CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/177053/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Lai, Peiwen, Zhong, Weizhi, Qin, Yipeng , Ren, Xiaohang, Wang, Baoyuan and Li, Guanbin 2025. LLM-driven multimodal and multi-Identity listening head generation. Presented at: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025, Nashville, USA, 11 - 15 June 2025.

Publishers page:

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

LLM-driven Multimodal and Multi-Identity Listening Head Generation

Peiwen Lai[1] Weizhi Zhong[1] Yipeng Qin[2] Xiaohang Ren[3] Baoyuan Wang[3] Guanbin Li[1,4,5*]
[1]Sun Yat-sen University [2]Cardiff University [3]Xiaobing.AI [4]Peng Cheng Laboratory
[5]Guangdong Key Laboratory of Big Data Analysis and Processing
{laipw5, zhongwzh5}@mail2.sysu.edu.cn, qiny16@cardiff.ac.uk, xiaomums@qq.com
zjuwby@gmail.com, liguanbin@mail.sysu.edu.cn

Abstract

Generating natural listener responses in conversational scenarios is crucial for creating engaging digital humans and avatars. Recent work has shown that large language models (LLMs) can be effectively leveraged for this task, demonstrating remarkable capabilities in generating contextually appropriate listener behaviors. However, current LLM-based methods face two critical limitations: they rely solely on speech content, overlooking other crucial communication signals, and they entangle listener identity with response generation, compromising output fidelity and generalization. In this work, we present a novel framework that addresses these limitations while maintaining the advantages of LLMs. Our approach introduces a Multimodal-LM architecture that jointly processes speech content, acoustics, and speaker emotion, capturing the full spectrum of communication cues. Additionally, we propose an identity disentanglement strategy using instance normalization and adaptive instance normalization in a VQ-VAE framework, enabling high-fidelity listening head synthesis with flexible identity control. Extensive experiments demonstrate that our method significantly outperforms existing approaches in terms of response naturalness and fidelity, while enabling effective identity control without retraining.

1. Introduction

Generating natural and responsive listener behaviors is crucial for creating engaging conversational agents and avatars, which has recently attracted increasing research interest due to its wide range of applications in human-computer interaction [21, 50, 53], digital humans [55, 56], virtual reality [18, 19], metaverse [8, 9, 30], etc. While humans naturally provide non-verbal feedback through facial expressions and head movements during conversations, synthesizing these subtle yet meaningful responses remains a significant challenge in computer vision and graphics.

*Corresponding author is Guanbin Li.

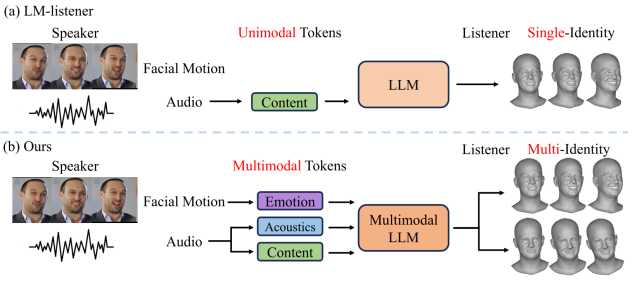

Figure 1. Comparison between LM-listener [24] and our method.

Several notable approaches have emerged in this rapidly evolving field, *e.g.*, Zhou *et al.* [55, 56] established a benchmark and a speaker-listener architecture for 3DMM [3] prediction; Learning2Listen [23] improved accuracy with VQ-VAE and motion-audio cross-attention; DIM [37] added self-supervised contrastive learning; while ELP [31] used emotional priors for more expressive outputs. A recent breakthrough came with the LM-listener [24], which shows that large language models (LLMs) can be effectively leveraged for listening head generation. By tapping into the rich semantic and contextual understanding capabilities of LLMs, LM-listener showed remarkable effectiveness in generating contextually appropriate listener responses from speaker speech content alone. This success revealed a promising direction: the strong prior knowledge embedded in LLMs about human communication patterns can be transferred to guide the generation of non-verbal responses, fundamentally advancing the field beyond traditional listening head generation approaches.

However, realizing the full potential of the LLM-based framework requires addressing two critical limitations in LM-Listener [24]. First, its unimodal approach (relying solely on speech content) fails to capitalize on the rich multimodal nature of human communication, where acoustic information and facial expressions significantly influence listener responses. Second, its entangled handling of listener identity and response generation not only compromises output fidelity but also limits the model's ability to

generalize across different listener identities. Addressing these limitations within an LLM framework presents unique challenges, as it requires careful consideration of how to effectively integrate multimodal information and disentangle identity control while preserving the powerful language understanding capabilities of LLMs.

In this paper, we present a novel framework based on LM-listener [24] that addresses these two challenges while maintaining the advantages of the LLM-based approach. First, we introduce a Multimodal-LM architecture that jointly processes speaker speech content, acoustics, and emotion (extracted from facial motion) to generate more contextually appropriate and natural listener responses. By incorporating speech acoustics through SpeechTokenizer [52] and discretizing facial motions into emotional states using EMOCA [10], our model captures the nuanced interplay between verbal and non-verbal communication cues. Second, we propose a novel identity disentanglement strategy that explicitly separates static identity features from dynamic facial motions during listening head generation. This separation is achieved through a carefully designed VQ-VAE architecture with instance normalization for identity erasure and adaptive instance normalization for identity injection, enabling high-fidelity response generation that can be easily adapted to different listener identities without retraining. Extensive experimental results demonstrate that our approach significantly outperforms existing methods in both the naturalness and the generalizability of generated listener behaviors. Our contributions include:

- We extend the promising LLM-based listening head generation framework to process multimodal inputs (speech content, acoustics, and emotion), enabling the model to capture the full spectrum of communication cues while preserving the advantages of language model priors.
- We develop a novel identity disentanglement approach for the VQ-VAE used in listening head generation using instance normalization and adaptive instance normalization, achieving high-fidelity output with flexible identity control without retraining.
- Extensive experiments demonstrate that our approach significantly outperforms existing methods in response naturalness and fidelity, while providing effective identity control across different listeners.

