Talking Together: Synthesizing Co-Located 3D Conversations from Audio

Mengyi Shan[1], Shouchieh Chang[2], Ziqian Bai[2], Shichen Liu[2], Yinda Zhang[2], Luchuan Song[3], Rohit Pandey[2], Sean Fanello[2], Zeng Huang[2]

[1]University of Washington, [2]Google, [3]University of Rochester

Abstract

We tackle the challenging task of generating complete 3D facial animations for two interacting, co-located participants from a mixed audio stream. While existing methods often produce disembodied "talking heads" akin to a video conference call, our work is the first to explicitly model the dynamic 3D spatial relationship, including relative position, orientation, and mutual gaze, that is crucial for realistic in-person dialogues. Our system synthesizes the full performance of both individuals, including precise lip-sync, and uniquely allows their relative head poses to be controlled via textual descriptions. To achieve this, we propose a dual-stream architecture where each stream is responsible for one participant's output. We employ speaker's role embeddings and inter-speaker cross-attention mechanisms are designed to disentangle the mixed audio and model the interaction. Furthermore, we introduce a novel eye gaze loss to promote natural, mutual eye contact. To power our data-hungry approach, we introduce a novel pipeline to curate a large-scale conversational dataset consisting of over 2 million dyadic pairs from in-the-wild videos. Our method generates fluid, controllable, and spatially aware dyadic animations suitable for immersive applications in VR and telepresence, significantly outperforming existing baselines in perceived realism and interaction coherence.

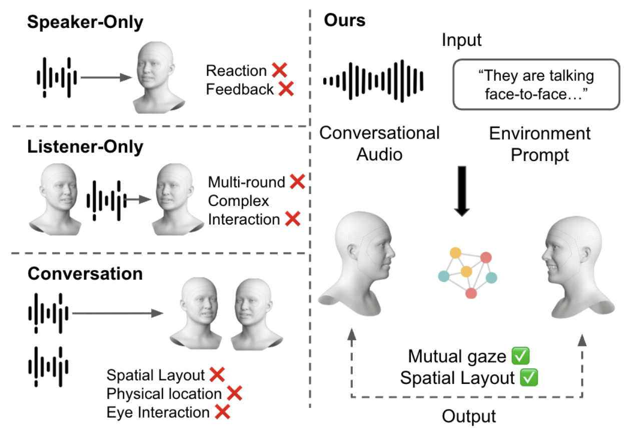

Figure 1. Our method in context. Left: Prior work generates isolated participants, such as Speaker-Only models that lack listener reactions, Listener-Only models that do not model multi-round interactions, or Conversation models that resemble a "video conference" call, failing to model a shared 3D space, spatial layout, or real, physically meaningful eye interaction. Right: Our method takes a single mixed conversational audio stream and a text-based environment prompt to generate the complete, co-located 3D performance of *both* participants. Our model is the first to explicitly synthesize this crucial spatial relationship, enabling realistic outputs with natural mutual gaze and a controllable spatial layout.

1. Introduction

Photorealistic digital humans [7, 13, 28, 29, 42, 48, 60, 67, 75] are foundational to immersive communication, poised to revolutionize applications ranging from virtual reality and telepresence to advanced human-computer interfaces [5, 24, 45, 72, 73]. At the heart of believable human interaction is dyadic conversation, a complex and subtle dance of verbal and non-verbal signals exchanged between two individuals in a shared space. Current audio-driven 3D animation research either focuses on speaker-only generation [12, 27, 46, 47, 63, 64, 69], or overlooks the physical co-location of participants, predominantly generating isolated "talking heads" that resemble a video conference call rather than an in-person interaction [31, 40, 56, 66]. This paradigm fails to capture the crucial non-verbal cues derived from a shared 3D space, such as responsive head movements and mutual eye gaze that signal turn-taking and engagement. These methods ignore the spatial relationship between speakers and produce disconnected animations that lack the organic flow of a genuine face-to-face dialogue, presenting a barrier to truly immersive social interaction.

We present a method for generating complete 3D facial animations of two co-located participants, including their dynamic spatial relationship, from a single mixed audio stream as shown in Fig. 1. A significant barrier to progress in this domain is the scarcity of large-scale, high-fidelity 3D training data that captures natural dyadic conversations.

We address this challenge by effectively leveraging a combined corpus: a massive collection of over 2 million interacting pairs for learning general conversational dynamics, and a separate, high quality collection of a single-person talking head. The latter, sourced from segments with superior resolution and camera angles, is critical. It allows us to train a specialized lip-sync expert model, essential for remedying the inherent lip-sync ambiguities and artifacts often found in in-the-wild conversational data, thereby enabling our subsequent two-stage training strategy.

Leveraging these curated datasets, we introduce a novel generative model designed specifically for dyadic interaction. The method is centered on a dual-stream architecture, where each stream synthesizes the 3D facial performance for one participant. To effectively disentangle the single mixed audio input and model turn-taking dynamics, we employ cross-attention mechanisms and speaker role embeddings. This allows our model to learn not just an individual's speech patterns, but also the reactive, non-verbal behavior of the listener. This architecture is trained in a two-stage process that utilizes both of our datasets: we first pre-train the model's shared backbone on our diverse multi-person data to ensure expressive interactive behaviors, and then fine-tune the entire interactive model on the smaller scale high-quality mixed data to improve lip-sync.

Furthermore, we introduce two novel mechanisms to explicitly model the spatial dynamics of the conversation. First, to promote realistic mutual eye contact, we apply a targeted auxiliary eye gaze loss. We curate our dataset by filtering for a high-quality subset of interactions with clear and meaningful eye gaze, and this loss is applied exclusively when training on these selected samples. Second, we integrate a text-based control mechanism for the participants' relative positioning. During training, the model is conditioned on the ground-truth global translation of both heads in the first frame. At inference, this becomes a powerful control interface: a user can provide a natural language prompt (e.g., "an intimate conversation" or "arguing across a table"), and a Large Language Model (LLM) predicts the initial 3D translations. This prediction is achieved in a few-shot manner, where the LLM is provided with several examples of text-to-translation pairs in its prompt.

Our key contributions can be summarized as follows:

- We introduce an automated pipeline to curate a large-scale dataset of dyadic conversations from in-the-wild videos, along with a high-fidelity single-speaker corpus for robust lip-sync training.
- We propose a dual-stream diffusion architecture with a shared U-Net backbone, cross-attention, and FiLM conditioning to effectively model speaker interaction and disentangle a single mixed audio track.
- We employ a mixed-data training strategy, pre-training on massive real conversational data and then fine-tuning on high-quality, synthetic conversation data to ensure both precise lip articulation and natural, interactive behaviors.
- We enable intuitive scene control through a few-shot, LLM-based text-to-3D spatial translation mechanism and promote realistic eye contact with a targeted auxiliary gaze loss on a curated data subset.

