DIM: Dyadic Interaction Modeling for Social Behavior Generation

Minh Tran* ⓘ, Di Chang* ⓘ, Maksim Siniukov, and Mohammad Soleymaniⓘ

University of Southern California, Institute for Creative Technologies
`soleymani@ict.usc.edu`

Abstract. Human-human communication is like a delicate dance where listeners and speakers concurrently interact to maintain conversational dynamics. Hence, an effective model for generating listener nonverbal behaviors requires understanding the dyadic context and interaction. In this paper, we present an effective framework for creating 3D facial motions in dyadic interactions. Existing work consider a listener as a reactive agent with reflexive behaviors to the speaker's voice and facial motions. The heart of our framework is Dyadic Interaction Modeling (DIM), a pre-training approach that jointly models speakers' and listeners' motions through masking and contrastive learning to learn representations that capture the dyadic context. To enable the generation of non-deterministic behaviors, we encode both listener and speaker motions into discrete latent representations, through VQ-VAE. The pre-trained model is further fine-tuned for motion generation. Extensive experiments demonstrate the superiority of our framework in generating listener motions, establishing a new state-of-the-art according to the quantitative measures capturing the diversity and realism of generated motions. Qualitative results demonstrate the superior capabilities of the proposed approach in generating diverse and realistic expressions, eye blinks and head gestures.

Keywords: Behavior Generation· Self-supervised Learning · Facial Motions

1 Introduction

Human nonverbal communication is like a dance. In human-human interaction, we concurrently encode and decode verbal and nonverbal messages, constantly attending to and reacting to our counterparts to maintain the flow of the conversation and engage in effective turn-taking [18]. Hence, a socially intelligent agent needs to learn the interactive nature of nonverbal communication to create a seamless and compelling experience. In this paper, we present a method that goes beyond a reactive agent that only considers speakers' motions and voice by learning the listener-speaker dyadic context through self-supervised masked pre-training on a large dataset of human-human dyadic interactions, i.e., CAN-DOR [54]. As illustrated in Figure 1, We aim to generate a listener's lifelike

---

* equal contribution
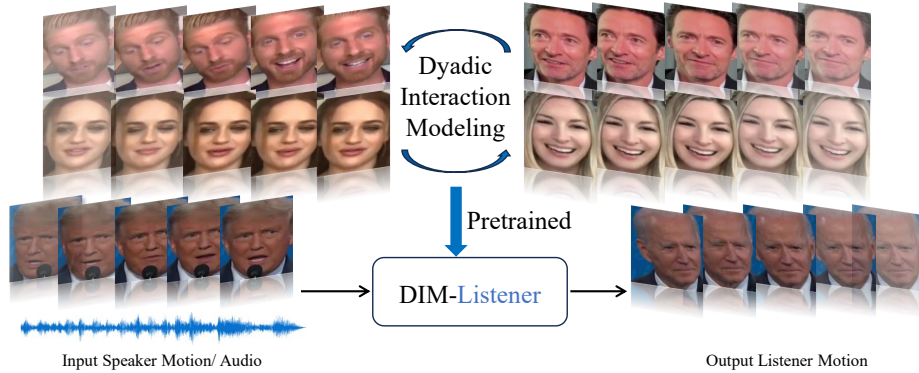
Fig. 1: We propose Dyadic Interaction Modeling, a pre-training strategy that jointly models speakers' and listeners' motions and learns representations that capture the dyadic context. We then utilize the pre-trained weights and feed multimodal inputs from the speaker into DIM-LISTENER. DIM-LISTENER is capable of generating photorealistic videos for the listener's motion.

and plausible head and facial movements in response to a speaker's motions and voice in a manner that resembles the contextualized bi-directional nature of human-human communication. Generating listener reactions has a wide range of applications, including in human-computer interaction [42, 77, 78], virtual reality [35, 41], the metaverse [6, 8, 63], and media forensics [30, 57–59]. Given the ability of DIM to model dyading interactions, it is also able to produce realistic speaker motions.

Despite considerable progress, existing methods are limited, particularly in capturing real-life conversation dynamics, and difficult to generalize across listeners. Relevant work focused on either generating speakers' or listeners' movements. Speaker motion generation [9, 11, 20, 33, 56, 60, 61, 66, 70] primarily focuses on producing videos where lip movements are synchronized with speech, a.k.a., lip syncing. Listener motion generation, on the other hand, automatically interprets the speaker's voice and nonverbal behaviors to generate the listener's nonverbal responses. Traditionally, listener motion generation have not fully addressed the intricacies of dyadic interactions, which involve more than simple reactions to the speaker's utterances including a multifaceted, concurrent exchange of signals and emotions [47, 62, 80]. They were also sometimes tuned to specific listeners and unable to generalize to others [47]. Earlier work used heuristics and simple machine learning models to generate a limited set of reactions, e.g., head nods [28]. The Responsive Listening Head Generation (RLHG) [80] approach mimics the regression techniques found in Speaker Generation [11], leading to a reduction in the unpredictability of responses and a smoothing over of listener motions. Ng et al. [47] proposed Learning2Listen (L2L) to mitigate this issue through a motion categorization strategy. In L2L, a codebook is learned through a VQ-VAE [45] that can be used to generate plausible and smooth

behaviors. Emotional Listener Portrait (ELP) [62] utilizes an Adaptive Space Encoder to map the discretized features combined with emotion to the listener's motion parameters. Similarly, Speaker motion generation techniques [22, 56, 74] have struggled to incorporate the feedback present in actual dyadic interactions. These shortcomings underscore a critical gap in the simulation of naturalistic human interactions, highlighting the need for a more holistic and dyadic modeling approach.

To address these challenges, we introduce Dyadic Interaction Modeling (DIM), an innovative pretraining strategy designed to enhance a model's capacity to simultaneously encode a unified representation from both speaker and listener behaviors. This strategy draws inspiration from the recognition of the importance of bidirectional communication in interaction modeling, as demonstrated by Nojavanasghari et al. [49], and Eskimez et al. [20], who explored aspects of interactive emotion recognition and speech-driven facial animation, respectively. Moreover, the concept of integrating audiovisual information into a cohesive representation, as explored by Karras et al. [36] and Zhou et al. [81], forms a foundational pillar for our approach, suggesting that a holistic and contextualized consideration of both audio and visual cues can significantly enhance animation quality.

