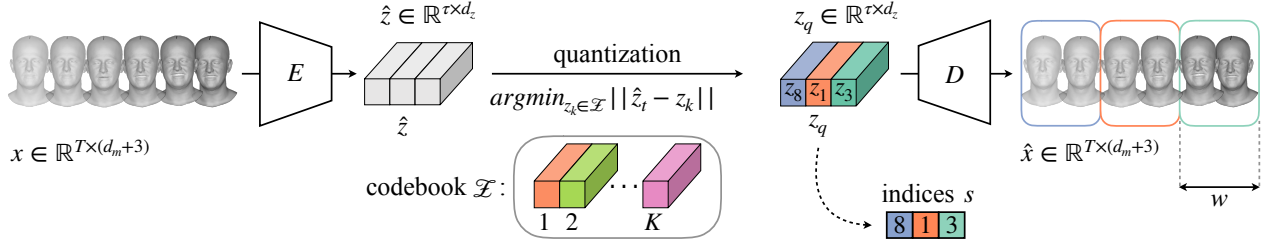
Figure 3. **Motion VQ-VAE that learns a discrete listener motion codebook.** The input is a T length sequence of raw listener facial motion (expression coefficients and 3D head rotations). The *transformer* sequence-encoder E compresses the input into an embedding that gets mapped to its closest quantized codebook element in $\mathcal{Z}$. The *transformer* decoder D decodes the quantized embedding into an approximate reconstruction of the input. We train on a reconstruction loss and commitment loss (Eq. 6). Not only does the VQ-VAE allow us to learn a representation robust to drift from autoregressive inference, it also enables non-deterministic motion synthesis.

worse results due to overly-long conditioning sequences. Applying cross-modal attention along a temporal sequence also allows different modalities to discover some temporal re-alignment [1]. This is especially helpful for encoding speaker inputs since a speaker's motion may not always align with their speech (*e.g.* delay for dramatic effect).

We compute the Queries Q_a for the cross-modal attention operation from the audio input, and the Keys K_m and Values V_m from the motion. We then apply a series of cross-modal attention blocks on the motion modality, where the audio queries are always computed from the raw audio:

$$\text{attention}_{m \to a} = \text{softmax}\left(\frac{Q_a K_m^\top}{\sqrt{d_k}}\right) V_m. \quad (7)$$

Here, d_k is the transformer hidden dimension. The cross-modal transformer outputs an intermediate embedding that incorporates information from both the audio and motion of the speaker. Additional convolutional layers temporally downsample the sequence to match the size of the quantized listener sequence. The final speaker encoding is an embedding $m' \in \mathbb{R}^{(\tau+1) \times d_k}$. We experimentally verify that this method of fusion outperforms others (Table. 1).

3.4. Listener Motion Predictor

We design a transformer-based predictor module, $\mathcal{P}$, to capture long-range correlations in the input data. Building off [40], we employ full-attention masking on the inputs, which has shown promising results in generating long-range motion in an auto-regressive manner. However, with our discrete latent code representation, our model is additionally able to capture multiple modes of outputs by predicting the distribution of possible next motions. Furthermore, we enable multi-modal inputs by means of cross-attention.

$\mathcal{P}$ takes as input the multimodal speaker embedding m' as well as the sequence of previously predicted listener motion. Rather than representing the listener quantized motion as a sequence of codebook vectors z_q, for the purpose of prediction we use the parallel representation of a sequence of corresponding codebook *indices*, $s = s_{1:\tau} \in \{1, ..., K\}^\tau$. Specifically, we discretize past continuous listener motion