2. Related Works

Generating Audio-based Talking Head. Audio-driven talking head generation has emerged as a prominent area of research in recent years. One major line of work focuses on predicting 3D parametric face models [6, 34], which can be converted into mesh vertices for downstream applications such as virtual or augmented reality [1, 2, 12, 18, 19, 27, 32, 50, 52, 57, 59, 64, 71]. In parallel, another line directly synthesizes pixel-level outputs by animating a static portrait according to input audio, enabling more flexible and expressive media generation [10, 36, 58, 68, 70]. Recent advancements, such as FaceFormer [20], CodeTalker [69], and SelfTalk [46], introduced geometry-based methods using facial mesh representations to enhance realism in 3D talking heads. UniTalker [17] improved generalization by training across multiple datasets and fine-tuning with minimal data, while ScanTalk [43] enabled 3D face animation with any topology, thus broadening application scenarios. Our approach builds upon 3D-based generation to capture structural consistency, while further incorporating relative spatial information to enable co-located rendering.

Generating Spatial-Aware Group Interactions. Recent research has progressed from generating isolated individuals to modeling multi-person interactions that respect spatial and social context. Early work explored multi-agent dynamics and collision-aware motion forecasting [38, 62], while more recent methods generate coordinated and semantically grounded interactions driven by text or motion cues [16, 23, 53, 54, 61]. Several approaches further incorporate scene and contact awareness, synthesizing physically plausible movements conditioned on 3D environments or human–object interactions [8, 14, 35]. Others model the communicative aspects of group behavior, generating expressive co-speech gestures and reactions within multi-person conversations [37]. Despite these advances, most existing frameworks focus on body dynamics and proxemics, with limited attention to integrating high-fidelity facial expressions and nuanced interpersonal cues, which is a key challenge for expressive, spatially aware group interaction generation.

Generating Conversation In human conversations, listeners convey vital non-verbal cues through facial expressions, head nods, and eye movements, which are crucial for natural and engaging interactions. Early works modeled such listener behaviors with neural networks [26, 30,

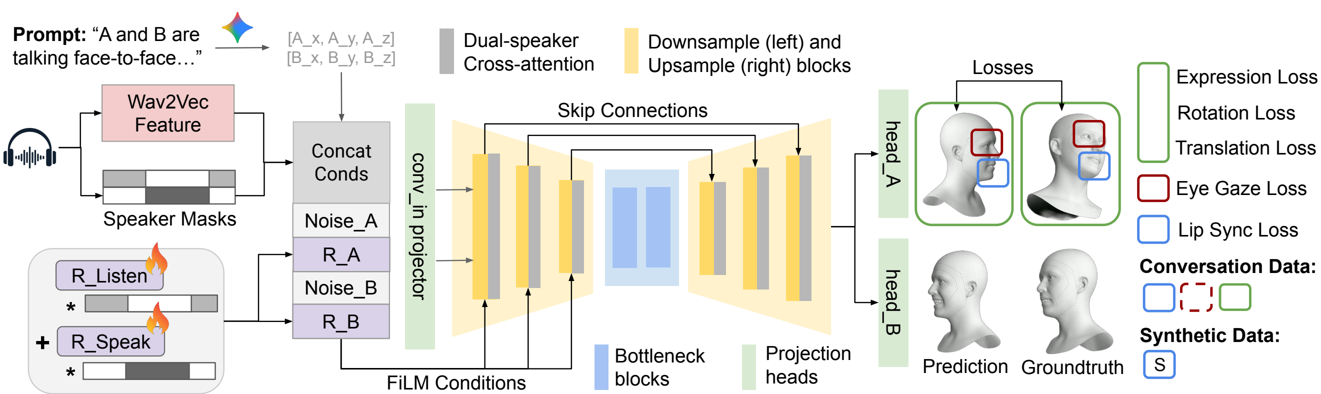

Figure 2. An overview of our dual-stream diffusion architecture. The model employs a shared U-Net backbone to process the noisy input streams for both participants in parallel. It is conditioned on features from the mixed audio, learnable role embeddings, and the speaker probability masks. Dual-speaker cross-attention layers within the decoder allow the two streams to exchange information, modeling the interaction dynamics. The network outputs the predicted 3D animation parameters for expression ($\boldsymbol{\psi}$), rotation ($\boldsymbol{\theta}$)), and translation ($\boldsymbol{t}$). For conversation data, we utilize losses for all expression, rotation and translation prediction, with auxiliary eye gaze loss on samples with large-scale head movements. For synthetic data, we finetune with losses only on speaker's lip parameters.

31, 33, 40, 41, 55, 56, 66, 74]. Zhou *et al.* [74] introduced the task of responsive listener generation with a Role Switcher alternating between speaker and listener roles, while Learning2Listen [39] produced brief reactions like nods and smiles from speech. In the video domain, INFP [76] learns conversational dynamics without explicit role assignments, and ARIG [25] employs a long-range autoregressive model for flexible interactions. Closely related works include Audio2Photoreal [41], which generates photorealistic avatars from audio but lacks behavioral reactivity, and DualTalk [49], which models dual-speaker conversations yet remains limited to "video conferencing"-style talking heads without spatially aware co-located movements. In contrast, our system addresses these gaps by jointly learning implicit speaker–listener roles, for fine-grained mutual interaction and by being the first to synthesize all the participants within a shared, dynamic 3D spatial content.

3. Method

Our full pipeline is illustrated in Fig. 2. We discuss our preliminaries, define the problem, and explain our model architecture as well as training paradigm below.

3.1. Preliminary

Diffusion Model Objectives. We employ a denoising diffusion model that learns to reverse a fixed Gaussian noise process. Our model G, is trained to predict the original clean data x_0 from a noised input x_t at timestep t, conditioned on a set of signals c. The objective is a simple L2 loss between the ground truth and the model's prediction:

$$\mathcal{L} = E_{x_0,t,x_t}\left[\|x_0 - G\left(x_t, t, c\right)\|_2^2\right]$$

To enable classifier-free guidance during inference, we randomly mask the audio portion of the condition c for 10% of the training samples.