Leveraging self-supervised contrastive learning and focusing on the reconstruction of masked-hidden units, DIM captures the intricate exchanges characteristic of human dyadic conversations. This approach is inspired by the success of self-supervised learning techniques in capturing complex data patterns, as evidenced by the works of Chen et al. [12] and Gong et al. [27], which have demonstrated significant advancements in data understanding without explicit labeling.

DIM framework learns dyadic context for generating realistic motions. This framework enables the generation of detailed and lifelike facial expressions and head motions, marking a significant advancement in listener (DIM-LISTENER) and speaker (DIM-Speaker) motion generation. Our method not only generates listener behaviors from speaker audio-visual inputs but could also adeptly produce speaker facial motions from speaker speech (DIM-Speaker), showcasing the framework's ability to model the bidirectional nature of human interactions.

Our main contributions can be summarized as follows:

- We introduce Dyadic Interaction Modeling (DIM), a pre-training strategy that enhances the model's ability to encode a unified representation from both speaker and listener behaviors through self-supervised contrastive learning, focusing on the reconstruction of masked-hidden units.
- Leveraging DIM, we develop DIM-LISTENER, a practical framework for listener motion generation in speaker-listener conversation with detailed and realistic facial expressions and head motions.
- DIM is further leveraged to generate speaker facial behaviors from speaker speech (DIM-Speaker).
- Extensive experiments and visualizations on motion generation demonstrate superior performance and effectiveness of our method.

2   Related Work

2.1   Speech-driven Speaker Generation

There is a wide body of work on automatic facial motion generation [4,7,25,37–40,51,64,69,72,73,83]. A large number of studies focus on 2D facial animation [9, 10, 14, 17, 21, 33, 53, 70, 76, 79]. Traditional procedural techniques [19, 44, 68, 75] rely on a framework of explicit rules for simulating the movements of a talking mouth, such as the use of dominance functions [44] for delineating speech control parameters.

These procedural strategies offer precise control over mouth movement accuracy but are labor-intensive due to the extensive manual tuning required. In contrast, data-driven methods [5, 15, 32, 36, 43, 52, 56, 65, 67] present an alternative, proposing various techniques to generate 3D facial animations with less manual intervention. Among these, Karras et al. [36] propose an end-to-end network with linear predictive coding to generate audio and novel latent codes with facial expression variations. In Zhou et al. [82], the proposed network predicts viseme curves by integrating phoneme groups, landmarks, and audio signals. VOCA [15] offers speaker-independent animation capturing diverse speaking styles, though it primarily affects the lower face. More recently, MeshTalk [56] has made strides in disentangling audio-correlated and uncorrelated facial motions through a categorical latent space. Innovations continue with FaceFormer [22] and CodeTalker [74], which leverage long-term audio context via transformer-based models and augment them with a VQ-VAE motion prior for autoregressive motion synthesis, respectively.

2.2   Speaker-driven Listener Generation

Compared to Speaker Generation, Listener Generation pays more attention to the feedback of the listener's motion to the speaker. Ahuja et al. [1] focus on the non-verbal body behaviors generation, [3] and Greenwood et al. [29] study synchronized conversational agent motion in dyadic conversation adapt speech. Recent methods focus on generating 3D facial motions with additional inputs from the listener, such as text [13] or speech [34, 35]. Song et al. [62] proposed ELP which learns the non-verbal listener motion in a dynamic communication. RealTalk [26] retrieves possible videos of the listener's face with a large language model. The most similar work to our approach is that of Learning2Listen [47], which regresses the discrete listener head motion with VQ-VAE [50]. However, L2L model is unidirectional and does not consider the dyadic context. Unlike this work, the L2L [47] fixes the pre-trained VQ-Decoder to decode continuous signals from discrete motion predictions, we integrate the VQ-Decoder into motion prediction and optimize both discrete units (discrete latent codes) and continuous motions.

3 Method

This section provides an overview of the proposed framework DIM-LISTENER. We first delve into the facial motion representation EMOCA [16], as discussed in Section 3.1. Subsequently, we revisit VQ-VAE [50], one of the core components in DIM-LISTENER, in Section 3.2. We further detail our novel pre-training strategy, Dyadic Interaction Modeling, designed to foster the model's proficiency in developing a unified representation for both speakers and listeners in dyadic conversations. This is followed by an explanation of the fine-tuning stage for DIM-LISTENER, presented in Section 3.3. We finally describe the rendering network [55], transforming the estimated motions into realistic videos in Section. 3.4.

3.1 Facial Motion Representation

Following [48], we utilize EMOCA [16], a 3D morphable model (3DMM), parameters as the representation for facial motions. EMOCA disentangles face shape, expression and head pose in three sets of parameters, allowing for capturing and expressing person-independent behaviors. Compared to DECA [24] used in previous work [46], EMOCA can better capture facial expressions and head poses. Following prior work [48, 62], we extract camera, shape, expression and pose codes from the speakers and listeners frame by frame. Note that only expression code and pose code are used for representing human facial motions while camera code and shape code are only used for face mesh visualization purposes.

3.2 Learning discrete Speaker-Listener motions

Formally, we train two separate VQ-VAE models, denoted as $VQ^{(s)}$ and $VQ^{(l)}$, to handle speaker and listener motions, respectively. For clarity, we present only the broad overview of the VQ-VAE workflow for encoding and reconstructing listener motions. In particular, the encoder $Enc_{VQ}^{(l)}$ processes the listener motions $l_{1:T}$ to produce the continuous latent representation $z_{cont}^{(l)} = Enc_{VQ}^{(l)}(l_{1:T})$. The latent representation is then quantized to $z^{(l)}$ using the nearest vector from the shared codebook C. The reconstruction of the listener motion is then given by: $\hat{l}_{1:T} = Dec_{VQ}^{(l)}(z^{(l)})$.