2. Related Work

2.1. Listener Response Generation

Conversational avatars involve not only speaker motion [2, 16, 25, 44, 54] but also responsive listener motion that provides non-verbal feedback to the speaker. Previous methods [1, 4, 7, 32] have focused on generating conversational agents that interact through speech [7], gestures [1, 32] or a combination of multiples modalities [4]. Recent methods [12, 23, 24, 31, 37, 55] have shifted focus to the synthesis of listening head motions, including facial expressions and head poses. For example, RLHG [55] propose the listener-centric task of listening head generation and present a new benchmark dataset and baseline. The baseline method consisting of a speaker encoder and listener decoder predicts the 3DMM [3] parameters of head pose and facial expression, which are further rendered into the listening video by a neural renderer. Learning2Listen [23] applies the VQ-VAE [39] to the domain of listener motion generation and predicts the listener motion in a quantized motion space via motion-audio cross-attention. Based on this, DIM [37] devised a pre-training strategy through self-supervised contrastive learning to learn a unified representation for listener motion generation. ELP [31] utilizes emotional priors to rearrange the latent space for emotional listener head generation. Recently, LM-listener [24] demonstrated that large language models (LLMs) can be effectively used for listening head generation, which can generate contextually appropriate listener responses from speech content alone. However, its unimodal approach and entangled handling of listener identity and response generation limit the naturalness and fidelity of its outputs, and its generalizability across diverse listener identities.

In this work, we address these two limitations by introducing a Multimodal-LM architecture to incorporate multimodal cues and an identity-disentangled VQ-VAE for flexible identity control.

2.2. Large Language Models

Large Language Models (LLMs) [6, 26, 35] have showcased remarkable capability in various vision and language tasks [15, 20, 29, 41, 43, 45], thanks to their scalable models and the large-scale dataset for training. Given the inherently multimodal nature of real-world environments, many studies [33, 45–48, 51] have focused on developing LLMs capable of perceiving or generating multimodal signals, leading to the emergence of multimodal LLMs (MM-LLMs). For example, AnyGPT [51] integrates new modalities into LLMs by converting various modalities into discrete representations, while keeping the model's architecture unchanged and performing multimodal understanding and generation. X-VILA [45] proposes a cross-modality alignment mechanism that can align the features of various modalities with the LLM textual embedding, facilitating cross-modality understanding, reasoning, and generation. GenArtist [42] utilizes MM-LLM as an agent to plan and invoke external tools through tree structure for unified image generation and editing. ShapeGPT [49] proposes a shape-included MM-LLM framework for 3D shape generation by discretizing continuous shapes into shape words.

In this work, we follow the MM-LLMs paradigm and propose a novel approach for listening head generation.

3. Method

In this work, we address two primary limitations of LM-listener [24]: (i) its unimodal approach, which introduces ambiguity in listener responses; and (ii) its entanglement of listener identity with response generation, which not only reduces the fidelity of the output but also makes it challenging to control or alter identities without retraining. To overcome these limitations, we propose: (i) integrating multimodal speaker information (*i.e.*, speech content, acoustics, and emotion) into speaker-listener modeling rather than relying solely on the speaker's unimodal speech content (Section 3.1); and (ii) disentangling listener identity control from response synthesis (Section 3.2), which not only effectively separates static components (*i.e.*, identity) from dynamic elements (*i.e.*, listener facial motion) in the generation process to improve fidelity, but also enables flexible response generation across multiple listener identities without retraining.

Notations. For an input speaker video containing T frames, denoted as $t = \{1, 2, \dots, T\}$, we represent its corresponding facial motion as $\mathbf{F}^s = \{f_1^s, f_2^s, \dots, f_T^s\}$ and audio as $\mathbf{A}^s$, where $f_t^s = [\psi_t^s, \theta_t^s]$ is defined using the 3DMM [3] facial expression parameters ψ and head pose parameters θ. Note that we intentionally exclude the identity-specific shape parameter β in f, as it remains static for each speaker and listener. This exclusion facilitates disentangled identity control and helps mitigate biases introduced by the speaker's identity. Similarly, we represent the output listener facial motion as $\mathbf{F}^l = \{f_1^l, f_2^l, \dots, f_T^l\}$. The 3DMM parameters for each video frame are extracted following the methods in [10, 11].

3.1. Multimodal Speaker-Listener Modeling

3.1.1. Multimodal Speaker Input

Speech typically comprises two key components: content, which includes phonemes and syllables, and acoustic information, including prosody, timbre, and stress patterns. Unlike LM-listener that relies solely on the unimodal content of speaker audio, our approach decouples the speaker's speech into distinct content and acoustics components. This separation allows our model to perform a more in-depth analysis, leveraging both content and acoustics to enhance its understanding of the speaker's intended meaning.

Speech Content. To represent the speech content, we leverage text-based encoding. Specifically, we employ a pretrained automatic speech recognition (ASR) model, Whisper [27], to transcribe the speaker audio $\mathbf{A}^s$ into text $\mathbf{W} \in \mathbb{R}^{N \times 1}$, where N is the number of text tokens corresponding to the input speaker video of T frames.

Speech Acoustics. Speech acoustics captures the delivery style of speech, which plays a critical role in conveying the speaker's emotions, attitudes, and intentions. To extract this information from $\mathbf{A}^s$, we utilize the pretrained SpeechTokenizer model [52], a residual vector quantization (RVQ) network with 8 quantizers. The first quantizer's output is interpreted as content tokens, while the outputs from the remaining quantizers primarily capture the acoustic features. To balance performance and computational efficiency, we select the output of the second quantizer as the representation of speech acoustics, denoted as $\mathbf{SA} \in \mathbb{R}^{M \times 1}$, where M is the number of acoustic tokens corresponding to the input speaker video of T frames.