$x = \mathbf{F}_{1:t}^L$ by encoding it via the pre-trained encoder E and quantization $\mathbf{q}$ (Section 3.2). We then obtain the sequence of *indices* of the nearest codebook entry per element, via $\mathbf{I}(\cdot)$, an element-wise inverse-lookup function that returns the index of a given codebook element

$$s_{1:\tau} = \mathbf{I}(\mathbf{q}(E(x))). \quad (8)$$

Given speaker input m' and listener input $s_{1:\tau}$, the predictor outputs $p(s_{\tau+1}) \in \mathbb{R}^K$, the multinomial distribution of the next listener codebook index across the K entries:

$$p(s_{\tau+1}) = \mathcal{P}(m', s_{1:\tau}). \quad (9)$$

We can then sample from $p(s_{\tau+1})$ to obtain an index k into the codebook $\mathcal{Z}$. We perform a codebook lookup to retrieve the corresponding quantized element z_k of listener motion, which we pass through the decoder D. The output is the predicted continuous future listener motion $\hat{y} = \hat{\mathbf{F}}_{t+1:t+1+w}^L$ of length w. We train our network with a cross entropy loss on the codebook index $s_{\tau+1}$:

$$\mathcal{L}_{\mathcal{P}} = \mathbb{E}_{y \sim p(y)}[-\log(p(s_{\tau+1}))], \quad (10)$$

where the target codebook index at $\tau + 1$ is computed from ground truth future facial motion $y = \mathbf{F}_{t+1:t+1+w}^L$.

At train time, we follow teacher-forcing and use ground truth listener motion y as past listener input. We randomly mask prior timesteps $\in [1, \tau]$ to facilitate autoregressive learning. At test time we input zeros for timesteps without prior listener predictions, and adjust the masking to ignore these timesteps. This allows us to autoregressively predict future listener motion for arbitrary length input. *No ground truth past listener motion is seen by the network at test time.*

4. In-the-wild Conversational Dataset

Due to the recent COVID-19 pandemic, videotaped interviews have migrated towards teleconferencing platforms that feature a split-screen panel with the host on one side of the screen and interviewee on the other. This setup is especially advantageous for studying face-to-face communication since
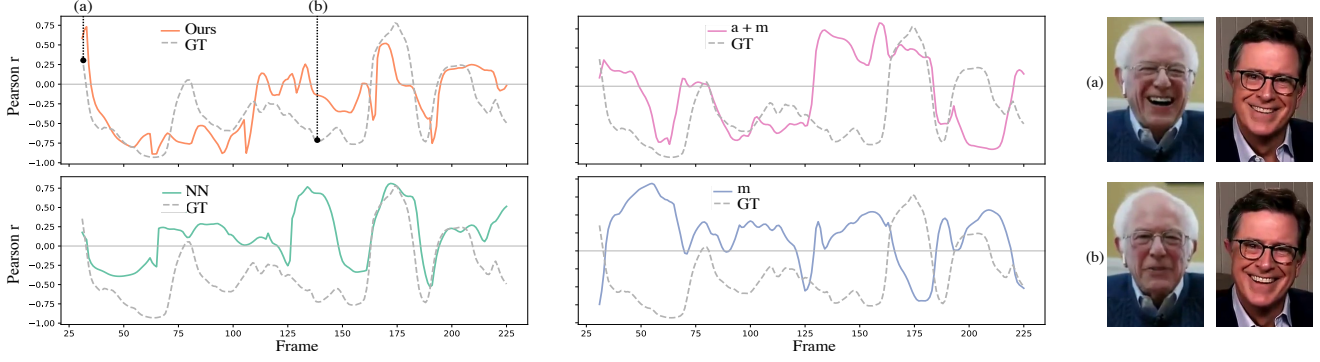
Figure 4. Synchrony of expressions between speaker and listener measured by PCC across a sequence. We convert the expression sequence to a 1D lip curvature time-series according to [21]. **Ours** best matches the synchrony seen in ground truth. **NN** produces sequences that are too synchronous with the speaker. **a+m** and **m** fail to follow the major trends seen in ground truth, such as periods of (a) high synchrony when both the listener and speaker are laughing, and (b) low/no synchrony when the speaker speaks and the listener continues smiling.

both individuals directly face the camera. To cover a broad range of expressions from diverse settings and people, we extract the facial motion and audio for 72 hours of videos from 6 YouTube channels. Each channel features a plethora of interviewees and hosts from a variety of backgrounds.