$VQ^{(l)}$'s training objective incorporates both reconstruction and codebook losses. The reconstruction loss $\mathcal{L}_{recon}$ aims to minimize the difference between the original and reconstructed motions:

$$\mathcal{L}_{recon}^{(l)} = \|l_{1:T} - \hat{l}_{1:T}\|_2^2, \tag{1}$$

where $\|\cdot\|_2$ denotes the squared Euclidean norm. The codebook loss $\mathcal{L}_{codebook}$, comprising the commitment loss and the codebook utilization loss, is defined as:

$$\mathcal{L}_{codebook}^{(l)} = \|sg[E(l_{1:T})] - z^{(l)}\|_2^2 + \beta\|sg[z^{(l)}] - E(l_{1:T})\|_2^2, \tag{2}$$
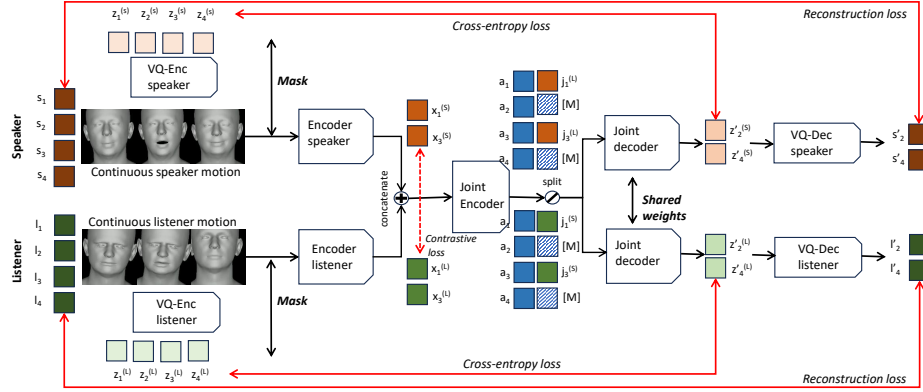
Fig. 2: **Dyadic Interaction Modeling** learns a unified speaker-listener representation from dyadic interactions. 1) The framework takes both the ground-truth speaker motion s and the listener motion l as input. 2) VQ-Encoders of speaker and listener then encode the motions to discrete units (discrete latent codes) $z^{(s)}$ and $z^{(l)}$. 3) The masked speaker's and listener's motions are further encoded and concatenated so that a unified representation is learned with contrastive loss. 4) Then, the split unified representation and speaker audio feature a are decoded into discrete unit predictions $z'^{(s)}$ and $z'^{(l)}$ supervised by cross-entropy loss. 5) Finally, the generated speaker motions s' and listener motions l' are decoded from these discrete unit predictions to optimize the reconstruction loss.

with $sg[\cdot]$ denoting the stop-gradient operator, and β is a hyperparameter that balances the commitment of the encoded vectors to the codebook. The overall loss combines these components:

$$\mathcal{L}_{\mathcal{VQ}}^{(l)} = \mathcal{L}_{recon}^{(l)} + \gamma \mathcal{L}_{codebook}^{(l)}, \tag{3}$$

where γ is a weighting factor for the codebook loss.

Following L2L [47] and ELP [62], which leverage discrete motions to model the complex interplay in human interactions and animate listener motion from speaker audiovisual inputs, we use VQ-VAE [50] to learn discrete motion codebooks for **both speaker and listener** behaviors, and use the learned codebooks to generate discrete latent codes for our Dyadic Interaction Modeling process.

3.3 Dyadic Interaction Modeling

Notations. We denote the input speaker motions as $s_{1:T}$, the input speaker audio features as $a_{1:T}$, the input listener motion as $l_{1:T}$, the encoders (and quantization modules) of the VQ-VAE models trained for speaker motions and listener motions as $Enc_{VQ}^{(s)}$ and $Enc_{VQ}^{(l)}$, respectively. We refer to the decoder of the VQ-VAE models trained for speaker motions and listener motions as $Dec_{VQ}^{(s)}$ and $Dec_{VQ}^{(l)}$, respectively.

The overall pipeline is illustrated in Figure 2. Given $s_{1:T}$ and $l_{1:T}$, DIM-LISTENER first generates $z^{(s)}$ and $z^{(l)}$ as the discrete motion sequences produced by the learned motion codebooks: $z^{(s)} = Enc_{VQ}^{(s)}(s_{1:T})$ and $z^{(l)} = Enc_{VQ}^{(l)}(l_{1:T})$, and uses these sequences as the discrete latent codes to later train the network. It is important to note that both $Enc_{VQ}^{(s)}$ and $Enc_{VQ}^{(l)}$ are frozen to produce consistent discrete latent codes. Then, we independently mask $p\%$ of random frames for both speakers and listeners, i.e.,

$$s_{masked} = Mask_p(s + E_s + E_s^P) \; ; \; l_{masked} = Mask_p(l + E_l + E_l^P) \qquad (4)$$

where E_s and E_l denote the learnable speaker and listener embeddings and E_s^P and E_l^P denote the sinusoidal positional embeddings. s and l are speaker motions and the listener motions from ground truth.

We then feed s_{masked} and l_{masked} to independent role-specific Transformer encoders $Enc^{(s)}$ and $Enc^{(l)}$ to produce $x_{masked}^{(s)}$ and $x_{masked}^{(l)}$. To further enhance $Enc^{(s)}$ and $Enc^{(l)}$ to extract robust features for speaker and listener inputs, we propose using contrastive learning to match speaker-listener pairs. We use the contrastive loss for the task

$$\mathcal{L}_c = -\frac{1}{N} \Sigma_{i=1}^N log[\frac{exp(x_i^{(s)} \cdot x_i^{(l)}/\tau)}{\Sigma_{i \neq k} \; exp(x_i^{(s)} \cdot x_k^{(l)}/\tau) + exp(x_i^{(s)} \cdot x_i^{(l)}/\tau)}] \qquad (5)$$

where $x_i^{(s)}$ and $x_i^{(l)}$ are the temporally mean-pooled feature representations of the ith sample of $x_{masked}^{(s)}$ and $x_{masked}^{(l)}$, respectively.