Emotion. Building on the Emotional Contagion theory [13, 40], which demonstrates the significant impact of speaker emotions on listener responses, we discretize the speaker's facial motion into a finite set of emotional states rather than their detailed expressions. To achieve this, we first downsample the sequence of speaker facial motions $\mathbf{F}^s = \{f_1^s, f_2^s, \dots, f_T^s\}$ by a rate of r, dividing $\mathbf{F}^s$ into T/r groups. For each group, we use the emotion recognition module (ER) from EMOCA [10] to predict the emotion probability distribution for each facial motion within the group. Then, we average the emotion probability distributions across all motions in the group. Finally, the emotion state with the highest probability is selected as the emotion token for that group's facial motions. This process is repeated for all groups, resulting in the final speaker emotion tokens $\mathbf{emo} \in \mathbb{R}^{(T/r) \times 1}$. Please see the supplementary materials for more details.

3.1.2. Multimodal-LM for Listener Generation

Leveraging the multimodal speaker inputs (*i.e.*, speech content $\mathbf{W}$, speech acoustics $\mathbf{SA}$, and emotion $\mathbf{emo}$) introduced above, we extend unimodal language models (LMs) to process multimodal input tokens as follows.

Specifically, we begin with a transformer-based LM, $\mathcal{G}$, which takes a sequence of text tokens as input and outputs a probability distribution over the vocabulary to predict the next token. Then,

- To adapt $\mathcal{G}$ for multimodal input, we randomly initialize word embeddings $E_{\text{MM}} \in \mathbb{R}^{(V_{\text{LR}}+V_{\text{SA}}+V_{\text{emo}}) \times d_w}$ and append it to the native word embedding of speech content (text) W in $\mathcal{G}$, where V_{LR}, V_{SA}, and V_{emo} are the number of listener response (facial motion) tokens, speaker speech acoustic tokens, and speaker emotion tokens, respectively; d_w is the dimension of token embeddings.
- To adapt $\mathcal{G}$ for listener response generation, we additionally append another randomly initialized affine projection layer to its output, producing a probability distribution over listener tokens.

Sequential Organization of Multimodal Input. As shown in Figure 2, we structure the input sequence S to $\mathcal{G}$ in interleaved order: $\mathbf{LR}_i, \mathbf{W}_{ri+1:r(i+1)}, \mathbf{SA}_{ri+1:r(i+1)}, \mathbf{emo}_{i+1}, \mathbf{LR}_{i+1}, \dots$, where $\mathbf{W}_{t_1:t_2}$ represents the words spoken be-

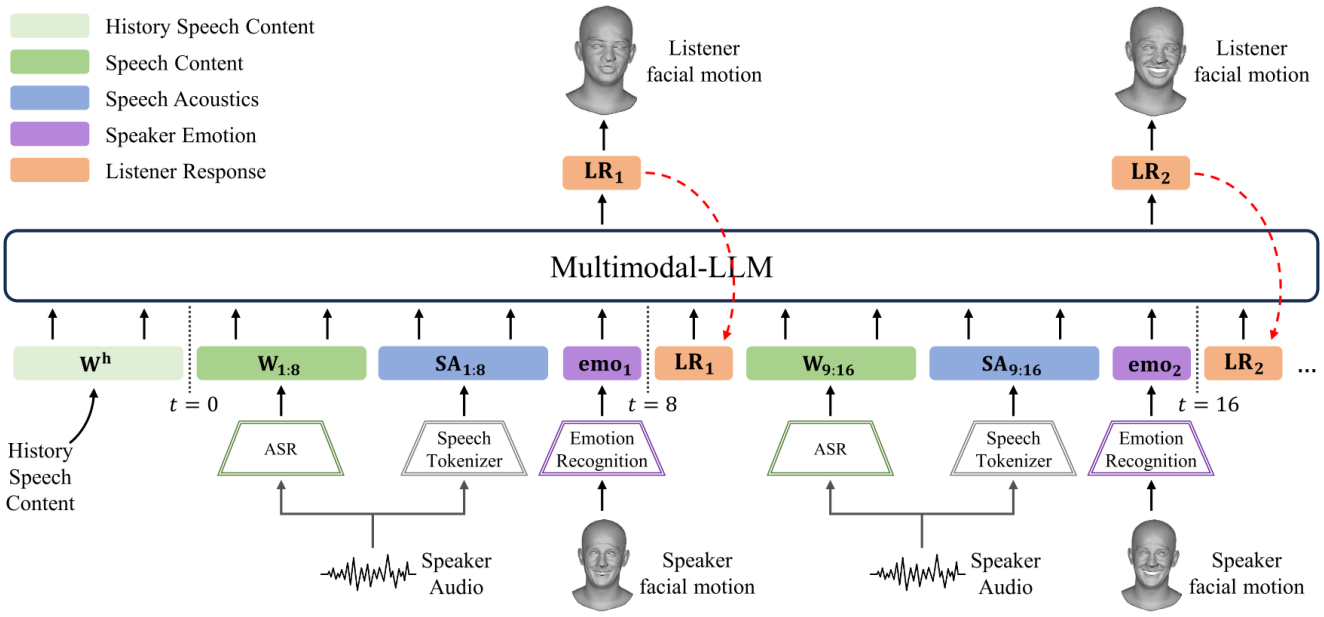

Figure 2. Illustration of the proposed Multimodal-LLM for listening head generation.

tween frames t_1 and t_2; $\mathbf{SA}_{t_1:t_2}$ corresponds similarly to the acoustic information within the same time frame. Note that the output $\mathbf{LR}_{i+1}$ of $\mathcal{G}$, generated based on the previous tokens as input, is fed back as input for the next prediction step, thereby preserving *causality* between speaker input and listener response. We also incorporate N' history speech content tokens $\mathbf{W}^h = \{w_1^h, w_2^h, \ldots, w_{N'}^h\}$ that occur before the first frame as additional contextual information. Except for speech content tokens, tokens from all other modalities include special start and end tokens to mark their boundaries. Finally, we have:

$$S = \{\mathbf{W}^h, \mathbf{W}_{1:r}, \mathbf{SA}_{1:r}, \mathbf{emo}_1, \mathbf{LR}_1, \mathbf{W}_{r+1:2r}, \mathbf{SA}_{r+1:2r}, \ldots, \mathbf{LR}_{T/r}\} \quad (1)$$

Training. The training of our Multimodal-LM consists of two stages, each of which uses cross-entropy loss to optimize the model for next-token prediction.