We leverage a state-of-the-art facial expression extraction method, DECA [18], to recover the 3D head pose and expression coefficients from in-the-wild videos. DECA estimates the pose, expression, and shape parameters according to the FLAME 3DMM [41]. The 3DMM defines 50 expression coefficients along with a 3D jaw rotation ($d_m = 53$), and 3D head rotation in Euler angles as described in Sec. 3.1. For audio, we use sound-source separation [45] to isolate the speaker's voice. We use these expressions, poses, and speaker-only audio as pseudo ground-truth to train our codebook (Eq. 6) and prediction model (Eq. 10). See Supp. for details. ***We release this large-scale, novel dataset.***

5. Experiments

We evaluate our model's ability to effectively translate the speaker's audio and motion into corresponding listener motion. We employ an extensive set of quantitative metrics to measure the realism, diversity, and synchrony of the listener's facial motion. Further, we perform a perceptual study to corroborate quantitative results. All evaluations are done against the raw ground truth listener motion y. We discuss person-agnostic listener models in Supp.

Implementation Details. We use $w = 8, T = 64, K = 200$, $d_z = 256, t = 32$. We add random masking of input past listener motion. While we train on many different input speaker identities, each codebook and predictor model is trained on a specific listener (*e.g.* person-specific listener behavior for any speaker input). For all, we use a train/val/test split of 70%/20%/10%. Quantitative results are aggregated over all listener models. At test-time, we use nucleus sampling [28].

To improve the visual perceptibility of our results, we also train a person-specific mesh-to-pixel visualization module

to directly translate 3DMM predictions to a picture of the listener (Figure 1). See Supp. and video. However, since photorealistic generation is not the main focus of our work, all evaluations are done on the 3D mesh reconstructions, which are the direct outputs of our model.

5.1. Experimental Setup

Evaluation Metrics. Quantifying motion realism is a difficult problem that cannot be reduced to a single metric. We thus evaluate our predictions along multiple axes based on a composition of metrics from prior work. Our evaluation suite is based on the notion that good listeners should display (1) realistic and (2) diverse motion that is (3) synchronous with the motion of the speaker. We assess expression and rotation separately according to these three pillars:

• *L2:* Distance to ground truth expression coefficients/pose.

• *Frechet distance for realism:* Motion realism measured by distribution distance between generated and ground-truth motion sequences following [40]. We directly calculate the Frechet distance (FD) [27] in the expression space $\mathbb{R}^{T \times d_m}$ or the head pose space $\mathbb{R}^{T \times 3}$ on the full motion sequence.

• *Variation for diversity:* Variance in motion across a sequence. We calculate the variance across the time series sequence of expression coefficents or 3D rotations.

• *SI for diversity:* Diverseness of predictions. As in [59], we empirically run k-means to cluster all listener expressions/rotations from training set. We compute avg. entropy (Shannon index) of the cluster id histogram of predicted sequences. $k = 15, 9$ for expression, rotation, respectively.

• *Paired FD for synchrony:* Quality of listener-speaker dynamics measured by distribution distances on listener-speaker *pairs* (P-FD). Calculated FD [27] on concatenated listener-speaker expression $\mathbb{R}^{T \times (d_m + d_m)}$/ pose $\mathbb{R}^{T \times (3+3)}$.