We further concatenated $x_{masked}^{(s)}$ and $x_{masked}^{(l)}$ along the feature dimension and forward the fused feature into a joint Transformer encoder to produce $j_{masked}^{(s-l)}$ and later split $j_{masked}^{(s-l)}$ back into separate speaker and listener features, i.e.,

$$[j_{masked}^{(s)} \; ; \; j_{masked}^{(l)}] = j_{masked}^{(s-l)} = Enc^{(joint)}([x_{masked}^{(s)} \; ; \; x_{masked}^{(l)}]) \qquad (6)$$

For the masked frames reconstruction task, we pad $j_{masked}^{(s)}$ and $j_{masked}^{(l)}$ with trainable mask tokens at the masked positions as $j^{(s)}$ and $j^{(l)}$. We further add learnable speaker and listener embeddings, along with sinusoidal positional embeddings, to $j^{(s)}$ and $j^{(l)}$ before concatenating them with extracted audio features for predicting the masked discrete motions via a joint Decoder.

$$z'^{(s)} = Dec^{(joint)}([j^{(l)} + E_l^D + E_l^p \; ; \; a_{1:T}])$$
$$z'^{(l)} = Dec^{(joint)}([j^{(s)} + E_s^D + E_s^p \; ; \; a_{1:T}]) \qquad (7)$$

The predicted discrete motions $z'^{(s)}$ and $z'^{(l)}$ are then forwarded into $Dec_{VQ}^{(s)}$ and $Dec_{VQ}^{(l)}$ for final continuous motion predictions.

$$\tilde{s} = Dec_{VQ}^{(s)}(z'^{(s)}) \; ; \; \tilde{l} = Dec_{VQ}^{(l)}(z'^{(l)}) \qquad (8)$$
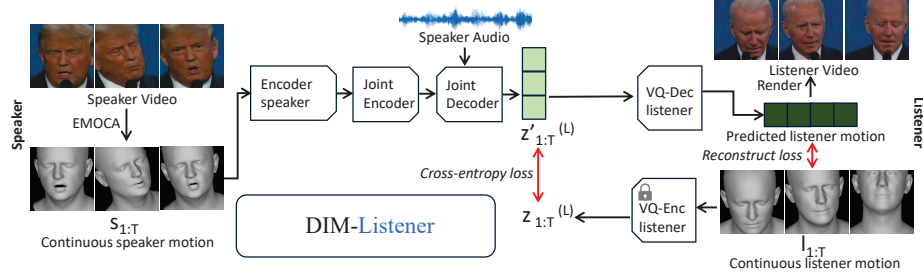
Fig. 3: For fine-tuning the model on listener motion generation, speaker input is not masked, and the listener input is entirely masked. We train the framework with the same cross-entropy loss and reconstruction loss from Dyadic Interaction Modeling while keeping the weights of listener VQ-Encoder fixed.

For reconstruction, we train our network on both a cross-entropy loss on the discrete motions and a reconstruction loss

$$
\begin{aligned}
\mathcal{L}_{rec}^{(s)} &= \Sigma_{t \in M} - log p(z_t^{(s)} | z_t'^{(s)}) + ||s_t - \tilde{s}_t||_2^2 \\
\mathcal{L}_{rec}^{(l)} &= \Sigma_{t \in M} - log p(z_t^{(l)} | z_t'^{(l)}) + ||l_t - \tilde{l}_t||_2^2
\end{aligned}
\tag{9}
$$

where M denotes the masked position during the masking process. Our final loss is defined as

$$
\mathcal{L} = \lambda_1 \mathcal{L}_c + \lambda_2 (\mathcal{L}_{rec}^{(s)} + \mathcal{L}_{rec}^{(l)})
\tag{10}
$$

where λ_1 and λ_2 are hyperparameters.

We then fine-tune the model for listener motion generation with initialized weights from Dyadic Interaction Modeling, as shown in Figure 3. During the fine-tuning phase where $s_{1:T}$ and $a_{1:T}$ are available but $l_{1:T}$ is the target prediction, we do not mask $s_{1:T}$ (i.e., $p_s = 0\%$) while mask entirely the listener input (i.e., $p_l = 100\%$). The VQ-Encoders are frozen while the remaining weights are fine-tuned with the reconstruction loss of the listener motions $\mathcal{L}_{rec}^{(l)}$.

3.4   Photorealistic Render

To visualize the real-human listener video from the predicted expression and pose EMOCA parameters, we re-train PIRenderer, a motion-to-video rendering network [55]. Collecting long-term person-specific portrait videos to train high-quality person-specific models is challenging and out of the scope of this work. Hence, we collect listener-speaker pairs and train a separate rendering network on pairs with the same speaker-listener identities. We feed the first frame of each video as the reference portrait image, together with the EMOCA parameters extracted from the later frames, into the renderer and train it with an image reconstruction loss. In this way, we use the predicted EMOCA motion prediction to drive the reference portrait and generate a vivid photorealistic video, as shown in the right side of Figure 3.

4   Experiments

4.1   Experimental Settings

Listener Behavior Generation. We evaluate the speaker-driven listener motion generation of DIM-LISTENER on ViCo [80] and LM_Listener [48] dataset.ViCo [80] consists of 483 video sequences featuring 50 unique listeners. The dataset that was proposed in LM_Listener [48] is an extended version of the dataset introduced in Learning2Listen [47] (L2L), which contains 2366 training segments, 222 validation segments, and 543 test segments of a single listener (Trevor Noah).

Speaker Behavior Generation. We evaluate the speech-driven speaker motion generation on BiWi [23]. BiWi [23] comprises affective speech recordings and the associated high-resolution, dynamic 3D facial geometry data. It features recordings from 14 participants who were instructed to read 40 sentences in English, recorded at a frame rate of 25 frames per second. On average, each recorded sequence is approximately 4.67 seconds.

Dyadic Interaction Modeling is performed on CANDOR [54] dataset. It consists of a collection of 1,656 conversations in English. With a total duration of over 850 hours, this large corpus includes more than 7 million words, encompassing rich audio, video, and speech transcripts. It enables detailed moment-to-moment analysis of vocal and facial expressions as well as semantic and conversational contexts. CANDOR provides speech diarization and alignment, which was used for speaker/listener designation at the utterance level. The dataset also provides separate speech tracks, so there is no overlapping speech.

Implementation Details For the pre-training stage, we use an Adam optimizer with a learning rate of $1e^{-5}$ and train the model on CANDOR [54] for 100 epochs. We use a masking ratio $p = 75\%$, which is determined with hyperparameter tuning with respect to the pre-training validation reconstruction loss. We use HuBERT [31] as our audio feature extractor. We use the same architectures for the Transformer encoders and decoders of DIM-LISTENER and VQ-VAE, which consists of a transformer with 8 hidden layers, 8 attention heads and an intermediate size of 768. For pre-processing of CANDOR, we use the provided transcript with the dataset to extract utterance-level samples of speaker-listener behaviors. For pre-processing of ViCo [2] and LM_Listener [48], we follow the pre-processing guidelines and train-test splits from prior works [2, 47, 62] for fair comparisons. For all experiments, we extract EMOCA [16], an expressive 3DMMs parameter set, from visual data and adopt it as facial representation for all methods and re-train them for the sake of fair comparison. We follow previous work [47,62,80] and adopt expression code (Exp) and pose code (Pose) of EMOCA [16] parameters as the description of the speaker and listener's motions. We use HuBERT [31] to extract audio features. Since HuBERT is already a robust pre-trained speech encoder, we do not use the audio features in our encoders and directly feed them into our model's decoder. For the implementation source code of our method, please see supplementary materials.