Stage 1. Speaker Understanding Pretraining.
We initialize $\mathcal{G}$ from a standard LM and fine-tune it to learn the semantics of the speaker acoustic and emotion tokens. In this stage, $\mathcal{G}$ processes a *speaker-only* input sequence:

$$S' = \{\mathbf{W}^h, \mathbf{W}_{1:r}, \mathbf{SA}_{1:r}, \mathbf{emo}_1, \mathbf{W}_{r+1:2r}, \mathbf{SA}_{r+1:2r}, \ldots, \mathbf{emo}_{T/r}\} \quad (2)$$

and outputs probabilities over the entire vocabulary. The loss is calculated for speech content tokens, speech acoustic tokens, and speaker emotion tokens:

$$\mathcal{L}_{\text{pre}} = -\sum_{j=1}^{N+M+T/r} \log \Pr[\mathcal{G}(S'_{1:N'+j-1}) = S'_{N'+j}] \quad (3)$$

which enables $\mathcal{G}$ to learn the probability distributions of various speaking styles and emotions.

Stage 2. Listener Response Fine-tuning.
In this stage, $\mathcal{G}$ takes S (Equation (1)) as input and outputs the probability distribution for each listener response token $\mathbf{LR}$. Here, we calculate the loss only for $\mathbf{LR}$:

$$\mathcal{L} = -\sum_{i=1}^{T/r} \log \Pr[\mathcal{G}(\mathbf{W}^h, \mathbf{W}_{1:ri}, \mathbf{SA}_{1:ri}, \mathbf{emo}_{1:i-1}, \mathbf{LR}_{1:i-1}) = \mathbf{LR}_i]. \quad (4)$$

which enables $\mathcal{G}$ to learn the probability distributions of various listener responses.

3.2. Listener Identity Disentanglement

Unlike [24] which requires the listener response decoder to simultaneously learn both listener identity and facial motion (which is a challenging task that compromises fidelity), we propose a novel strategy to explicitly inject listener identity information into the decoding process, thereby facilitating the decoder training. By disentangling listener identity from response (facial motion) synthesis, our approach not only improves the fidelity of the generated responses but also significantly enhances the model's generalization to multiple identities without requiring retraining.

Specifically, we formulate the training of the listener response decoder as a proxy reconstruction task using VQ-VAE [39], whose architecture is introduced as follows:

Overview of VQ-VAE. In a nutshell, our VQ-VAE aims to reconstruct a given sequence of listener facial motions $\mathbf{F}^l$

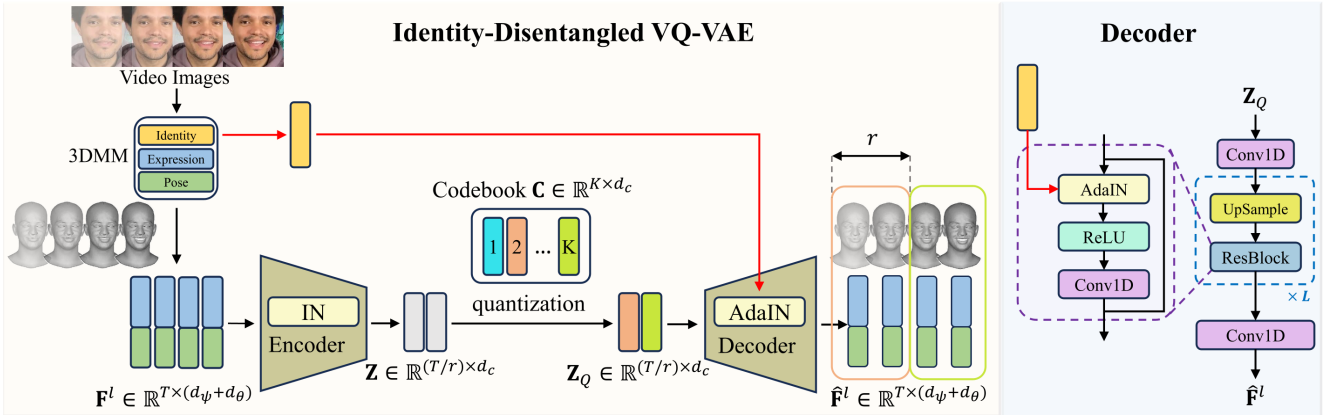

Figure 3. Network architecture of the proposed identity-disentangled VQ-VAE for listening head generation.

using a codebook of listener response token embeddings as their latent representations. As shown in Figure 3 (left), our VQ-VAE consists of an encoder, a decoder, and a codebook $\mathbf{C} \in \mathbb{R}^{K \times d_c}$ of $\mathbf{LR}$ embeddings, where $K = V_{\text{LR}}$ and d_c is the dimension of $\mathbf{LR}$ embeddings in VQ-VAE. Each $\mathbf{LR}$ embedding corresponds to an atomic listener facial motion. Then, the encoder takes the sequence of listener facial motions $\mathbf{F}^l = \{f_1^l, f_2^l, \ldots, f_T^l\}$ as input and produces a sequence of latent features $\mathbf{Z} = \{z_1, z_2, \ldots, z_{T/r}\}$, $z_i \in \mathbb{R}^{d_c}$, where r is the downsampling rate described earlier. Then, a deterministic and parameter-free quantizer function Q is applied to map each z_i into an $\mathbf{LR}$ embedding in $\mathbf{C}$ using nearest-neighbor matching:

$$\mathbf{LR}_j = Q(z_i) = \arg \min_{1 \le j \le K} \|z_i - \mathbf{LR}_j\|^2, \qquad (5)$$

Finally, the decoder takes as input the quantized latent features $\mathbf{Z}_Q = \{Q(z_1), Q(z_2), \ldots, Q(z_{T/r})\}$ along with the injected listener identity information β^l, and outputs the reconstructed listener facial motions $\hat{\mathbf{F}}^l = \{\hat{f}_1^l, \hat{f}_2^l, \ldots, \hat{f}_T^l\}$.