• *PCC for synchrony:* Pearson correlation coefficient (PCC), popular metric used to quantify *global* synchrony in psychology [6,50]. Measures how a listener covaries with a speaker

	Expression						Rotation					
	L2 ↓	FD ↓ (10^3)	variation	SI	P-FD ↓ (10^3)	PCC	L2 ↓	FD ↓ (10^2)	variation	SI	P-FD ↓ (10^2)	PCC
GT			*2.90*	*2.61*		*0.09*			*0.81*	*1.96*		*0.008*
NN motion	45.76	20.66	2.79	2.36	21.94	0.02	6.44	2.78	0.90	1.91	4.06	0.006
NN audio	52.67	31.98	2.70	2.41	33.81	0.02	7.61	5.54	0.93	2.01	6.87	0.007
Random	54.58	43.53	2.76	2.49	45.25	0.01	8.14	6.51	0.90	1.94	7.83	0.005
Median	43.18	64.48	0.00	0.00	64.77	-	6.35	15.50	0.00	0.00	15.50	-
Mirror	53.90	43.56	3.73	2.99	75.30	1.00	7.80	6.01	1.22	1.99	16.88	1.000
Delayed Mirror	53.95	43.79	3.78	2.88	76.72	0.98	7.82	5.99	1.31	1.93	16.96	0.999
LFI [33]	50.07	43.63	1.15	1.33	54.34	0.89	9.00	9.80	0.17	1.07	12.36	0.034
Random Expression	129.34	524.69	62.23	1.17	526.46	0.00	27.67	257.06	62.39	1.06	257.16	0.002
Ours Random Walk	52.68	40.45	1.99	2.26	42.55	0.01	7.14	5.74	0.60	1.37	7.79	0.001
Ours	**33.16**	**3.55**	2.01	2.48	**5.15**	0.07	**4.75**	**0.81**	0.62	1.82	**0.87**	0.008

Table 1. **Baselines.** Comparison against ground truth annotations (GT) on in-the-wild data. ↓ indicates lower is better; for no arrow, closer to GT is better. We bold best performances that are statistically significant. For FD and P-FD, results shown in units indicated above.

over a 1D time series. We calculate lip curvature [21] to measure smile synchrony (Fig. 4). For rotation, we measure synchrony in up/down head motion (nods).

• *TLCC for synchrony:* We further analyze the leader-follower relationship between our generated listeners and the input speakers by calculating the time lagged cross correlation (TLCC) [6]. For $x \in [0, 60]$ frames (up to 2s) we shift the speaker forward by x frames and calculate the correlation on the delayed speaker and corresponding listener. The peak correlation indicates when the two time series are most synchronized. We also use this analysis to find the optimal delay for **Mirror Delay** baseline below.

Baselines. We compare to the following baselines:
• **NN motion:** A segment-search method commonly used for synthesis in graphics. Given an input speaker motion, we find its nearest neighbor from the training set and use its corresponding listener segment as the prediction. We found NN on the full 64-frame sequence to work better than NN on smaller subsequences that are then interpolated together.

• **NN audio:** Same as above, but we find NN via audio embeddings obtained from a pretrained VGGish [26] model.

• **Random**: Return a randomly-chosen 64-frame motion sequence of a listener from the training set.

• **Median**: Simple yet strong baseline exploiting prior that listener is often still. Median expression/pose from train set.

• **Mirror**: Return the speaker's motion smoothed.

• **Delayed Mirror**: Here we mirror the speaker's smoothed motion delayed by 17 frames (≈ 0.5 s). While [17] delayed by 90 frames, we analytically found the optimal lag according to time lagged cross correlation as discussed above.

• **Let's Face It (LFI)** [33]: SOTA interlocutor-aware 3D avatar generation re-trained on our data. Details in Supp.

• **Random Expression**: Walk over 3DMM space; returns a random face at each timestep.

• **Ours Random Walk**: Walk over codebook indices.

