Human-Preference Aligned Listener Facial Expression Generation in Natural Dyadic Interaction

Xu Chen*
Rui Gao*
{chenxu, gaorui}@bit.edu.cn
Beijing Institute of Technology
Beijing, China
Shenzhen MSU-BIT University
Shenzhen, China

Che Sun✉
sunche@smbu.edu.cn
The Guangdong Laboratory of Machine Perception and Intelligent Computing
Shenzhen MSU-BIT University
Shenzhen, China

Xinjie Zhang
Haoyu Zhang
{zhangxinjie, zhanghaoyu}@bit.edu.cn
Beijing Institute of Technology
Beijing, China
Shenzhen MSU-BIT University
Shenzhen, China

Zhi Gao
Yuwei Wu✉
{gaozhibit, wuyuwei}@bit.edu.cn
Beijing Institute of Technology
Beijing, China
Shenzhen MSU-BIT University
Shenzhen, China

Yunde Jia
jiayunde@smbu.edu.cn
The Guangdong Laboratory of Machine Perception and Intelligent Computing
Shenzhen MSU-BIT University
Shenzhen, China

Abstract

Listener facial expression generation in natural dyadic interaction requires responses that are not only realistic, but also consistent with social norms and emotional expectations, which are often reflected in human preferences. However, human-preference alignment remains largely underexplored in listener facial expression generation. In this paper, we propose a human-preference aligned listener facial expression generation method by leveraging human feedback to produce emotionally and socially appropriate expressions for natural dyadic interaction. A key idea of our method is to frame the generation of listener facial expressions as an action learning problem, allowing human feedback to assess the quality of responses without being confounded by the listener's appearance. Specifically, we first train a vision-language-action model via supervised fine-tuning to map the speaker's multimodal cues to low-dimensional facial actions. We then construct a human preference dataset by collecting annotations over sampled listener responses, and further introduce a human-feedback reinforcement learning strategy to optimize the model via direct preference optimization (DPO). Experiments on the L2L-Trevor and RealTalk benchmarks show that our method improves emotional alignment and social appropriateness while maintaining competitive motion quality.

CCS Concepts

• **Human-centered computing** → **Interaction paradigms**; • **Computing methodologies** → *Reinforcement learning*; • **Social and professional topics** → User characteristics.

*Both authors contributed equally to this research.

MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835359

Keywords

Listener Facial Expression Generation, Dyadic Interaction, Human Preference Alignment

ACM Reference Format:
Xu Chen, Rui Gao, Che Sun, Xinjie Zhang, Haoyu Zhang, Zhi Gao, Yuwei Wu, and Yunde Jia. 2026. Human-Preference Aligned Listener Facial Expression Generation in Natural Dyadic Interaction. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835359

1 Introduction

Listener facial expression generation in dyadic interaction aims to generate the listener's facial responses based on the speaker's multimodal cues, such as speech, language, and visual dynamics, to enable more human-like expression interactions, which has drawn

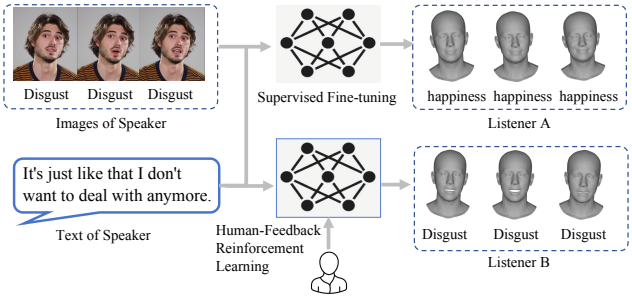

Figure 1: Examples of facial expression generation with (Listener B) and without (Listener A) human preference alignment. Listener B, aligned via human feedback, generates a disgust expression consistent with the Speaker's emotion, whereas the unaligned Listener A generates an inconsistent happiness expression.