Table 1: Quantitative Comparison of Listener Generation on ViCo [80] Dataset. † denotes the corresponding method has been pre-trained on the CANDOR [54] dataset. ∗ denotes the method didn't release any code, and we re-implement it from scratch on our own.

Method	FD↓		P-FD↓		MSE↓		SID		Var		rPCC↓	
	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose
Random	72.88	0.12	75.82	0.12	2.05	0.03	4.68	3.83	1.49	0.02	0.09	0.17
Nearest Audio	65.77	0.10	68.84	0.10	1.77	0.03	4.81	3.93	1.49	0.02	0.07	0.11
Nearest Motion	42.41	**0.06**	45.33	**0.06**	1.27	0.02	4.62	3.74	1.49	0.02	0.08	0.12
ELP∗ [62]	47.17	0.08	47.48	0.08	0.98	0.02	1.76	1.66	1.49	0.02	0.09	**0.01**
RLHG [80]	39.02	0.07	40.18	0.07	0.86	**0.01**	3.62	3.17	1.52	0.02	0.08	0.02
L2L [47]	33.93	**0.06**	35.88	**0.06**	0.93	**0.01**	2.77	2.66	0.83	0.02	**0.06**	0.08
L2L† [47]	31.03	**0.06**	33.02	**0.06**	0.87	**0.01**	3.53	3.05	0.83	0.02	**0.06**	0.09
GT	-	-	-	-	-	-	5.03	4.07	0.93	0.01	-	-
DIM-LISTENER †	**23.88**	0.06	**24.39**	0.06	**0.70**	0.01	3.71	2.35	1.53	0.02	**0.06**	0.03

Baselines We compare the performance of DIM-LISTENER on Listener Generation with random and nearest neighbor baselines in addition to the following state-of-the-art methods.

- RLHG [80] is a regression method. We use their released official code for a fair comparison.
- L2L [47] utilizes a discrete codebook to synthesize motion patterns. We use their released official code for a fair comparison.
- ELP [62] extracts speaker-style features and emotion vectors. Due to the unavailability of the source code, we re-implement the baseline following the paper. We provide our re-implementation in the supplementary materials.

The random and nearest neighbor baselines are also widely reported in previous work [47, 62, 80], which include:

- We randomly select the facial and head motion parameters in the training data, and inject random small perturbations into the normal distribution.
- We use the smoothed speaker's motion as the listener's motion.
- For arbitrary input speaker motion or audio input, we find its nearest neighbor from the training set and use its corresponding listener motion as output.

Metrics We re-implement all baselines and train them with EMOCA 3DMM as output motion representation. We then evaluate them using the metrics reported in LM_Listener [48] and ELP [62], namely the *Frechet Distance* (FD), *Paired FD* (P-FD) for synchrony , *Mean Squared Error* (MSE), *SI for Diversity* (SID), *Variance* (Var), and *Residual Pearson Correlation Coefficient* (rPCC). We provide more details about the used metrics in the supplementary materials.

Table 2: Quantitative Comparison of Listener Generation on LM_Listener [48] Dataset. † denotes the corresponding method has been pre-trained on the CANDOR [54] dataset. * denotes the method didn't release any code, and we re-implement it from scratch on our own.

Method	FD↓		P-FD↓		MSE↓		SID		Var		rPCC↓	
	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose
Random	63.37	0.10	67.14	0.10	2.14	0.03	4.06	3.49	1.57	0.01	0.08	0.06
Mirror	69.25	0.16	87.11	0.19	1.95	0.03	4.06	3.54	1.51	0.01	0.74	0.86
Nearest Motion	56.00	0.07	58.94	0.08	2.04	0.02	3.94	3.46	1.62	0.01	0.13	0.13
ELP* [62]	22.06	**0.01**	22.09	**0.01**	0.45	**0.01**	5.07	3.64	0.49	0.01	0.09	0.04
RLHG [80]	30.50	0.02	31.74	0.02	0.69	**0.01**	3.84	3.65	0.94	0.01	0.09	0.03
L2L [47]	23.56	0.02	23.94	0.02	0.75	**0.01**	3.50	3.00	0.58	0.01	0.12	0.04
L2L† [47]	21.63	0.02	22.03	0.02	0.73	**0.01**	3.50	3.00	0.67	0.01	0.14	0.03
GT	-	-	-	-	-	-	4.97	3.69	0.78	0.01	-	-
DIM-LISTENER †	**12.49**	**0.01**	**12.95**	0.02	**0.29**	**0.01**	4.53	2.93	0.94	0.01	**0.04**	**0.02**

4.2 Benchmark Performances

Quantitative Results. We retrain RLHG [80], L2L [47] and ELP [62] on the ViCo dataset [80]. Since the Dyadic Interaction Modeling of our pipeline is conducted on CANDOR [54], we also pre-train L2L [47], a recent state-of-the-art method, which adopts a similar approach using VQ-VAE [50] in the model, on CANDOR [54], for fair comparison. Table 1 shows the quantitative comparison of Listener Generation on ViCo [80]. Based on the metrics presented in Table 1, it is clear that our proposed DIM-LISTENER outperforms existing methods by a significant margin. Our method is about 1.5 times better than the current state-of-the-art method L2L [47] in terms of the most important metrics *FD* and *P-FD* of expressions in the Listener Generation task. The performance gain of L2L after being trained on CANDOR [54] remains limited compared to the gain of our method, which shows that the effectiveness of Dyadic Interaction Modeling is as a result of our novel framework rather than additional data. In measuring motion diversity (Var, SID), DIM-LISTENER generates diverse listener reactions to the speaker. We expect that listeners' movements will showcase a greater range of diversity when the *FD* and *MSE* are kept to a minimum. As demonstrated in Table 1, it is evident that even when FD and MSE are performing well (close to ground truth), DIM-LISTENER remains highly effective in both the Var and SID, which benefits from our proposed Dyadic Interaction Modeling between speakers and listeners. The RPCC measures the motion synchronization between the speaker and the listener, and it can be found that our method achieves competitive performance.