Listener Identity Erasure and Injection. To disentangle listener identity from response synthesis, we need to i) erase the identity information from the encoder and ii) inject the identity information back into the decoder. Specifically, we treat listener identity as a "style" control for response synthesis and have:

- To obtain an identity-agnostic codebook of $\mathbf{LR}$, we employ instance normalization layers [38] in the encoder to eliminate identity-specific variations, enabling the model to focus on the content of facial motions rather than individual style, thereby facilitating the codebook learning.
- To ensure accurate facial motion reconstruction from the identity-agnostic $\mathbf{LR}$, we incorporate adaptive instance normalization (AdaIN) layers [17] in the decoder to inject identity information back into the response synthesis process (Figure 3, right). Specifically, the 3DMM identity parameter β^l of a listener is introduced as a conditional input to adjust the mean and variance of the feature maps during decoding, thereby enabling flexible control and altering of listener identities without retraining.

Training. Following the common practice [24, 39], we use a set of different losses to train our VQ-VAE, including: $\mathcal{L}_{embed}$, $\mathcal{L}_{commit}$, $\mathcal{L}_{rec}$, and $\mathcal{L}_{veloc}$. Exponential moving average and codebook reset operations are also employed to enhance the efficiency and stability of the training process. Please see the supplementary materials for more details.

Remark. Note that the listener response $\{\mathbf{LR}_i\}_{i=1}^{T/r}$ used in training our Multimodal-LM (Section 3.1.2) is obtained by applying the trained encoder to $\mathbf{F}^l = \{f_i^l\}_{i=1}^{T}$.

4. Experiment

4.1. Experimental Settings

Datasets. We use two open-source listener datasets: L2L-trevor [24] and RealTalk [12] to evaluate our method. L2L-trevor is a single-listener (Trevor Noah) dataset introduced in Learning2Listen [23] and enhanced in LM-listener [24]. RealTalk is a multi-listener dataset proposed in [12]. Following the preprocessing in [24], we process the datasets in three steps. First, we segment the raw videos into 8-second segments to ensure sufficient context and perform speech separation [5, 34] to identify listener segments. Next, we use EMOCA [10, 11] to extract 3DMMs of both the speaker and listener from the videos, further recognizing the speaker's emotions. Finally, we employ Whisper [27] and SpeechTokenizer [52] to extract time-aligned speech transcriptions and acoustic information. This results in 2,366 training, 222 validation, and 543 test segments in L2L-trevor, and 2,714 training, 238 validation, and 808 test segments in RealTalk.

Implementation Details. (i) For VQ-VAE, the codebook size V_{LR} is 256, the dimension of the codebook embedding d_c is 512, and the downsampling rate r is 8. The weights for $\mathcal{L}_{commit}$ and $\mathcal{L}_{veloc}$ are set to 0.02 and 0.5 respectively, with the rest set to 1. (ii) For the language model, we in-

Method	L2 ↓	FD ↓	Variation	Diversity	P-FD ↓	L2 Affect(10^2) ↓
GT			0.1148	2.6053		
Random	0.6791	32.6036	0.1035	2.4601	33.7893	11.1136
NN Facial Motion	0.5682	26.4138	0.0896	2.2948	27.6294	9.6732
NN Speech Content	0.5232	23.9410	0.0884	2.2834	25.0246	7.8624
Naive Random Walk	0.7103	29.5126	0.2664	4.8417	31.5963	9.8946
LM-listener (GPT2) [24]	0.4485	18.5156	0.1215	2.8880	19.9911	6.4928
LM-listener (LLAMA2) [24]	0.4345	17.6299	0.1189	2.9374	19.1583	6.3992
Ours Random Walk	0.3456	11.9217	0.0961	2.7620	13.3772	2.9964
Ours (GPT2)	**0.2848**	**9.8093**	0.0762	2.3280	**11.0835**	2.6575
Ours (LLAMA2)	0.2910	10.0949	0.0704	2.2960	11.3908	**2.5797**

Table 1. Quantitative comparison on the L2L-trevor dataset [24].

stantiate $\mathcal{G}$ as GPT2-Medium [26] and LLAMA2-7B [36], using full parameter fine-tuning and LoRA [14] fine-tuning respectively. We use an AdamW [22] optimizer in training. During testing, we use greedy sampling to predict motion tokens from the language model. Please see the supplement for more details.

Metrics. Following [24], we evaluate our method based on realism (*L2* and *Frechet Distance (FD)*), diversity (*Variation* and *Diversity*), and synchrony (*Paired FD (P-FD)* and *L2 Affect*). Please see the supplement for more details.

Baselines. We compare ours to the following baselines:

- **Random**: Return a random listener sequence (train set).
- **NN Facial Motion**: For an input facial motion sequence, return the corresponding listener sequence of its nearest neighbor (cosine similarity) in the training set.
- **NN Speech Content**: Same as above, but we find NN via text embeddings obtained from Sentence-BERT [28].
- **Naive Random Walk**: Randomly walk over codebook indices. The codebook is from the naive VQ-VAE used in [23, 24] without the listener identity disentanglement.
- **Ours Random Walk**: Same as above, but the codebook is from our identity-disentangled VQ-VAE.
- **LM-listener** [24]: The state-of-the-art listener response generation method. It uses a naive VQ-VAE and predicts listener responses using only speaker's speech contents.

4.2. Quantitative Results

L2L-trevor Dataset. As Table 1 shows, **Ours** significantly outperforms all baselines across a range of metrics: i) For realism, measured by *L2* and *FD*, our approach yields listener responses and their distributions closest to the ground truth, with improvements of 38% and 45% over [24], respectively. ii) Our model also excels in diversity, producing listener responses with variability comparable to real data. iii) Notably, **Ours** achieves high performance on *P-FD* and *L2 Affect* for synchrony, demonstrating that it effectively captures conversational dynamics, generating listener responses with facial motions that synchronize well with the speaker. Interestingly, **Ours Random Walk** significantly outperforms **Naive Random Walk** and even surpasses **LM-listener** [24], indicating that our identity VQ-VAE can generate realistic, individualized listener motions, even with random sampling from the codebook.