increasing attention in human–computer interaction [25, 49]. Recent advances [5, 14, 28, 49] in dyadic facial expression generation leverage deep generative models [15], such as diffusion networks [4, 13, 26] and generative adversarial networks [9], to generate listeners' facial expressions conditioned on the speaker's multimodal cues, and can produce realistic listener behaviors. However, natural listener facial expression generation requires more than visual realism. Appropriate listener responses should also conform to social norms and emotional expectations, thereby aligning with human preferences. This alignment is crucial, as evidence from both social psychology [38] and conversation analysis [12] suggests that expressions misaligned with contextual or social norms (e.g., inappropriate emotion or disregard for "preference organization" in turns) can disrupt interactional flow and reduce user acceptance. A typical example of this limitation can be seen in Listener A in Figure 1, where the listener generates a happiness expression while the speaker conveys disgust, likely leading to conversational dissonance and a breakdown in social rapport. In this paper, we focus on listener facial expression generation aligned with human preference, aiming to ensure that listener expressions are emotionally appropriate and socially compatible with human expectations.

A natural way to achieve alignment with human preference is to incorporate human feedback into the learning process, allowing the model to move beyond merely imitating observed behaviors toward generating expressions that better reflect human judgment. However, directly using the feedback to generate facial expression in dyadic interaction is non-trivial. First, generated expressions are often entangled with identity and appearance, making it challenging to obtain unbiased human feedback that truly reflects the expression quality. This entanglement causes human raters to confuse visual realism or appeal with the quality of expressions. Consequently, the human feedback would become sensitive to appearance rather than to the objective expressions, limiting its reliability as a learning signal for human-preference alignment. Second, even with reliable annotations, it remains difficult to transform such feedback into an effective supervision signal for sequential facial expression generation, where appropriate responses must continuously adapt to the speaker's evolving multimodal behaviors and conversational context.

In this paper, we propose a novel two-stage method of listener facial expression generation aligned with human preference by leveraging human feedback to produce contextually and emotionally appropriate listener expressions for natural dyadic interaction. Our key idea is to formulate listener facial expression generation as an action learning problem, where facial parameters are predicted in a facial action space rather than assessed directly from the listener's visual appearance. This formulation helps human feedback focus more on expression quality and social appropriateness, thereby reducing the influence of listener appearance on preference judgments. Specifically, we use a vision–language–action model and use a supervised fine-tuning strategy to map the speaker's facial motion and language cues into controllable low-dimensional expression actions represented on a 3D morphable model. We further introduce a human-feedback reinforcement learning strategy to integrate imitation of desirable expressions with human-feedback-guided refinement, enabling iterative optimization of alignment quality. By introducing human-feedback reinforcement learning, our listener (exemplified by Listener B in Figure 1) generates aligned expressions, such as generating disgust expressions in response to the speaker's disgust emotion, which significantly improves interactional coherence and perceived social appropriateness. As demonstrated in the bottom row of Figure 1, our method achieves superior alignment with the speaker's "disgust" cues, fostering a more natural and socially resonant dyadic interaction.

We evaluate our method by generating listener expressions on two widely used datasets: L2L-Trevor [28], and RealTalk [8]. Experimental results show that our method consistently outperforms strong baselines in terms of emotional alignment and social appropriateness. In particular, while the supervised fine-tuning stage provides strong motion fidelity, the subsequent human-feedback optimization further improves affective metrics and perceptual quality, demonstrating that preference-driven learning effectively shifts the optimization target from simple geometric reconstruction to socially and emotionally appropriate listener expression generation. These findings are further supported by user studies, where our method achieves the highest ratings across the evaluated metrics.

The main contributions of this work are summarized as follows:

- We introduce human-preference alignment into listener facial expression generation for natural dyadic interaction, aiming to generate listener responses that are more socially and emotionally appropriate.
- We propose a method that formulates listener facial expression generation as an action learning problem to reduce the influence of listener appearance, and further combines supervised fine-tuning with human-feedback preference optimization based on DPO to improve emotional alignment and social appropriateness.