Face Motion Representation. Our work utilizes a 3DMM-based parametric face model, where a 3D head mesh is defined by identity ($\boldsymbol{\beta}$), expression ($\boldsymbol{\psi}$), pose ($\boldsymbol{\theta}$), and translation ($\mathbf{t}$) parameters. The final facial geometry, $M(\boldsymbol{\beta}, \boldsymbol{\psi})$, is generated by adding a linear combination of identity and expression bases to a mean shape template ($\mathbf{S}$):

$$M(\boldsymbol{\beta}, \boldsymbol{\psi}) = \mathbf{S} + \sum_i \beta_i \mathbf{S}_i + \sum_j \psi_j \mathbf{E}_j, \tag{1}$$

where $\mathbf{S}_i$ and $\mathbf{E}_j$ are the principal components derived from a large dataset of high-quality 3D scans. This mesh is then articulated by pose and translation parameters using a standard Linear Blend Skinning rig to get animated geometry.

3.2. Problem Definition

Our primary objective is to synthesize a realistic 3D animation of a dyadic conversation from a single audio source. We formulate this as a sequence-to-sequence task where the input is a mixed audio waveform, $\mathbf{A} \in \mathbb{R}^T$, containing the speech of two participants. For each participant, the output is the concatenation of three sequences: a sequence of facial expression vectors, $\boldsymbol{\psi} \in \mathbb{R}^{L\times 63}$; a sequence of skeletal poses, $\boldsymbol{\theta} \in \mathbb{R}^{L\times 4\times 3}$, representing the orientation of four key skeletal joints (neck, head, left and right eyes); and a sequence of head translations, $\mathbf{t} \in \mathbb{R}^{L\times 3}$, representing the global position in metric meters. The desired model output is a synchronized sequence of animation parameters of length L for both participants, $\mathbf{x} = \{\mathbf{x_A}, \mathbf{x_B}\}$ where each $\mathbf{x} = \text{concat}(\psi, \theta, \mathbf{t}) \in \mathbb{R}^{L\times 78}$.

3.3. Modeling Dyadic Interaction

To model dyadic interaction, we propose a conditioned, dual-stream cross-attention diffusion architecture. We explain the modeling details below.

Audio Speaker Masking. Given the mixed waveform, we predict two speaking probability masks $\mathbf{m}_A, \mathbf{m}_B \in [0,1]^{T\times 1}$ corresponding to the speaking probability of each participant. These masks provide temporal guidance for disentangling speech and non-verbal behavior, and are used as conditioning inputs. Notably, the masks are not strictly accurate; this mild imperfection introduces beneficial noise that stabilizes training and improves robustness to real conversational overlap. At training time, we pre-computed and saved these masks instead of computing during the training loop (explained in Sec. 3.6). At inference time, we apply the audio separation and voice activity detection modules to estimates per-frame speaker activity on-the-fly.

Shared Dual-Stream Architecture. We adopt a dual-stream design with shared weights across participants. A single U-Net backbone processes the noisy inputs $\mathbf{x_{t,A}}$ and $\mathbf{x_{t,B}}$ in parallel, promoting a unified, speaker-agnostic representation of facial motion. The shared generator $\mathcal{G}$ predicts denoised outputs for both streams as:

$$\hat{\mathbf{x}}_{0,A},\, \hat{\mathbf{x}}_{0,B} = \mathcal{G}(\mathbf{x}_{t,A},\, \mathbf{x}_{t,B},\, t,\, \mathbf{m}_A,\, \mathbf{m}_B)\,, \tag{2}$$

This formulation ensures coherent evolution of both outputs while enabling downstream cross-attention layers to capture speaker–listener interactions.

Interaction Modeling via Cross-Attention. To capture mutual influence between participants, we insert cross-attention layers within the shared U-Net decoder, enabling bidirectional information exchange. Let $\mathbf{h}_A$ and $\mathbf{h}_B$ denote the intermediate feature maps for the two streams. At each layer, the features of one participant are updated by attending to the other:

$$\mathbf{h}'_A = \text{Attention}(\mathbf{Q}_A,\, \mathbf{K}_B,\, \mathbf{V}_B)\,, \tag{3}$$

$$\mathbf{h}'_B = \text{Attention}(\mathbf{Q}_B,\, \mathbf{K}_A,\, \mathbf{V}_A)\,, \tag{4}$$

where $\mathbf{Q}_A = \mathbf{W}_Q\mathbf{h}_A$, $\mathbf{K}_B = \mathbf{W}_K\mathbf{h}_B$, and $\mathbf{V}_B = \mathbf{W}_V\mathbf{h}_B$. This symmetric exchange encourages each stream to model reactive, conversational behaviors while preserving speaker-specific representations.

Speaker Role Embedding. To encode conversational intent, we introduce two learnable embedding vectors, $\mathbf{e}_{\text{speak}}$ and $\mathbf{e}_{\text{listen}}$, which are shared by both participants. At each animation timestep k, a dynamic role vector is calculated for each person by linearly interpolating these embeddings. The weighting $\mathbf{m}^{(k)}$ is the scalar speaker probability for that person at that frame, derived from their full speaker mask vector (e.g., $\mathbf{m}_A$).

$$\mathbf{e}^{(k)}_{\text{role}} = \mathbf{m}^{(k)}\mathbf{e}_{\text{speak}} + (1-\mathbf{m}^{(k)})\mathbf{e}_{\text{listen}}. \tag{5}$$

This resulting vector provides a continuous representation of a participant's interaction state. These role embeddings for both participants serve as key components of the model's overall conditioning signal, defined next.

Dynamic Dual Conditioning. We condition the shared U-Net on a comprehensive set of multimodal cues capturing the conversational state and participant identity. These cues are consolidated into a single conditioning vector $\mathbf{c}^{(k)}$, which is dependent on the animation timestep k. This vector concatenates the Wav2Vec [4] audio features $\mathbf{a}^{(k)}$, the dynamic role embeddings for both participants $\mathbf{e}^{(k)}_{\text{role},A}$ and $\mathbf{e}^{(k)}_{\text{role},B}$, and the speaker probability masks $\mathbf{m}^{(k)}_A$ and $\mathbf{m}^{(k)}_B$. This vector $\mathbf{c}^{(k)}$ guides generation through two paths: (1) it is concatenated with the noisy latent input $\mathbf{x}$ before the diffusion U-Net, injecting global multimodal context, and (2) it modulates intermediate features $\mathbf{h}$ via FiLM:

$$\text{FiLM}(\mathbf{h}, \mathbf{c}^{(k)}) = (\boldsymbol{\gamma}(\mathbf{c}^{(k)}) + 1) \odot \mathbf{h} + \boldsymbol{\beta}(\mathbf{c}^{(k)}),$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are learned functions that adaptively shift and rescale activations. These mechanisms jointly enable frame-dependent control over "speaking" and "listening" behaviors while preserving visual identity and coherence.