5.2. Quantitative Results

Table 1 shows our proposed method outperforms all other competing methods across a variety of metrics. Overall, **Ours** achieves the best balance of performance across the various metrics. Rather than evaluating on L2 performance alone, our full suite of metrics provides a well-rounded view of the qualities of good listeners. For instance, while **Median** performs competitively against **Ours** on L2, it suffers in terms of motion diversity (variation, SI). As a result, this baseline produces less realistic listeners, as noted by our realism metrics (FD, P-FD). However, more variation in the facial gestures is not necessarily better. While **NN motion**, **NN audio**, and **Random** produce diversity similar to real motion, the expression synchrony (PCC) for these baselines is severely lacking. The incongruous listeners hinder the realism of the dyad as a whole (P-FD). That said, a mime that mirrors the speaker like **Mirror** and **Mirror Delay** looks uncanny due to excessive variation and synchrony. **Ours** delicately balances realism, diversity, and synchrony.

The weaker performance of **LFI** [33] demonstrates the advantages of our approach. **LFI** [33] was far less robust when re-trained on our in-the-wild data. Unable to learn realistic listener motion, **LFI** [33] defaulted to mirroring the speaker, resulting in excessively high synchrony (PCC) and worse realism (FD, P-FD). Even when evaluated on the LFI [33] dataset, ours outperforms. These results and visual comparisons in Supp.

Additionally, we quantitatively demonstrate a major advantage of our method's VQ-VAE in learning a robust and realistic manifold of listener motion. **Ours Random Walk** is competitive against **Random**, where we sample full sequences of *real motion*. It significantly outperforms **Random Expression**, where we randomly sample static expressions and rotations at each timestep. This demonstrates that random walks along the codebook still produce realistic motion, though it may not be in sync with the speaker.

Finally, the average TLCC calculated for **GT** and **Ours**

	audio	motion	VQ	CA	Expression						Rotation					
					L2↓	FD↓ (10^3)	variation	SI	P-FD↓ (10^3)	PCC	L2↓	FD↓ (10^2)	variation	SI	P-FD↓ (10^2)	PCC
GT							*2.90*	*2.61*		*0.09*			*0.81*	*1.96*		*0.008*
NoVQ $a+m$	✓	✓	✗	✓	36.06	16.60	0.55	1.69	18.49	0.05	4.99	3.64	0.17	1.21	3.78	0.006
m	✗	✓	✓	-	38.32	4.10	1.91	2.46	5.69	0.12	5.47	0.96	0.57	1.80	1.02	0.009
a	✓	✗	✓	-	39.37	4.11	1.93	2.47	5.86	0.06	5.80	0.91	0.61	1.78	0.98	0.007
$a+m$	✓	✓	✓	✗	38.05	4.01	1.93	2.45	5.67	0.11	5.50	0.87	0.58	1.84	0.93	0.009
Full	✓	✓	✓	✓	**33.16**	**3.55**	2.01	2.48	**5.15**	0.07	**4.75**	**0.81**	0.62	1.82	**0.87**	0.008

Table 2. **Ablations.** Effect of ablating key components of our method. ↓ indicates lower is better; for no arrow, closer to GT is better. CA denotes cross-attention. We bold best performances that are statistically significant. For FD and P-FD, results shown in units indicated above.

were both ≈ 17 frames, both reflecting an average listener response time of ≈ 0.5s. As mentioned above, we use this response time as the optimal delay for **Mirror Delay** baseline. See Supp. for full analysis.

Model Ablations. Table 2 quantifies the contributions of each component of our method. In **NoVQ a+m**, we remove the VQ-VAE and use raw listener motion as the input and output representations. **NoVQ a+m** produces unrealistic, overly smoothed sequences. Adding the VQ-VAE gives a significant performance boost, which further confirms the importance of the codebook in generating realistic motion. Furthermore, we demonstrate that utilizing both audio and motion as input **a+m** via concatenation slightly improves performance over using just one or the other (**a** and **m**). However, **Ours** achieves a more substantial improvement when combining both modalities via cross-attention (**CA**). See Supp. for details of ablation architectures.

5.3. Qualitative Results

To corroborate our quantitative results and gain insight into how our synthesized listeners perceptually compare to real motion, we conducted an A/B test on Amazon Mechanical Turk. Since all quantitative trends were consistent across all listener identities, we randomly chose a single identity for the evaluation. We visualized listener motion using videos of grayscale 3D facial meshes.