2 Related Work

2.1 3D Talking Head Generation

Existing 3D talking head generation methods have primarily focused on single/dyadic-role and audio-driven facial animation synthesis. Early works focus on lip-synchronized facial animations aligning to the speaker's audio [2, 11]. Recent works such as FaceFormer [7] and CodeTalker [41] introduce the Transformer architectures to capture long-term dependency of multimodal information for generating facial meshes with geometric structure, achieving high-precision lip synchronization and realistic 3D talking head generation. To further improve generation quality and efficiency, recent research has made improvements from both architectural and training strategy perspectives, such as SadTalker [46], TalkingMachines [24], and so on. Besides, MultiTalk [37] constructs a cross-lingual dataset and improves the accuracy of mouth shape synchronization under different language inputs by introducing multilingual style embeddings. ConsistentAvatar [43] introduces a diffusion-based neural render to generate temporal, 3D-aware, and expression-consistent talking head avatars. DualTalk [32] extends audio-driven 3D talking head generation to dyadic interaction modeling.

Prior to our work, Avatar Forcing [18] introduces preference alignment into 3D talking head generation, enhancing expressiveness through direct preference optimization (DPO). However, its alignment is constrained by reliance on synthetic proxy data rather

than real human feedback, which limits its ability to capture the subtle and complex social norms as well as emotional expectations in human-to-human interaction. In contrast, our work focuses on listener facial expression generation in dyadic interaction and constructs preference supervision directly from real human feedback annotations. By formulating listener facial expression generation as an action learning problem, our method reduces the influence of listener appearance on human feedback and enables preference optimization toward socially and emotionally appropriate listener behaviors.

2.2 Listener Modeling

The goal of listener modeling is to generate listeners' non-verbal feedback (such as nodding, smiling) based on the speaker's multimodal inputs, or to achieve dynamic switching between both roles. In the listener feedback generation task, L2L [29] first uses VQ-VAE to discretely encode head motion sequences, generating non-deterministic listening reactions. Subsequently, RLHG [48] proposes the ViCo dataset and establishes a baseline model based on sequential decoding. Further, ELP [36] and CustomListener [23] enhance reaction diversity and controllability from the perspectives of emotional priors and text control, respectively. In the domain of dyadic interaction modeling, some prominent methods have emerged, e.g., DIM [40] and INFP [49], etc. DIM [40] generates non-deterministic facial motions in dyadic interactions by utilizing a VQ-VAE to encode actor motions into discrete latent representations. INFP [49] introduces an audio-driven interactive head generation framework that operates through a two-stage pipeline.

Existing listener modeling methods (e.g., L2L [29], UniFaRN [21], REA-Listener [47], DIM [40], MMLHG [19]) have achieved remarkable progress. Most of them rely on imitation learning from real-world datasets and treat all samples in a dataset as equally "correct", failing to distinguish between high-quality social interactions and neutral or inattentive listener states. In contrast, our method explicitly learns from human-preference feedback to generate listener expressions that are not only realistic but also socially and emotionally appropriate.

2.3 Preference Alignment in Generative Tasks

Preference alignment has emerged as a key problem for steering generative models toward outputs that better reflect human judgment, and has achieved foundational success in natural language processing [3, 34, 35]. The study of preference alignment has subsequently been extended to other modalities. In audio and speech synthesis, methods such as BATON [22] and SpeechAlign [45] have demonstrated that incorporating human-centric utility metrics can enhance prosodic quality and emotional expressiveness beyond the limits of conventional maximum-likelihood training. A similar trend is observed in the vision-language domain, where models like SHAPE [17] employ preference-based tuning to improve the relevance and coherence of generated descriptions.

These advances underscore the effectiveness of preference-driven learning across modalities, while the alignment of human preferences in facial expression generation remains unexplored. A core challenge is to obtain human feedback that reliably reflects expression quality. Our method addresses this challenge by formulating listener facial expression generation as an action learning problem in an action space, thereby reducing the influence of listener appearance on human feedback and enabling preference optimization toward socially and emotionally appropriate responses.