3.4. Multi-Stage Training

Conversational datasets exhibit poor lip synchronization due to low resolution and unavoidable occlusions as people might be facing each other in nature, and thus cannot be captured perfectly by the camera. To achieve accurate lip motion while retaining interactive dynamics, we adopt a two-stage training strategy with mixed datasets.

Stage 1: Pre-training on Interaction Data. We first pretrain the dual-stream model on large-scale conversational data to establish general audio–visual alignment and interaction understanding. This stage captures speaker–listener dynamics across diverse settings, learning rotation .

Stage 2: Fine-tuning with High-Quality Lip Data. Next, we fine-tune the model using a combination of (1) high-resolution single-speaker datasets with precise lip motion and (2) an augmented conversation data subset with high landmark confidence after super-resolution. For (1), we apply the L2 reconstruction loss only to the 20 parameters of the expression vector ψ corresponding to lip and jaw articulation for the speaking participant, and set all other losses (e.g., rotation, translation, and non-lip expression) for both participants to zero. This staged process leverages large-scale conversational diversity for interaction grounding while preserving accurate lip synchronization through targeted fine-tuning on clean, high-quality visual data.

3.5. Controllable Spatial Relationship

To ground the conversation in 3D space and enable plausible interactions like eye contact, we explicitly condition the model on the participants' spatial layout.

Training with Relative Translation. The model is trained to predict motion relative to a center-normalized starting position. At training time, we provide the ground-truth first-frame translations for both speakers, $\mathbf{t}_A^{(0)}$ and $\mathbf{t}_B^{(0)}$, as an additional conditioning signal. This static spatial layout is concatenated into the main conditioning vector $\mathbf{c}^{(k)}$, providing constant spatial context across all animation timesteps k. The model's prediction target for the translation component of $\mathbf{x}_0$ is then the *delta* from this first frame: $\Delta\mathbf{t}^{(k)} = \mathbf{t}^{(k)} - \mathbf{t}^{(0)}$. This normalizes the problem, forcing the model to learn the movements including translation changes, head poses and eye rotations given absolute first frame position.

Inference-Time Control via LLM. This conditioning becomes a powerful control mechanism at inference. We leverage a Large Language Model (LLM) [51] to enable intuitive, text-based scene control. We employ a few-shot prompting strategy, where the LLM is provided with several human-annotated in-context examples of text-to-translation pairs (e.g., mapping a description like "Standing side-by-side" to a set of 3D coordinates for $t_A^{(0)}$ and $t_B^{(0)}$). This conditions the model to generate a structured coordinate set for a new, user-provided prompt. The full prompt structure and examples are detailed in the supplementary material.

Auxiliary Eye Gaze Loss. To promote realistic gaze behavior, our loss function computes the cosine similarity between the predicted gaze direction vector $\mathbf{g}_{pred}$ and the ground-truth gaze vector $\mathbf{g}_{gt}$. Specifically, we first convert the eye rotation parameters for each eye into a 3D gaze-forward vector. The final gaze direction vector, g, is then computed as the average of the left and right eye directions. This allows the model to learn the full spectrum of natural gaze, including both eye contact and aversion, by mimicking the ground truth. Crucially, we apply a higher loss weight *selectively* to conversational samples that exhibit large-scale head movements (top 20% head rotation variance of the dataset), which intuitively should exhibit more meaningful eye contact behaviors to learn from.

3.6. Dataset Curation

A key challenge for training our model is the lack of a large-scale dataset containing 3D motion of co-located, interacting individuals. To address this, we curate two large-scale datasets with complementary strengths: a massive corpus of in-the-wild dyadic conversations for learning natural interactions, and a novel synthetic dubbing dataset with perfect ground truth for speaker activity and lip motion. Our full data generation pipeline is visualized in Fig. 3. We additionally compare our datasets with existing open-sourced datasets in Tab. 1, and visualize distribution characteristics of our conversational dataset in Fig. 4

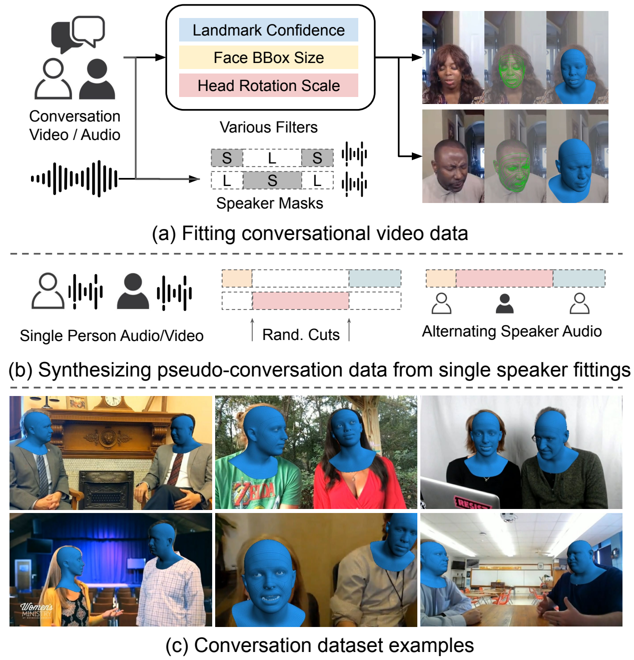

Figure 3. Our Data Curation Pipeline. This figure illustrates our two-pronged approach to dataset creation. (a) Dyadic Conversation Dataset: We process raw conversational videos through various filters to reconstruct 3D facial parameters and extract speaker masks. (b) Synthetic Dubbing Dataset: We generate clean pseudo-conversations by taking single-person videos, applying random cuts, and re-assembling them into new alternating-speaker audio tracks. (c) Dataset Example: A sample of 3D reconstructions overlaid on videos from our curated dyadic dataset.

3.6.1 Dyadic Conversation Dataset

This dataset captures complex non-verbal dynamics and spatial relationships of real-world interactions. Our data begins with a vast corpus of online conversational videos containing two people. Refer to Fig. 3 (a) for the steps, and (c) for results overlaid on original videos.

Scenario Filtering. We first automatically filter out scenarios where participants are not in the same physical space (e.g., video conference calls). We accomplished this by inspecting the background color distribution of left and right halves of the video, discarding videos that contain two separated scenes next to the middle seam.

Quality Control Filtering. We aggressively filter the dataset by discarding videos where multiple frames contain heavily occluded, overly small or blurred faces. We take advantage of both the face bounding box sizes and the confidence values while predicting facial landmarks.