We also retrain RLHG [80], L2L [47] and our model on LM_Listener [48] dataset. Table. 2 demonstrates the quantitative comparison. We can observe that *FD* and *P-FD* of expressions of DIM-LISTENER is about two times better than L2L [47] and three times better than RLHG [80]. We also achieve state-of-the-art
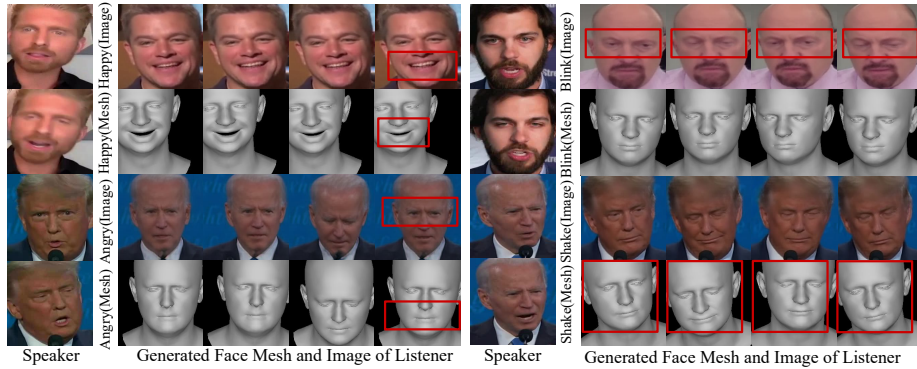
Fig. 4: Comparison with L2L [47](pre-trained on CANDOR [54]), ELP [62], and RLHG [2]. Our method can generate diverse head movements while maintaining facial expressions that better align with speakers' sentiments.

in all metrics evaluating the data distribution distances or accuracy between generation and ground-truth while maintaining the ability to generate more diverse motions.

Qualitative Results. We first compare with the photorealistic results of RLHG [80] and L2L [47] in Figure 4. We feed the EMOCA [16] 3DMM predictions and the listener appearance into PIRenderer [55] and render the corresponding images frame by frame. Since there is no guideline or source code on the renderer's settings in L2L [47], we use PIRenderer [71] official code for training and generation. We fine-tune a unique renderer for each speaker-listener pair and render the listener image sequence using the same fine-tuned weight of each pair for all approaches. Compared to RLHG [80], L2L [47], and ELP [62]: 1) our method demonstrates more expressive behaviors that synchronize with speakers' and target listeners' sentiments (first example),

Table 3: Ablation Analysis of Listener Generation on ViCo [80] and LM_Listener [48] Datasets. VQ represents the VQ-VAEs for speakers and listeners, DIM represents Dyadic Interaction Modeling on CANDOR [54], $Dec_{VQ}^{(l)}$ represents unfreezing listener's VQ decoder, $\mathcal{L}_c$ represents contrastive loss, and S-L represents the two-branch architecture that jointly models speaker and listener behaviors.

VQ	DIM	$Dec_{VQ}^{(l)}$	$\mathcal{L}_c$	S-L	ViCo MSE↓	FD↓	LM_Listener MSE↓	FD↓
✓	✓	✓	✓	✓	**0.62**	**24.15**	**0.27**	**12.94**
✗	✓	✗	✓	✓	0.92	42.52	0.66	33.07
✓	✗	✓	✓	✓	0.75	33.84	0.60	27.43
✓	✓	✗	✓	✓	0.81	39.63	0.69	31.93
✓	✓	✓	✗	✓	0.65	28.47	0.32	17.92
✓	✓	✓	✗	✗	0.72	30.92	0.47	21.27

2) our method can generate diverse head movements compared to other methods (second example). We also demonstrate our method's ability to model emotions and facial behaviors of our model in Figure 5. In particular, we show that DIM-

Fig. 5: DIM-Listener can generate diverse listener emotions (e.g. Happy and Angry) and facial behaviors (e.g. Eyes Blinking and Head Shaking).

Listener can generate multiple types of social signals such as smiles, head shaking and blinks. These suggest the benefits of Dyadic Interaction Modeling of DIM-Listener on the large-scale CANDOR [54] dataset with more diverse dyadic behaviors.

4.3   Ablation Analysis

Table 3 illustrates the contribution of each component to DIM-Listener. Initially, we note that VQ-VAE stands out as the most crucial module in our framework. Furthermore, our observation that directly predicting listener's motions as continuous signals yields less diverse motions aligns with prior research in L2L [47] and ELP [62]. As expected, Dyadic Interaction Modeling on large-scale dyadic behaviors significantly enhances performance for both ViCo and LM_Listener datasets. In contrast to the pre-training approach in L2L [47] that fixes the VQ-Decoder to decode continuous signals from discrete motion predictions, integrating the VQ-Decoder into the listener motion prediction model notably enhances performance, ranking as the second most important component in DIM-Listener. While $\mathcal{L}_c$ ranks as the least important component, it still contributes significantly to performance by encouraging role-specific encoders to extract meaningful information from input signals relevant not only to the current person but also the conversational partner. Finally, we demonstrate the benefits of jointly modeling both speaker and listener behaviors with $S - L$, which aligns with existing literature indicating that speaker and listener behaviors correlate and mutually influence each other. we also provide modality ablation as well as online listener behavior generation results in the supplementary material.

4.4   Speaker Generation
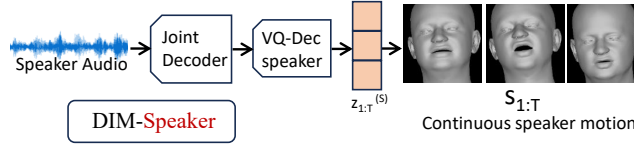
Fig. 6: For Speech-Driven Speaker Generation, we fine-tune the pre-trained weights of the joint decoder and speaker VQ-Decoder from DIM. The generated speaker's motion from DIM-SPEAKER is optimized by the same cross-entropy and reconstruction loss as DIM-LISTENER.

As mentioned in the previous sections, since we introduce a dyadic framework capturing contextualized interactions among listeners and speakers, our approach can also animate speaker motion from the speaker's speech input. We name it DIM-SPEAKER, and the fine-tuning pipeline is shown in Figure 6. For this task, we compared the results to several state-of-the-arts methods, e.g. MeshTalk [56], FaceFormer [22] and CodeTalker [74] on BiWi [23] dataset.