Method	L2 ↓	FD ↓	Variation	Diversity	P-FD ↓	L2 Affect(10^2) ↓
GT			0.0260	1.3087		
Random	0.3127	15.7005	0.0437	1.7969	16.0617	21.3266
NN Facial Motion	0.2951	15.0987	0.0274	1.3456	15.3962	19.1231
NN Speech Content	0.2773	13.7154	0.0408	1.7148	14.0675	16.0370
Naive Random Walk	0.2253	10.2483	0.0549	2.1963	10.7152	14.0559
LM-listener [24]	0.2026	9.6016	0.0313	1.5817	9.9899	11.6484
Ours Random Walk	0.1168	4.7305	0.0355	1.4534	5.5045	6.4791
Ours	**0.0860**	**3.3939**	0.0130	1.0426	**3.6768**	**5.2537**

Table 2. Quantitative comparison on the RealTalk dataset [12].

Generalization across LM models. To demonstrate the generalizability of our approach, we evaluate it using two language models: GPT2-Medium [26] and LLAMA2-7B [36]. As Table 1 shows, **Ours** consistently outperforms **LM-listener** [24] across both models, confirming that the performance gains achieved are not model-specific. Notably, there is minimal performance difference between **Ours (GPT2)** and **Ours (LLAMA2)**, likely due to dataset size limitations and variations in fine-tuning procedures. Balancing performance and computational cost, we proceed with GPT2-Medium [26] for subsequent experiments.

RealTalk Dataset. As Table 2 shows, **Ours** achieves the highest performance across a range of metrics as well. Notably, **Ours** outperforms **LM-listener** [24] by 60% in *L2* and 65% in *FD*, a substantial improvement over the gains of 38% and 45% observed on the single-listener L2L-trevor dataset mentioned above. This highlights our model's enhanced ability to generate realistic facial motions across diverse listener styles, facilitated by incorporating listener identity information. Additionally, our approach sets a new state-of-the-art on metrics that assess distribution distances between listener-speaker pairs (*P-FD*), and on metrics evaluating listener facial affect accuracy (*L2 Affect*) between generated and ground-truth data.

4.3. Qualitative Results

Listener Response Consistency. We show our method's ability to generate listener responses aligned with the emotional context of a conversation by comparing it to LM-listener [24] across different emotional scenarios (Fig. 4). For convenience, speaker speech style descriptions are used to represent speech acoustics. In the first two scenarios, where the speaker's sarcastic humor is difficult to capture from text alone, our method successfully perceives the speaker's emotional state (*happy*) and style (*humorous*), generating a synchronized response (*laughter*). In the third and fourth scenarios, although the speaker uses positive words (*best* and *love*), their serious facial expressions and tone prompt our model to generate *calm* expressions rather than positive ones. By integrating multimodal cues, our method accurately captures the speaker's emotional state, producing listener responses that align with the context.

Results Generated with Different Identities. As shown in Fig. 5, unlike naive VQ-VAE [23, 24], ours disentan-

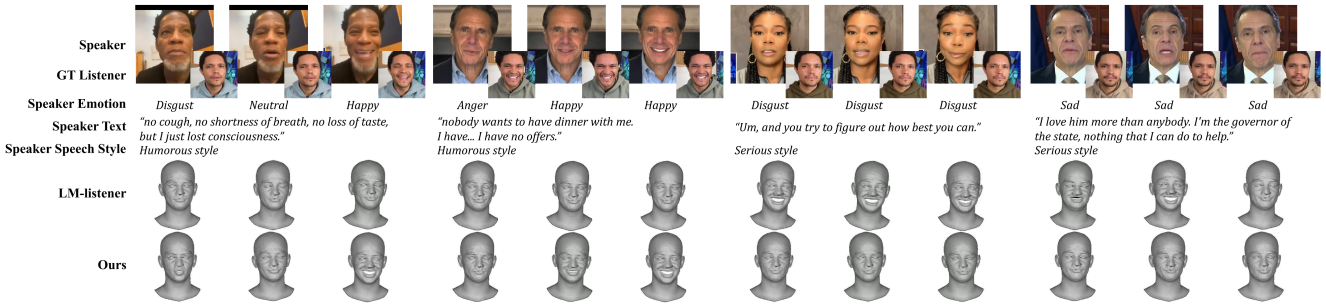

Figure 4. Qualitative comparison on the L2L-trevor dataset [24]. Please zoom in for the best view.

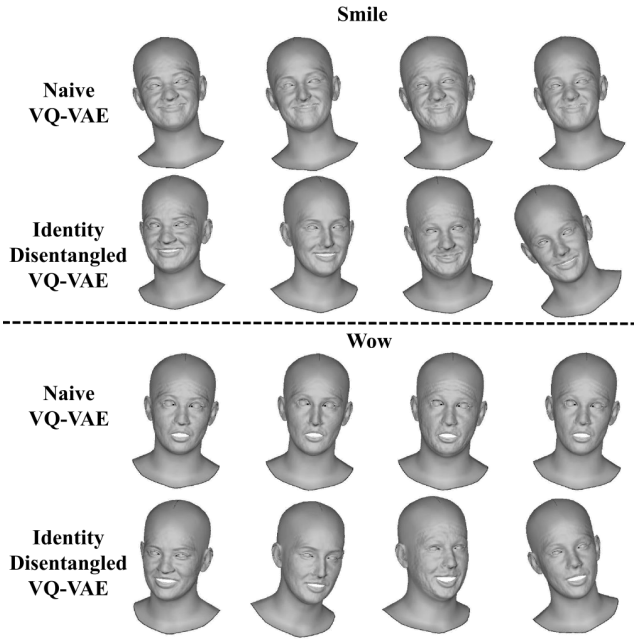

Figure 5. Qualitative comparison of listener responses generated with different identities on the RealTalk dataset [12].

gles listener identity from response synthesis, enabling flexible response generation across multiple listener identities without retraining. Specifically, the naive VQ-VAE produces fixed expressions and poses for a given **LR** token, disregarding identity differences. In contrast, our identity-disentangled VQ-VAE adapts the same token to generate identity-specific expressions and poses, capturing unique listener styles, *e.g.*, open-mouth versus closed-mouth smiles (Fig. 5, top), thereby learning and reconstructing more realistic, individualized listener facial motions.