Audio Source Separation and Masking. We use Looking to Listen [15] as our source separation model to isolate each speaker's audio. We then incorporate voice activity detection based on WebRTC [44], which generates binary speaker probability masks for each frame index.

Table 1. Comparison of 3D talking head datasets. Existing datasets differ in scale, interaction diversity, and lip accuracy. Our Dyadic Conversation and Synthetic Dubbing datasets combine large-scale interactive scenes with accurate lip motion and identity consistency, enabling joint learning of interaction and high-fidelity speech animation.

Datasets	Duration	Identities	Interaction	Multi-Round Conversations	Spatial Relationship	Simultaneous Talking
VOCASET [11]	0.5h	12	✗	✗	✗	✗
BIWI [21]	1.44h	14	✗	✗	✗	✗
ViCO [74]	1.6h	92	✓	✗	✗	✗
L2L [39]	72h	6	✓	✗	✗	✗
Lm_listener [40]	7h	4	✓	✗	✗	✗
RealTalk [22]	8h	-	✓	✗	✗	✗
DualTalk [49]	50h	1000+	✓	✓	✗	✗
Our Conversation	**50,000+h**	**10k+**	✓	✓	✓	✓
Our Synthetic	**50,000+h**	**10k+**	✗	✗	✗	✓

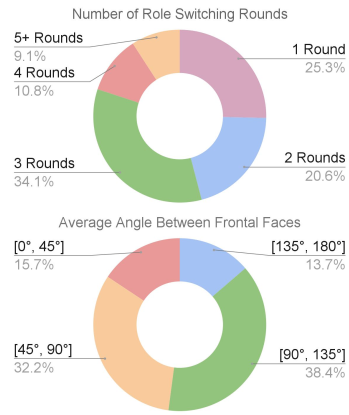

Figure 4. Distribution visualization of the conversational dataset.

Video Quality Enhancement. We detect face bounding boxes and apply a face-specific super-resolution network similar to GLEAN [9] to specified face patches in the video. This enhances the clarity and detail of facial features.

3D Face Reconstruction. For each valid frame, we apply a robust 3D face reconstruction method to fit our parametric model to every detected face. This step yields per-frame expression parameters as well as the full 6DoF head pose, including global translation and orientation with respect to a fixed camera coordinate system for both faces. Absolute scales are estimated by assuming an average inter-ocular distance, allowing us to recover translation in approximate metric meters. We include a temporal term in the optimization process to ensure smooth transitioning between frames.

3.6.2 Synthetic Dubbing Dataset

While our dyadic dataset captures natural interactions, it presents two key limitations for high-fidelity training. First, it suffers from frequent occlusions; as participants naturally face each other, a single camera often fails to capture clear lip movements from all angles. Second, the speaker masks are imperfect, particularly in resolving ambiguous, overlapping speech. To generate a clean dataset with perfect ground-truth supervision, we introduce a synthetic dubbing pipeline. This pipeline uses a large corpus of high-quality, frontal-facing single-person videos, sourced from YouTube, which has highly accurate lip motion.

We randomly sample and cut speech segments from different videos, temporally aligning them to create synthetic two-person dialogues. By selectively muting one speaker while the other speaks, we simulate realistic conversational turn-taking. This "dubbing" process simulates complex scenarios, including simultaneous speech, with precision. Crucially, because the source for each speaker is isolated, we obtain perfect ground-truth speaker activity masks and inherit high-fidelity lip motion from the clean source data. This dataset is invaluable for training the model to handle overlapping audio and perform precise lip-sync. Fig. 3 (b) shows the process of synthesizing pseudo-conversation data from single speaker fittings.

4. Experiments

We present a thorough and detailed evaluation of our models below, including explanation of baselines and metrics.

4.1. Implementation details

Our model is implemented in JAX. Audio is sampled at 16 kHz and converted to 768-dimensional Wav2Vec 2.0 features [3]. Training is performed on 16 A100 GPUs for 200,000 steps using AdamW, a cosine noise schedule, batch size $B = 1024$, and learning rate $l = 1\text{e}{-4}$. At inference time we use a classifier-free guidance weight of 2.5. The total loss combines diffusion reconstruction losses on expression ($\lambda_{expr} = 1$), rotation ($\lambda_{rot} = 8$), and translation ($\lambda_{trans} = 1$), with auxiliary regularization terms for vertex velocity ($\lambda_{vel} = 1$) and eye gaze ($\lambda_{gaze} = 5$). Audios are 10 seconds each, and all sequences are 25fps ($L = 250$). Evaluation are performed on a test set of 2048 audio-face pairs not seen during training. Users can use Gemini [51] to predict initial 3D translations from text. For our quantitative results, we use the groundtruth translation as our condition to allow for per-vertex comparison with the test data. See Appendix for architecture and training details.

4.2. Baselines

We compare our approach with three categories of baselines to evaluate different aspects of conversational modeling.

Single-Speaker Models. We first compare with single-person talking head models that take clean audio as input. To apply them to conversational audio, we perform audio segmentation to obtain two separated audio tracks, which are then fed independently into each single-person

Table 2. Quantitative comparison of our method against baselines. We only compare translation results with retrieval-based methods, as other baselines inherently do not predict translations. Metrics: FD/P-FD (on renders) measure realism/interaction. MSE (on parameters) and vMSE (on vertices) measure accuracy for FULL motion, EXPression, ROTation, TRANSLation, EYE gaze, and LIP motion. SID measures diversity for LIStener and SPEaker roles. Lower is better (↓) for FD, P-FD, MSE, and vMSE; higher is better (↑) for SID.

Methods	FD ↓	P-FD ↓	MSE ↓					vMSE ↓		SID ↑			
	FULL	FULL	EXP	TRANSL	ROT	EYE	LIP	SPE	LIS	FULL	EXP	ROT	TRANSL
CodeTalker [69]	47.23	70.54	10.47	-	14.28	3.07	2.95	12.49	6.85	0	0	0	-
SelfTalk [46]	43.58	53.98	8.21	-	11.59	2.47	2.41	10.98	6.13	1.68	1.27	1.39	-
FaceFormer [19]	52.66	59.84	13.89	-	12.34	2.96	2.84	10.47	6.44	1.59	0.43	0.86	-
Ours (Single)	19.58	29.03	6.32	-	6.74	1.23	1.14	**6.86**	5.23	2.23	1.40	1.61	-
Listen-R	63.74	68.75	11.03	8.90	10.93	2.58	2.27	7.35	7.98	1.84	2.39	2.18	**2.98**
Listen-A	65.09	41.57	12.57	7.67	7.98	2.32	1.91	7.68	4.89	1.29	2.26	2.57	1.96
Listen-M	33.40	29.06	9.42	9.06	9.81	2.79	2.12	7.01	4.63	1.03	2.47	**2.86**	2.08
DIM [65]	55.09	45.20	14.56	-	10.67	2.96	2.35	11.79	6.21	0.73	1.84	1.31	-
DualTalk [49]	28.41	38.29	9.91	-	8.42	2.11	2.50	8.32	6.88	1.57	1.95	1.79	-
L2L [39]	38.92	66.13	11.32	-	10.15	2.35	2.94	11.21	5.71	1.78	1.58	1.12	-
Ours	**10.43**	**18.24**	**4.03**	**2.09**	**3.50**	**0.98**	**0.35**	7.99	**2.29**	**2.28**	**2.48**	1.97	2.45