Table 4: Quantitative Comparison of Speech-Driven Speaker Generation on BiWi [23] Dataset.

Method	LVE↓ ($\times 10^{-4}$)	FDD↓ ($\times 10^{-4}$)
MeshTalk [56]	6.14	2.53
FaceFormer [22]	5.40	1.37
CodeTalker [74]	5.93	2.42
DIM-SPEAKER	**3.86**	**0.98**

We retrain these methods and compare the output motion parameters in the FLAME point cloud space, which consists of 5023 vertices converted from EMOCA 3DMMs. We follow the past work and adopt *Lip Vertex Error (LVE)*, which measures the discrepancy in lip movement for a generated sequence compared to the ground truth by calculating the maximum Euclidean (L2) distance error for the lip vertices in each frame, averaged over all frames, and *upper-face dynamics deviation(FDD)*, which evaluates the variation of facial dynamics for a motion sequence, as evaluation metrics.

As shown in Table. 4, we outperform previous methods in this evaluation space, which demonstrated the effectiveness of our framework for the Speaker Generation task.

5   Conclusion

In this study, we introduced Dyadic Interaction Modeling, a novel self-supervised pre-training strategy designed to improve the model's capability to effectly encode representations from both speaker and listener motions via large-scale dyadic data. Our comprehensive experiments and detailed visualizations demonstrate the capability of DIM-LISTENER and DIM-SPEAKER in generating accurate and realistic motions for both listeners and speakers.

Acknowledgement

References

1. Ahuja, C., Ma, S., Morency, L.P., Sheikh, Y.: 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. pp. 74–84 (2019)
2. Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.): Computer Vision – ECCV 2022. Springer (2022). https://doi.org/10.1007/978-3-031-19769-7
3. Bohus, D., Horvitz, E.: Facilitating multiparty dialog with gaze, gesture, and speech. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction. pp. 1–8 (2010)
4. Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K.: Real-time facial animation with image-based dynamic avatars. ACM Transactions on Graphics **35**(4) (2016)
5. Cao, Y., Tien, W.C., Faloutsos, P., Pighin, F.: Expressive speech-driven facial animation. ACM Transactions on Graphics **24**(4), 1283–1302 (2005)
6. Cerekovic, A., Aran, O., Gatica-Perez, D.: Rapport with virtual agents: What do human social cues and personality explain? IEEE Transactions on Affective Computing **8**(3), 382–395 (2016)
7. Chang, Z., Hu, W., Yang, Q., Zheng, S.: Hierarchical semantic perceptual listener head video generation: A high-performance pipeline. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 9581–9585 (2023)
8. Chen, L., Cao, C., De la Torre, F., Saragih, J., Xu, C., Sheikh, Y.: High-fidelity face tracking for ar/vr via deep lighting adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13059–13069 (2021)
9. Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., Xu, C.: Talking-head generation with rhythmic head motion. In: Proceedings of the European Conference on Computer Vision. pp. 35–51. Springer (2020)
10. Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European Conference on Computer Vision. pp. 520–535 (2018)
11. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7832–7841 (2019)
12. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)

13. Chu, H., Li, D., Fidler, S.: A face-to-face neural conversation model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7113–7121 (2018)
14. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Asian Conference on Computer Vision. pp. 251–263. Springer (2016)
15. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10101–10111 (2019)
16. Danecek, R., Black, M.J., Bolkart, T.: EMOCA: Emotion driven monocular face capture and animation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20311–20322 (2022)
17. Das, D., Biswas, S., Sinha, S., Bhowmick, B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In: European Conference on Computer Vision. pp. 408–424. Springer (2020)
18. DeVito, J.A.: Interpersonal communication book, The, 13/E. Pearson, London, UK (2013)
19. Edwards, P., Landreth, C., Fiume, E., Singh, K.: Jali: an animator-centric viseme model for expressive lip synchronization. ACM Transactions on Graphics) **35**(4), 1–11 (2016)
20. Eskimez, S.E., Zhang, Y., Duan, Z.: Speech driven talking face generation from a single image and an emotion condition. IEEE Transactions on Multimedia (2021)
21. Fan, B., Wang, L., Soong, F.K., Xie, L.: Photo-real talking head with deep bidirectional lstm. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4884–4888. IEEE (2015)
22. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18770–18780 (2022)
23. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: A 3-d audio-visual corpus of affective communication. IEEE Transactions on Multimedia **12**(6), 591–598 (2010)
24. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. ACM Transactions on Graphics, (Proc. SIGGRAPH) **40**(8) (2021), `https://doi.org/10.1145/3450626.3459936`
25. Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C., Agrawala, M.: Text-based editing of talking-head video. ACM Transactions on Graphics **38**(4), 1–14 (2019)
26. Geng, S., Teotia, R., Tendulkar, P., Menon, S., Vondrick, C.: Affective faces for goal-driven dyadic communication. arXiv preprint arXiv:2301.10939 (2023)
27. Gong, Y., Rouditchenko, A., Liu, A.H., Harwath, D., Karlinsky, L., Kuehne, H., Glass, J.: Contrastive audio-visual masked autoencoder. arXiv preprint arXiv:2210.07839 (2022)
28. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Intelligent Virtual Agents (IVA). pp. 125–138. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
29. Greenwood, D., Laycock, S., Matthews, I.: Predicting head pose in dyadic conversation. In: International Conference on Intelligent Virtual Agents. pp. 160–169. Springer (2017)
30. He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., Liu, Z.: Forgerynet: A versatile benchmark for comprehensive forgery analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4360–4369 (2021)

31. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)

32. Hussen Abdelaziz, A., Theobald, B.J., Dixon, P., Knothe, R., Apostoloff, N., Kajareker, S.: Modality dropout for improved performance-driven talking faces. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 378–386 (2020)

33. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 14080–14089 (2021)

34. Jonell, P., Kucherenko, T., Ekstedt, E., Beskow, J.: 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)

35. Jonell, P., Kucherenko, T., Henter, G.E., Beskow, J.: 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. pp. 1–8 (2020)

36. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics **36**(4), 1–12 (2017)

37. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics **37**(4), 1–14 (2018)

38. Kucherenko, T., Nagy, R., Yoon, Y., Woo, J., Nikolov, T., Tsakov, M., Henter, G.E.: The genea challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In: Proceedings of the 25th International Conference on Multimodal Interaction. pp. 792–801 (2023)