3 Method

3.1 Problem Formulation

We formulate listener facial expression generation in dyadic interaction as a sequential action learning problem. Given a sequence of the speaker's multimodal inputs, denoted as $\mathbf{S}_{1:t} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_t\}$, where each state $\mathbf{s}_i = (\mathbf{I}_i, \mathbf{L}_i)$ comprises visual frames $\mathbf{I}_i$ and language content $\mathbf{L}_i$, our goal is to learn a policy π_θ that generates appropriate listener facial parameters $\mathbf{A}_t \sim \pi_\theta(\cdot|\mathbf{S}_{1:t})$ at each time step.

The facial parameters $\mathbf{A}_t = [\mathbf{a}_t^{\text{exp}}; \mathbf{a}_t^{\text{pose}}]$ consist of expression coefficients $\mathbf{a}_t^{\text{exp}} \in \mathbb{R}^{D^{\text{exp}}}$ and head pose parameters $\mathbf{a}_t^{\text{pose}} \in \mathbb{R}^{D^{\text{pose}}}$, and they are used to render the 3D face mesh $\mathcal{M}_t$ through the FLAME model [20] while maintaining the fixed identity parameters $\mathbf{a}^{\text{shape}}$, given by

$$\mathcal{M}_t = \text{FLAME}(\mathbf{a}^{\text{shape}}, \mathbf{a}_t^{\text{exp}}, \mathbf{a}_t^{\text{pose}}). \tag{1}$$

This formulation directly serves our goal of human-preference alignment. By defining the listener's response as a policy output in a sequential interaction, our method is suitable for feedback-driven learning, allowing the model to adapt based on human judgments. Moreover, the use of identity-agnostic facial parameters as actions focuses the policy on learning transferable expression dynamics and establishes a disentangled space for collecting human feedback that targets social appropriateness rather than visual appearance.

3.2 Overview of the Method

We propose a facial expression generation method aligned with human preferences. As illustrated in Figure 2, our method operates in two stages. In the first stage, the vision-language-action (VLA) model takes the speaker's images $\{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}$ and the corresponding text content $\{\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_T\}$ as inputs to predict the listener's facial expression actions $\mathbf{A}_t = [\mathbf{a}_t^{\text{exp}}; \mathbf{a}_t^{\text{pose}}]$. During this stage, the model is trained via Supervised Fine-Tuning (SFT) to imitate the ground-truth listener actions, thereby acquiring a foundational capability for facial expression synthesis. In the second stage, we introduce a Human-Feedback Reinforcement Learning strategy. We utilize the policy (i.e., the VLA model trained in the SFT stage) to sample N candidate listener actions $\{\mathbf{A}_{1:T}^1, ..., \mathbf{A}_{1:T}^N\}$. Subsequently, these sampled actions are rendered into visual listener responses to facilitate human assessment. These generated responses, together with the ground-truth (GT) response $\mathbf{A}_{1:T}^*$, are then evaluated and ranked by human annotators. According to these rankings, we designate the highest-scored response as the human-preferred (Pre) sample and the lowest-scored response as the human-dispreferred (Dispre) sample. Finally, we employ the Direct Preference Optimization (DPO) algorithm to optimize the policy using these selected preference pairs.

3.3 Vision-Language-Action Model

Our vision-language-action model consists of three key components: multimodal input encoding, action de-tokenizer, and a large