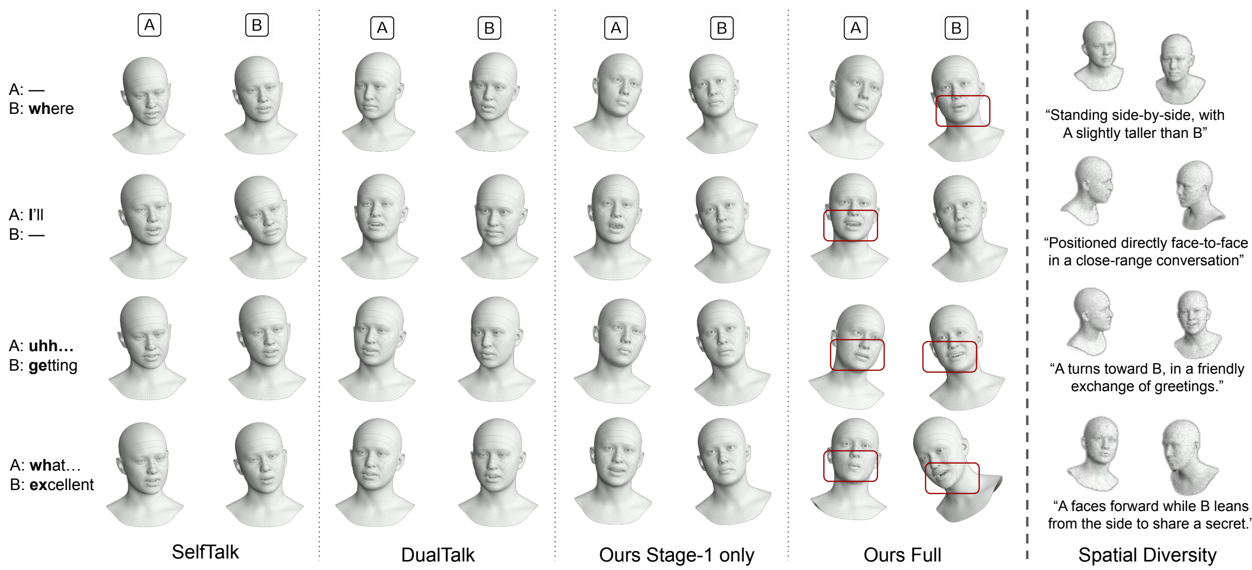

Figure 5. Left: Result of our model compared with baselines, shown as a few representative frames from one sequence. Top two rows feature speaker and listener interaction; Bottom two rows feature our model's ability to deal with short simultaneous speaking sequences. Right: Run inference with a variety of spatial layout prompts and render the first frame in a shared space. Red squares are used to mark significant speaker lip movements. This could be significantly easier to view as videos in supplementary.

model. These methods produce realistic speech-driven motion but lack any listener reaction or inter-speaker dependency. We include three such baselines with code available, CodeTalker [69], FaceFormer [19] and SelfTalk [46], and a single-person model trained on our single-person data.

Speaker + Listener Model. We further evaluate baselines that combine a single-speaker motion generation model (for the active speaker) with separate listener motion models. We test our model on DIM [65] and DualTalk [49]. We additionally compare with one random listener motion baseline (Listen-R) and two retrieval-based baselines that searches for a nearest-neighbor listener sequences from the training set based on either audio or speaker motion similarity (Listen-A, Listen-M). These comparisons assess how effectively each approach captures natural listener feedback without simultaneous modeling of interactions.

Two-Speaker Interaction Models. Finally, we compare with L2L [39], which still doesn't generate both speakers simultaneously, but trains separate models for a specific identity that can generate either speaking or listening behaviors.

4.3. Quantitative Evaluation

We employ a similar set of metrics as DualTalk [49]. We omit the exact definition here and leave them to Appendix.

Baselines. Quantitative comparison with baselines are shown in Table 2. Compared with all baselines, our method

Table 3. Ablation study on different model components. Lower FD, P-FD, and MSE (EXPression, TRANslation or ROTation) indicate better performance when evaluated with groundtruth.

Ablation Setting	FD ↓	P-FD ↓	EXP ↓	TRAN ↓	ROT ↓
Single-Person Only	50.45	50.05	10.01	2.32	3.88
w/o 2nd Stage	60.12	64.44	7.73	1.71	2.94
w/o Speaker Emb.	35.92	35.93	7.18	1.80	2.87
w/o Cross-Attn	30.49	40.87	6.87	1.54	2.98
w/o Gaze Loss	37.46	42.90	7.33	2.59	2.77
Full Model (Ours)	**21.71**	**22.56**	**5.97**	**1.50**	**2.48**

achieves the best performance across nearly all metrics on test sets, confirming the effectiveness of our system in capturing expressive, speaker-specific motion. Improvements in expression and pose MSE further demonstrate our cross-attention mechanism's performance in coordinating head motion between speakers. Our single person baseline slightly outperforms the conversation model on speaker vertex prediction, but worse in listener and interaction modeling. Retrieval-based methods occasionally generate more diverse rotations and translations but with low fidelity.

Ablations. We conduct ablations to assess each component's impact (Tab. 3). Training solely on single-person data performs poorly, as it provides no guidance for listener behavior. Conversely, training only on conversation data degrades facial expression accuracy, confirming this data source has high-quality interactions but low-quality expressions. Omitting the speaker-role embedding yields unnatural listener responses. Removing cross-attention fragments turn-taking and increases FD/P-FD, confirming its role in modeling interaction. Finally, excluding the eye-gaze loss reduces mutual attention realism and increases both P-FD and expression MSE. These results demonstrate that all of our datasets, training stages, and architecture jointly enhance quantitative accuracy and perceptual quality.