Participants watched a series of video pairs. In each pair, one video was generated from our model; the other was produced by an ablation or a baseline. Participants were then asked to identify the video containing the listener that looks like they are listening and paying *more* attention to the speaker. Videos of 8 seconds each of resolution 849×450 (downsampled from 1132×600 in order to fit two videos vertically stacked on different screen sizes) were shown, and after each pair, participants were given unlimited time to respond. Since the most tell-tale moments for when a listener is truly listening are during defining moments (speaker tells a joke, shares a sad story, *etc.*) that illicit strong responses, we manually curated such notable moment sequences from our held-out test data. We then randomly sampled 50 from these sequences and predicted a corresponding listener 3D facial motion sequence using each method. For every test

sequence, each A/B comparison was made by 3 evaluators.

We compared our strongest baseline **NN motion** and ablation **a+m** to our proposed model and recorded the percentage of times our method was preferred over the baseline models or vice versa. **Ours** significantly outperformed. 75.3% of the total 150 evaluators preferred **Ours** over **NN**, and 71.1% preferred **Ours** over **a+m**. These statistics reflect the quantitative trends in Table 2. Furthermore, in a comparison against avatars rendered from ground truth listeners, evaluators preferred **Ours** 50.1% of the time. This *highlights the perceptual realism of our predicted listener motion*.

6. Discussion

In this work, we explored the synchronicity of motion between a speaker and a listener. To this end, we employ a motion-audio cross-attention transformer to handle the multiple modalities of speaker inputs. Furthermore, we enable non-deterministic motion synthesis with a VQ-VAE. Trained on a novel, in-the-wild dataset of dyadic conversations, our method autoregressively outputs convincing 3D listener facial motion that correlate with a given speaker.

While videotaped teleconferencing data lends itself to data collection, it has inherent limitations (*e.g.* no eye contact, time delays introduced by remote connections, *etc.*). A future direction would be to apply this study to in-person conversations, which would allow us to incorporate gaze. Furthermore, as we only model listener motion in response to a speaker, modeling the full dyadic cycle of back-and-forth effect remains for future work. While our goal is to understand conversational dynamics, we discuss concerns for misuse of this technology in Supp. Please see Supp. for result video, per-listener results, implementation details, ablation architectures, multiple mode output evaluation, *etc.*

References

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020. 4, 5

[2] 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. 3

[3] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 4

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 4

[5] 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

[6] Steven Boker, Minquan Xu, Jennifer Rotondo, and Kadijah King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods*, 7:338–55, 10 2002. 6, 7

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008. 2

[8] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 2013. 4

[9] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 3

[10] 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, page 413–420, New York, NY, USA, 1994. Association for Computing Machinery. 2

[11] Justine Cassell and Kristinn R. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538, 1999. 1

[12] Tanya L Chartrand and John A Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999. 1, 3

[13] Jin Hyun Cheong, Zainab Molani, Sushmita Sadhukha, and Luke J Chang. Synchronized affect in shared experiences strengthens social connection, Jan 2020. 2

[14] Hang Chu, Daiqing Li, and Sanja Fidler. A face-to-face neural conversation model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7113–7121, 2018. 3

[15] W. S. Condon and W. D. Ogston. Sound film analysis of normal and pathological behavior patterns. *The Journal of Nervous and Mental Disease*, 143(4), 1966. 2

[16] 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. 4

[17] Will Feng, Anitha Kannan, Georgia Gkioxari, and Larry Zitnick. Learn2smile: Learning non-verbal interaction through observation. *IROS*, 2017. 2, 3, 4, 7

[18] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021. 6

[19] Guy Gaziv, Lior Noy, Yuvalal Liron, and Uri Alon. A reduced-dimensionality approach to uncovering dyadic modes of body motion in conversations. *PloS one*, 12(1):e0170786, 2017. 2