User Study. To validate our quantitative results, we conducted a user study with 33 volunteers. Each volunteer watched a series of videos, each comprising a speaker video paired with three types of listener responses generated by our method, LM-listener [24], and synthesized ground truth, presented in random order. Volunteers ranked the responses based on how well they aligned with the emotional context and synchronized with the speaker. Our method significantly outperformed LM-listener [24], with volunteers preferring it in 86.4% of instances, mirroring the quantitative trends shown in Tab. 1. Additionally, when compared with ground truth, our method was preferred 65.8% of the time, highlighting the perceptual realism of our generated listener motions.

Input Modality	L2 ↓	FD ↓	Variation	Diversity	P-FD ↓	L2 Affect(10^2) ↓
GT			0.1148	2.6053		
W	0.4485	18.5156	0.1215	2.8880	19.9911	6.4928
W + iSC	0.4531	18.7325	0.1123	2.8307	20.2486	6.4200
W + SA	0.4477	18.3777	0.1232	2.9453	19.8907	6.6520
W + emo	0.4320	17.9176	0.1141	2.8006	19.3585	5.9971
W + iSC + emo	0.4510	18.7733	0.1071	2.7377	20.2696	6.7678
W + SA + emo	0.4027	16.5647	0.0992	2.6329	17.9687	5.5912

Table 3. Ablation study of the speaker's different modality tokens on the L2L-trevor dataset [24].

4.4. Ablation Study

4.4.1. Effectiveness of Multimodal Tokens

To assess the effectiveness of the speaker's different modality tokens (Sec. 3.1.1), we experiment with various combinations of speech content (**W**), speech acoustics (**SA**), and emotion (**emo**). Note that we also include in our test the implicit speech content (**iSC**), which is the output of the first quantizer in SpeechTokenizer [52], to further demonstrate the effectiveness of speech content.

As Tab. 3 shows, using **W** as a baseline, we find that combinations including **iSC**, such as **W + SiC** and **SC + iSC + emo**, yield no performance gains. This is likely because **iSC** is semantically redundant with **W**, adding complexity without new information. Similarly, **W + SA** shows limited improvement, as low-level **SA** information is challenging for the model to leverage for listener response generation. In contrast, **W + emo** significantly improves listener affect accuracy (*L2 Affect*) since **emo** clearly conveys the emotional context through distinct semantic labels. The combination **W + SA + emo** achieves the best results across all metrics, suggesting that **emo** enhances the model's understanding of low-level **SA**, creating a complementary effect. Note that **iSC + SA + emo** is not included as it makes no

Method	MultiModal	Identity	L2 ↓	FD ↓	Variation	Diversity	P-FD ↓	L2 Affect(10^2) ↓
GT					0.1148	2.6053		
Naive	×	×	0.4485	18.5156	0.1215	2.8880	19.9911	6.4928
MM only	√	×	0.4027	16.5647	0.0992	2.6329	17.9687	5.5912
ID only	×	√	0.3009	10.4572	0.0847	2.4312	11.7611	2.9486
Full (Ours)	√	√	0.2848	9.8093	0.0762	2.3280	11.0835	2.6575

Table 4. Ablation study of the different components (Multimodal-LM and Identity-disentangled VQ-VAE) on the L2L-trevor dataset [24].

use of the learned priors in large language models in interpreting speaker input.

4.4.2. Choice of IN/AdaIN for Disentangling Identity

To validate our choice of Instance Normalization (IN) and Adaptive Instance Normalization (AdaIN) for identity disentanglement, we test VQ-VAE models with different normalization configurations: none (∅), instance/layer normalization (**IN** or **LN**), and adaptive instance/layer normalization (**AdaIN** or **AdaLN**). Specifically, (∅, ∅) denotes no normalization in both encoder and decoder, while our identity-disentangled VQ-VAE (**IN**, **AdaIN**) applies IN in the encoder and AdaIN in the decoder. We assess performance using *Reconstruction* (Rec.), *Commitment* (Commit.), and *Perplexity* (PPL), with details in the supplement.

As Tab. 5 shows, i) Incorporating normalization layers significantly improves reconstruction performance for listener response generation. Specifically, normalization layers use the input motion sequence's mean and variance to enhance results, whereas the naive VQ-VAE uses the dataset-wide mean and variance, which degrades performance when handling input with diverse distributions. ii) The adaptive normalization layer in the decoder notably boosts *Commit.* by allowing the model to adjust the mean and variance of the output sequence according to identity, improving the reconstruction of listener facial motions with varying distributions. Without adaptive normalization, the codebook bears the responsibility of denormalization, which harms token quality and increases the distance to the encoder output. iii) Our identity-disentangled VQ-VAE achieves improvements of 66% and 87% in *Rec.* on the L2L-trevor [24] and RealTalk [12] datasets, respectively, demonstrating its ability to better represent diverse listener motion distributions. iv) Finally, we observe minimal performance differences between instance and layer normalization. We opt for instance normalization due to its prevalence in identity transfer [17].

4.4.3. Effectiveness of Different Components

To justify the effectiveness of our proposed Multimodal-LM and Identity-disentangled VQ-VAE, we conduct an ablation study on them. Specifically, we use **Naive** to denote the LM-listener [24] baseline, which uses a naive VQ-VAE and predicts listener motion from speech content only; we use **MM only** to denote the incorporation of our Multimodal-LM only (Sec. 3.1); and we use **ID only** to denote the integration of our Identity-disentangled VQ-VAE only (Sec. 3.2); and **Full** denotes the proposed method.