4.4. Qualitative Evaluation

We prepare extensive qualitative results showing our model's performance comparing with baselines (single speaker and speaker + listener) and alternative designs (without the second stage). We refer the readers to the videos in the supplementary files for illustration. In Fig. 5, we present generated frames with a single input audio .

Conversation Generation. Top two rows in Fig 5 visualize complete conversational sequences generated from mixed audio. Our method produces smooth, coherent facial motion and clear turn-taking behavior. Bottom two rows in Fig 5 show that our model disentangles overlapping speech and generates plausible mouth motion for both speakers. When both participants speak concurrently, each stream maintains accurate lip articulation and independent head motion while preserving the conversational context.

Table 4. Human evaluation preference rate (%). Higher is better. Participants selected the best model from four options.

Method	Lip	Speaker M	Listener M	Interact	Gaze
SelfTalk	0.9	0.9	1.6	1.6	2.4
DualTalk	3.9	6.3	7.2	5.6	7.9
Ours (S1)	15.9	19.0	18.2	21.4	21.4
Ours	**79.3**	**73.8**	**73.0**	**71.4**	**68.3**

LLM-guided Spatial Relationship Control. Right column in Fig. 5 demonstrates the controllability of relative positioning through text prompts. Our system composes 3D layouts while maintaining realistic head translation and gaze behavior. Our model also consistently produces mutual eye contact and context-aware gaze shifts. While text-to-translation is not the focus of our evaluation, we provide a full user study on its effectiveness in the Appendix.

4.5. Human Evaluation

Automatic metrics cannot fully capture the naturalness and mutual responsiveness of co-located conversational animation, so we performed a forced-choice user study. Nineteen Participants were shown 14 groups of clips of two-person conversations. Specifically, we test it on our model, our model with first stage only, the best conversation model DualTalk [49] and the best single person model SelfTalk [46]. For each clip, participants were asked to select the single best result according to five aspects:

- **Speaker Lip Quality:** The realism and accuracy of the lip-sync for whoever is actively speaking.
- **Speaker Movements:** The naturalness and vividness of the speaker's head movements (excluding the lip region).
- **Listener Movements:** The realism and appropriateness of the listener's head movements and expression.
- **Interaction Quality:** The naturalness of the non-verbal interaction between the two subjects, including mutual head turning, reactions, etc.
- **Eye Gaze Quality:** The realism of eye movements for both participants during communication.

As shown in Tab. 4, our model was selected as the best most frequently across all criteria. It outperformed baselines significantly in terms of interaction and eye gaze. Training with the second stage provides greatly improves lip sync, also aligning with the quantitative metrics.

5. Conclusion

We address the challenging task of generating complete 3D facial animations for two co-located, interacting participants from a single mixed audio stream. Our framework explicitly models the dynamic 3D spatial relationship between speakers, a crucial yet often overlooked aspect of in-person dialogue. Enabled by a large-scale dataset built through in-the-wild video curation and a synthetic dubbing pipeline, our dual-stream generative model with cross-attention and an eye gaze loss produces realistic, controllable, and spatially coherent dyadic interactions.

References

[1] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models. *arXiv preprint arXiv:2312.08459*, 2023. 2

[2] Anonymous Author(s). NeRFFaceSpeech: One-shot audio-driven 3d talking head synthesis via generative priors + nerf. *arXiv preprint arXiv:2405.05749*, 2024. 2

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12449–12460, 2020. 6

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

[5] Yunpeng Bai, Yanbo Fan, Xuan Wang, Yong Zhang, Jingxiang Sun, Chun Yuan, and Ying Shan. High-fidelity facial avatar reconstruction from monocular video with generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4541–4551, 2023. 1

[6] 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. 2

[7] Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, Hao Yan, et al. Human conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents*, pages 29–63, 2000. 1

[8] Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[9] Kelvin C. K. Chan, Xiangyu Xu, Xintao Wang, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for image super-resolution and beyond, 2022. 6

[10] Baptiste Chopin, Tashvik Dhamija, Pranav Balaji, Yaohui Wang, and Antitza Dantcheva. Dimitra: Audio-driven diffusion model for expressive talking head generation. *arXiv preprint arXiv:2502.17198*, 2025. 2

[11] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10101–10111, 2019. 6

[12] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, page 1–13. ACM, 2023. 1, 2

[13] Stephan Diederich, Alfred Benedikt Brendel, Stefan Morana, and Lutz Kolbe. On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, 23(1):96–138, 2022. 1

[14] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[15] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):1–11, 2018. 5

[16] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[17] Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model. *arXiv preprint arXiv:2408.00762*, 2024. 2

[18] Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3D facial animation through a unified model. In *Proceedings of the European Conference on Computer Vision*, 2024. 2

[19] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3D facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 2, 7

[20] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 2

[21] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010. 6

[22] Scott Geng, Revant Teotia, Purva Tendulkar, Sachit Menon, and Carl Vondrick. Affective faces for goal-driven dyadic communication. *arXiv preprint arXiv:2301.10939*, 2023. 6

[23] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[24] Shreyank N Gowda, Dheeraj Pandey, and Shashank Narayana Gowda. From pixels to portraits: A comprehensive survey of talking head generation techniques and applications. *arXiv preprint arXiv:2308.16041*, 2023. 1

[25] Ying Guo, Xi Liu, Cheng Zhen, Pengfei Yan, and Xiaoming Wei. Arig: Autoregressive interactive head generation for real-time conversations, 2025. 3

[26] Hung-Hsuan Huang, Masato Fukuda, and Toyoaki Nishida. Toward rnn based micro non-verbal behavior generation for virtual listener agents. In *Social Computing and Social Media. Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I 21*, pages 53–63. Springer, 2019. 2

[27] Hyung Kyu Kim, Sangmin Lee, and Hak Gu Kim. Memorytalker: Personalized speech-driven 3d facial animation via audio-guided stylization. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1, 2

[28] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17978, 2023. 1

[29] Lizi Liao, Grace Hui Yang, and Chirag Shah. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3452–3455, 2023. 1

[30] Jin Liu, Xi Wang, Xiaomeng Fu, Yesheng Chai, Cai Yu, Jiao Dai, and Jizhong Han. Mfr-net: Multi-faceted responsive listening head generation via denoising diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6734–6743, 2023. 2

[31] Xi Liu, Ying Guo, Cheng Zhen, Tong Li, Yingying Ao, and Pengfei Yan. Customlistener: Text-guided responsive interaction for user-friendly listening head generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2415–2424, 2024. 1, 3