[20] 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. 3

[21] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A Efros. A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–7, 2015. 6, 7

[22] 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

[23] David Greenwood, Stephen Laycock, and Iain Matthews. Predicting head pose in dyadic conversation. In *International Conference on Intelligent Virtual Agents*, pages 160–169. Springer, 2017. 3

[24] Yuval Hart, Efrat Czerniak, Orit Karnieli-Miller, Avraham E. Mayo, Amitai Ziv, Anat Biegon, Atay Citron, and Uri Alon. Automated video analysis of non-verbal communication in a medical setting. *Frontiers in Psychology*, 7:1130, 2016. 2

[25] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 3

[26] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee*

international conference on acoustics, speech and signal processing (icassp), pages 131–135. IEEE, 2017. 7

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017. 6

[28] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 6

[29] 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

[30] 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. 3

[31] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 4

[32] 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. 3

[33] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020. 2, 3, 4, 7

[34] 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. 3

[35] Yvonne Jung, Arjan Kuijper, Dieter W Fellner, Michael Kipp, Jan Miksatko, Jonathan Gratch, and Daniel Thalmann. Believable virtual characters in human-computer dialogs. In *Eurographics (State of the Art Reports)*, pages 75–100, 2011. 2

[36] Adam Kendon. Movement coordination in social interaction: Some examples described. *Acta psychologica*, 32:101–125, 1970. 1, 3

[37] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018. 3

[38] Marianne LaFrance. Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. *Social Psychology Quarterly*, 42(1):66–70, 1979. 1, 3

[39] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020. 3

[40] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv preprint arXiv:2101.08779*, 2021. 3, 5, 6

[41] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 4, 6

[42] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 4

[43] 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

[44] 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, 3, 4

[45] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*, 2018. 3, 6

[46] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee, 2009. 4

[47] Ken Perlin and Athomas Goldberg. Improv: A system for scripting interactive actors in virtual worlds. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 205–216, 1996. 3

[48] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3

[49] Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony of head- and body-movement in psychotherapy: different signals have different associations with outcome. *Frontiers in Psychology*, 5:979, 2014. 1

[50] Marcel Riehle, Jürgen Kempkensteffen, and Tania M. Lincoln. Quantifying facial expression synchrony in face-to-face dyadic interactions: Temporal dynamics of simultaneously recorded facial emg signals. *Journal of Nonverbal Behavior*, 41(2):85–102, Jun 2017. 6

[51] Paul Ruvolo, Daniel Messinger, and Javier Movellan. Infants time their smiles to make their moms smile. *PLOS ONE*, 10(9):1–10, 09 2015. 2

[52] José Serra, Verónica Orvalho, and Darren Cosker. Behavioural facial animation using motion graphs and mind maps. In *Proceedings of the 9th International Conference on Motion in Games*, pages 161–166, 2016. 2

[53] Sinan Sonlu, Uğur Güdükbay, and Funda Durupinar. A conversational agent framework with multi-modal personality expression. *ACM Transactions on Graphics (TOG)*, 40(1):1–16, 2021. 2

[54] Colwyn Trevarthen and M. Bullowa. Communication and cooperation in early infancy: A description of primary intersubjectivity. *Before Speech (Cambridge)*, pages 321–347, 01 1979. 2

[55] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 4

[56] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6309–6318, 2017. 2, 3, 4

[57] Auriel Washburn, Rachel W Kallen, Maurice Lamb, Nigel Stepp, Kevin Shockley, and Michael J Richardson. Feedback delays can enhance anticipatory synchronization in human-machine interaction. *PloS one*, 14(8):e0221275, 2019. 4

[58] Li Zhang, Noah Snavely, Brian Curless, and Steven M Seitz. Spacetime faces: High-resolution capture for˜ modeling and animation. In *Data-Driven 3D Facial Animation*, pages 248–276. Springer, 2008. 2

[59] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6194–6204, 2020. 6

[60] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, 2018. 4