Dataset	Method	Rec. ↓	Commit. ↓	PPL. ↑
L2L-trevor	∅, ∅	0.60	2.10	3.53
	LN, LN	0.19	0.90	3.71
	LN, AdaLN	0.22	0.15	3.67
	IN, IN	0.17	0.68	4.00
	IN, AdaIN	0.21	0.18	3.79
RealTalk	∅, ∅	0.49	1.25	3.38
	LN, LN	0.05	0.27	3.82
	LN, AdaLN	0.08	0.05	3.68
	IN, IN	0.04	0.20	3.98
	IN, AdaIN	0.06	0.08	3.97

Table 5. Justification of choice of IN/AdaIN for identity disentanglement on the L2L-trevor [24] and RealTalk [12] datasets.

As shown in Tab. 4, **MM only** primarily improves *L2 Affect*, suggesting that our Multimodal-LM enhances the model's understanding of the speaker, helping generate listener responses aligned with the emotional context. **ID only** allows the identity-disentangled VQ-VAE to produce more realistic listener responses, yielding significant improvements. Finally, **Full** achieves the best performance across all metrics, demonstrating that these components complement each other effectively. Additional results on RealTalk [12] are provided in the supplementary materials.

5. Conclusion

In this work, we propose a novel framework that advances listening head generation by addressing two key limitations of LLM-based approaches: i) Our method extends the LLM framework to jointly consider speech content, acoustics, and speaker emotion through a carefully designed token organization strategy; ii) Our identity disentanglement approach using IN and AdaIN in the VQ-VAE framework enables high-fidelity listening head synthesis while providing flexible identity control without model retraining. Extensive experiments demonstrate the effectiveness of our approach.

Acknowledgement

This work is supported in part by the National Key R&D Program of China (2024YFB3908503, 2024YFB3908500), in part by the National Natural Science Foundation of China under Grant NO. 62322608 and in part by the CCF-Kuaishou Large Model Explorer Fund (NO. CCF-KuaiShou 2024007).

References

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

[2] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21263–21273, 2024. 2

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. 1, 2, 3

[4] Dan Bohus and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8, 2010. 2

[5] Hervé Bredin. pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Conference of the International Speech Communication Association*, pages 1983–1987. ISCA, 2023. 5

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 2

[7] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, pages 413–420, 1994. 2

[8] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. Rapport with virtual agents: What do human social cues and personality explain? *IEEE Transactions on Affective Computing*, 8(3):382–395, 2016. 1

[9] Lele Chen, Chen Cao, Fernando De la Torre, Jason Saragih, Chenliang Xu, and Yaser Sheikh. High-fidelity face tracking for ar/vr via deep lighting adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13059–13069, 2021. 1

[10] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 2, 3, 5

[11] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022. 3, 5

[12] Scott Geng, Revant Teotia, Purva Tendulkar, Sachit Menon, and Carl Vondrick. Affective faces for goal-driven dyadic communication. *arXiv preprint arXiv:2301.10939*, 2023. 2, 5, 6, 7, 8

[13] Elaine Hatfield, Richard L. Rapson, and Yen-Chi Lam Le. Emotional contagion and empathy. In *The social neuroscience of empathy*. Citeseer, 2009. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022. 6

[15] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 2

[16] Ricong Huang, Peiwen Lai, Yipeng Qin, and Guanbin Li. Parametric implicit face representation for audio-driven facial reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12759–12768, 2023. 2

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 5, 8

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

[19] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[21] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems*, 35:21386–21399, 2022. 1

[22] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[23] 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. 1, 2, 5, 6

[24] 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, 2, 3, 4, 5, 6, 7, 8

[25] 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. 2

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 6

[27] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 3, 5

[28] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 6

[29] Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023. 2

[30] Luchuan Song, Guojun Yin, Bin Liu, Yuhui Zhang, and Nenghai Yu. Fsft-net: face transfer video generation with few-shot views. In *2021 IEEE International Conference on Image Processing*, pages 3582–3586. IEEE, 2021. 1

[31] Luchuan Song, Guojun Yin, Zhenchao Jin, Xiaoyi Dong, and Chenliang Xu. Emotional listener portrait: Realistic listener motion simulation in conversation. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20782–20792. IEEE, 2023. 1, 2

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

[33] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434, 2024. 2

[34] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 3927–3935, 2021. 5

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[37] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dim: Dyadic interaction modeling for social behavior generation. In *European Conference on Computer Vision*, pages 484–503. Springer, 2024. 1, 2

[38] D Ulyanov. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[39] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 4, 5

[40] Gerben A Van Kleef. How emotions regulate social life: The emotions as social information (easi) model. *Current directions in psychological science*, 18(3):184–188, 2009. 3

[41] Duomin Wang, Bin Dai, Yu Deng, and Baoyuan Wang. Agentavatar: Disentangling planning, driving and rendering for photorealistic avatar agents. *arXiv preprint arXiv:2311.17465*, 2023. 2

[42] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37:128374–128395, 2024. 2

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2

[44] Zhenhua Wu, Linxuan Jiang, Xiang Li, Chaowei Fang, Yipeng Qin, and Guanbin Li. Hierarchically controlled deformable 3d gaussians for talking head synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 2

[45] Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, et al. X-vila: Cross-modality alignment for large language model. *arXiv preprint arXiv:2405.19335*, 2024. 2

[46] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[47] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[48] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 2

[49] Fukun Yin, Xin Chen, Chi Zhang, Biao Jiang, Zibo Zhao, Wen Liu, Gang Yu, and Tao Chen. Shapegpt: 3d shape generation with a unified multi-modal language model. *IEEE Transactions on Multimedia*, 2025. 2

[50] Jun Yu and Chang Wen Chen. From talking head to singing head: a significant enhancement for more natural human computer interaction. In *2017 IEEE International Conference on Multimedia and Expo*, pages 511–516. IEEE, 2017. 1

[51] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024. 2

[52] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023. 2, 3, 5, 7

[53] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 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*, pages 3661–3670, 2021. 1

[54] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 2

[55] 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. 1, 2

[56] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, and Tiejun Zhao. Interactive conversational head generation. *arXiv preprint arXiv:2307.02090*, 2023. 1