[32] Liying Lu, Tianke Zhang, Yunfei Liu, Xuangeng Chu, and Yu Li. Audio-driven 3d facial animation from in-the-wild videos. *arXiv preprint arXiv:2306.11541*, 2023. 2

[33] Cheng Luo, Siyang Song, Weicheng Xie, Micol Spitale, Linlin Shen, and Hatice Gunes. Reactface: Multiple appropriate facial reaction generation in dyadic interactions. *arXiv preprint arXiv:2305.15748*, 2023. 3

[34] Marcel Lüthi, Thomas Gerig, Clément Jud, and Thomas Vetter. Flame: Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36 (6):1–17, 2017. 2

[35] Sihan Ma, Qiong Cao, Jing Zhang, and Dacheng Tao. Contact-aware human motion generation from textual descriptions. In *arXiv preprint*, 2024. 2

[36] Dongchan Min, Minyoung Song, Eunji Ko, and Sung Ju Hwang. Styletalker: One-shot style-based audio-driven talking head video generation. *arXiv preprint arXiv:2208.10922*, 2022. 2

[37] Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[38] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael J. Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[39] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022. 3, 6, 7

[40] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10083–10093, 2023. 1, 3, 6

[41] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024. 3

[42] Arthur Niswar, Ee Ping Ong, Hong Thai Nguyen, and Zhiyong Huang. Real-time 3d talking head from a synthetic viseme dataset. In *Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry*, pages 29–33, 2009. 1

[43] Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillere, Stefano Berretti, and Mohamed Daoudi. Scantalk: 3d talking heads from unregistered scans. *arXiv preprint arXiv:2403.10942*, 2024. 2

[44] Misa Ogura and Matt Haynes. X-vector based voice activity detection for multi-genre broadcast speech-to-text, 2021. 5

[45] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2023. 1

[46] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023. 1, 2, 7, 8

[47] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023. 1

[48] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024. 1

[49] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025. 3, 6, 7, 8

[50] Ziqiao Peng et al. Emotalk: Speech-driven emotional disentanglement for 3D face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[51] M. Reid, R. Anil, A. Badia, and et al. Gemini 1.5: Unlocking multimodal understanding across modalities. *arXiv preprint arXiv:2403.05530*, 2024. 5, 6

[52] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3D face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5782–5791, 2021. 2

[53] Shihao Sang, Xiaokang Yang, Yujin Chen, Xiaoyi Dong, Yujun Cai, and Yong-Jin Liu. Timotion: Temporal and interactive modeling for human-human motion generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[54] Mengyi Shan, Lu Dong, Yutao Han, Yuan Yao, Tao Liu, Ifeoma Nwogu, Guo-Jun Qi, and Mitch Hill. Towards open-domain text-driven synthesis of multi-person motions. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[55] Luchuan Song, Guojun Yin, Zhenchao Jin, Xiaoyi Dong, and Chenliang Xu. Emotional listener portrait: Neural listener head generation with emotion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20839–20849, 2023. 3

[56] Siyang Song, Micol Spitale, Cheng Luo, Germán Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth André, et al. React2023: The first multiple appropriate facial reaction generation challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9620–9624, 2023. 1, 3

[57] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3D facial animation synthesis using diffusion. *Proceedings of ACM on Multimedia Interactive Experiences (MIG)*, 2023. 2

[58] Jiacheng Su, Kunhong Liu, Liyan Chen, Junfeng Yao, Qingsong Liu, and Dongdong Lv. Audio-driven high-resolution seamless talking head video editing via stylegan. *arXiv preprint arXiv:2407.05577*, 2024. 2

[59] Xusen Sun, Longhao Zhang, Hao Zhu, Peng Zhang, Bang Zhang, Xinya Ji, Kangneng Zhou, Daiheng Gao, Liefeng Bo, and Xun Cao. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. *arXiv preprint arXiv:2312.01841*, 2023. 2

[60] Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. Laughtalk: Expressive 3d talking head generation with laughter. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6404–6413, 2024. 1

[61] Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[62] Jan Tanke, Yi Zhou, Mohamed Omran, et al. Social diffusion: Long-term multiple human motion anticipation. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[63] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20621–20631, 2023. 1

[64] Balamurugan Thambiraja, Malte Prinzler, Sadegh Aliakbarian, Darren Cosker, and Justus Thies. 3diFACE: Synthesizing and editing holistic 3d facial animation. In *International Conference on 3D Vision 2025*, 2025. 1, 2

[65] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dim: Dyadic interaction modeling for social behavior generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXVII*, page 484–503, Berlin, Heidelberg, 2024. Springer-Verlag. 7

[66] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dyadic interaction modeling for social behavior generation. *arXiv preprint arXiv:2403.09069*, 2024. 1, 3

[67] Haoyu Wu, Ziqiao Peng, Xukun Zhou, Yunfei Cheng, Jun He, Hongyan Liu, and Zhaoxin Fan. Vgg-tex: A vivid geometry-guided facial texture estimation model for high fidelity monocular 3d face reconstruction. *arXiv preprint arXiv:2409.09740*, 2024. 1

[68] Yifan Xie, Fei Ma, Yi Bin, Ying He, and Fei Yu. Audio-driven talking face video generation with joint uncertainty learning. *arXiv preprint arXiv:2504.18810*, 2025. 2

[69] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 1, 2, 7

[70] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 2

[71] Hongyun Yu, Zhan Qu, Qihang Yu, Jianchuan Chen, Zhonghua Jiang, Zhiwen Chen, Shengyu Zhang, Jimin Xu, Fei Wu, Chengfei Lv, and Gang Yu. Gaussiantalker: Speaker-specific talking head synthesis via 3d gaussian splatting. *arXiv preprint arXiv:2404.14037*, 2024. 2

[72] Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, et al. Nofa: Nerf-based one-shot facial avatar reconstruction. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–12, 2023. 1

[73] Rui Zhen, Wenchao Song, Qiang He, Juan Cao, Lei Shi, and Jia Luo. Human-computer interaction system: A survey of talking-head generation. *Electronics*, 12(1):218, 2023. 1

[74] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: a benchmark dataset and baseline. In *European Conference on Computer Vision*, pages 124–142. Springer, 2022. 3, 6

[75] Xukun Zhou, Fengxin Li, Ziqiao Peng, Kejian Wu, Jun He, Biao Qin, Zhaoxin Fan, and Hongyan Liu. Meta-learning empowered meta-face: Personalized speaking style adaptation for audio-driven 3d talking face animation. *arXiv preprint arXiv:2408.09357*, 2024. 1

[76] Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. Infp: Audio-driven interactive head generation in dyadic conversations. *arXiv preprint arXiv:2412.04037*, 2024. 3