39. Lahiri, A., Kwatra, V., Frueh, C., Lewis, J., Bregler, C.: Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2755–2764 (2021)

40. Li, H., Yu, J., Ye, Y., Bregler, C.: Realtime facial animation with on-the-fly correctives. ACM Transactions on Graphics **32**(4), 42–1 (2013)

41. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021)

42. Liu, X., Wu, Q., Zhou, H., Du, Y., Wu, W., Lin, D., Liu, Z.: Audio-driven co-speech gesture video generation. arXiv preprint arXiv:2212.02350 (2022)

43. Liu, Y., Xu, F., Chai, J., Tong, X., Wang, L., Huo, Q.: Video-audio driven real-time facial animation. ACM Transactions on Graphics **34**(6), 1–10 (2015)

44. Massaro, D., Cohen, M., Tabain, M., Beskow, J., Clark, R.: Animated speech: research progress and applications. Audiovisual Speech Processing p. 309–345 (2012)

45. Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP). pp. 2227–2231. IEEE (2017)

46. Ng, E.: Learning2listen, `https://evonneng.github.io/learning2listen/`

47. Ng, E., Joo, H., Hu, L., Li, H., Darrell, T., Kanazawa, A., Ginosar, S.: Learning to listen: Modeling non-deterministic dyadic facial motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20395–20405 (2022)
48. Ng, E., Subramanian, S., Klein, D., Kanazawa, A., Darrell, T., Ginosar, S.: Can language models learn to listen? In: Proceedings of the International Conference on Computer Vision (ICCV) (2023)
49. Nojavanasghari, B., Huang, Y., Khan, S.: Interactive generative adversarial networks for facial expression generation in dyadic interactions. arXiv preprint arXiv:1801.09092 (2018)
50. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6309–6318 (2017)
51. Palmero, C., Barquero, G., Junior, J.C.J., Clapés, A., Núñez, J., Curto, D., Smeureanu, S., Selva, J., Zhang, Z., Saeteros, D., et al.: Chalearn lap challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results. In: Understanding Social Behavior in Dyadic and Small Group Interactions. pp. 4–52. PMLR (2022)
52. Pham, H.X., Wang, Y., Pavlovic, V.: End-to-end learning for 3d facial animation from speech. In: Proceedings of the ACM International Conference on Multimodal Interaction. pp. 361–365 (2018)
53. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 484–492 (2020)
54. Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., Knox, D., Liebscher, A., Marin, S.: The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. Science Advances **9**(13), eadf3197 (2023)
55. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13759–13768 (2021)
56. Richard, A., Zollhöfer, M., Wen, Y., de la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1173–1182 (2021)
57. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179 (2018)
58. Song, L., Fang, Z., Li, X., Dong, X., Jin, Z., Chen, Y., Lyu, S.: Adaptive Face Forgery Detection in Cross Domain. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 467–484. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022). `https://doi.org/10.1007/978-3-031-19830-4_27`
59. Song, L., Li, X., Fang, Z., Jin, Z., Chen, Y., Xu, C.: Face forgery detection via symmetric transformer. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4102–4111 (2022)
60. Song, L., Liu, B., Yin, G., Dong, X., Zhang, Y., Bai, J.X.: Tacr-net: Editing on deep video and voice portraits. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 478–486 (2021)
61. Song, L., Liu, B., Yu, N.: Talking face video generation with editable expression. In: Image and Graphics: 11th International Conference, ICIG 2021, Haikou, China, August 6–8, 2021, Proceedings, Part III 11. pp. 753–764. Springer (2021)

62. Song, L., Yin, G., Jin, Z., Dong, X., Xu, C.: Emotional listener portrait: Realistic listener motion simulation in conversation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20782–20792. IEEE (2023)
63. Song, L., Yin, G., Liu, B., Zhang, Y., Yu, N.: Fsft-net: face transfer video generation with few-shot views. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 3582–3586. IEEE (2021)
64. Song, S., Spitale, M., Luo, C., Barquero, G., Palmero, C., Escalera, S., Valstar, M., Baur, T., Ringeval, F., André, E., et al.: React2023: The first multiple appropriate facial reaction generation challenge. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 9620–9624 (2023)
65. Stan, S., Haque, K.I., Yumak, Z.: Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In: Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. pp. 1–11 (2023)
66. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics **36**(4), 1–13 (2017)
67. Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J., Matthews, I.: A deep learning approach for generalized speech animation. ACM Transactions on Graphics **36**(4), 1–11 (2017)
68. Taylor, S.L., Mahler, M., Theobald, B.J., Matthews, I.: Dynamic units of visual speech. In: Proceedings of the ACM SIGGRAPH/Eurographics conference on Computer Animation. pp. 275–284 (2012)
69. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: Proceedings of European Conference on Computer Vision. pp. 716–731. Springer (2020)
70. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. International Journal of Computer Vision **128**(5), 1398–1413 (2020)
71. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. arXiv preprint arXiv:1808.06601 (2018)
72. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. ACM Transactions on Graphics **30**(4), 1–10 (2011)
73. Woo, J., Fares, M., Pelachaud, C., Achard, C.: Amii: Adaptive multimodal interpersonal and intra-personal model for adapted behavior synthesis. arXiv preprint arXiv:2305.11310 (2023)
74. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12780–12790 (2023)
75. Xu, Y., Feng, A.W., Marsella, S., Shapiro, A.: A practical and configurable lip sync method for games. In: Proceedings of Motion on Games. pp. 131–140 (2013)
76. Yi, R., Ye, Z., Zhang, J., Bao, H., Liu, Y.J.: Audio-driven talking face video generation with learning-based personalized head pose. arXiv preprint arXiv:2002.10137 (2020)
77. Yu, J., Chen, C.W.: From talking head to singing head: a significant enhancement for more natural human computer interaction. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 511–516. IEEE (2017)
78. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021)
79. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceed-

ings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4176–4186 (2021)

80. Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T., Mei, T.: Responsive listening head generation: a benchmark dataset and baseline. In: European Conference on Computer Vision. pp. 124–142. Springer (2022)

81. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions on Graphics (TOG) **39**(6), 1–15 (2020)

82. Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., Singh, K.: Visemenet: Audio-driven animator-centric speech animation. ACM Transactions on Graphics **37**(4), 1–10 (2018)

83. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. In: Computer Graphics Forum. pp. 523–550 (2018)