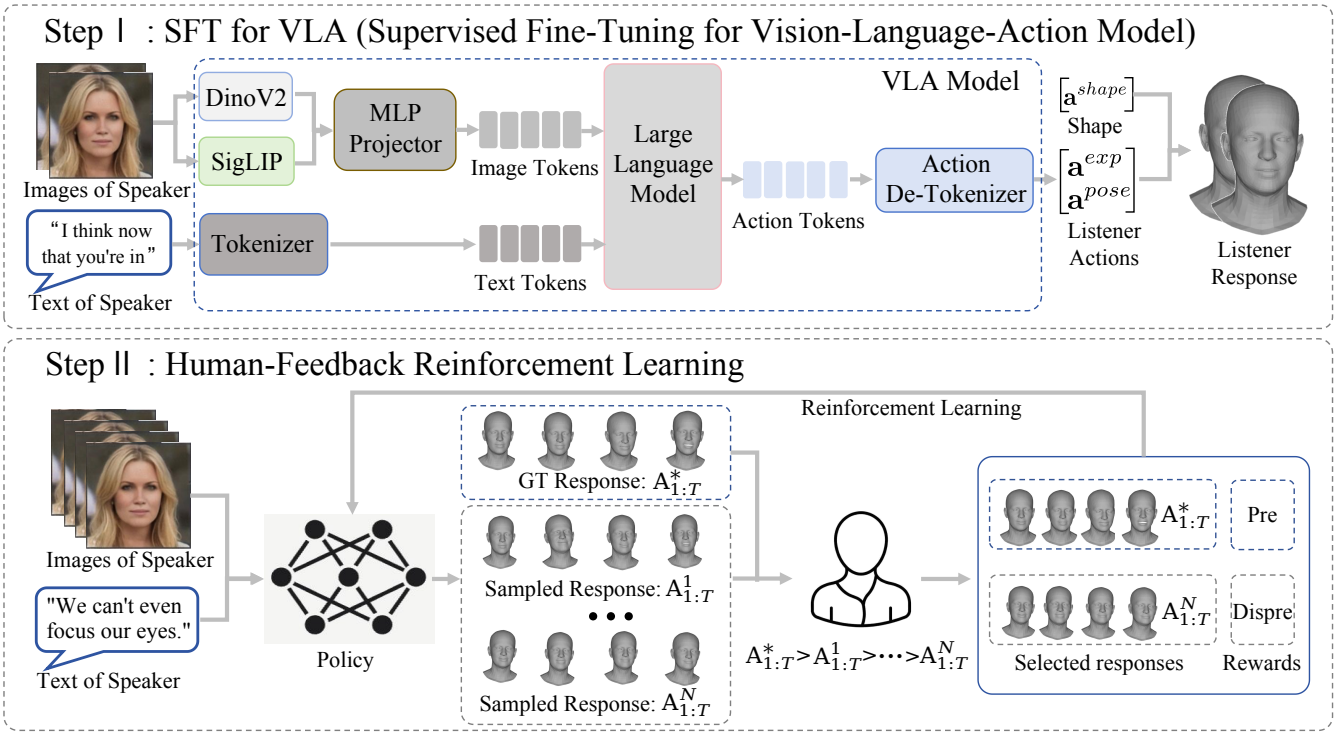

Figure 2: Overall framework of our method.

language model (LLM) backbone. In this work, we use the 7B-parameter LLaMA 2 language model as the backbone [39].

3.3.1 Multimodal Input Encoding. Given a video sequence of the speaker containing T frames $\{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}$, we first extract comprehensive visual features. These features are fed into a projector to map them into the input space of the LLM, yielding a sequence of image tokens.

To capture both fine-grained facial dynamics and global semantic context, we employ a dual-stream visual encoder. For each frame $\mathbf{I}_t$, we extract visual features using pre-trained DINO [30] and SigLIP [44] models, given by

$$\mathbf{v}_t^{\text{dino}} = \text{DINO}(\mathbf{I}_t), \quad \mathbf{v}_t^{\text{siglip}} = \text{SigLIP}(\mathbf{I}_t), \tag{2}$$

where $\mathbf{v}_t^{\text{dino}} \in \mathbb{R}^{d_v}$ captures pose and subtle expression details, while $\mathbf{v}_t^{\text{siglip}} \in \mathbb{R}^{d_v}$ encodes global affective semantics and social cues. The motivation for adopting this dual-branch encoder is that effective listener response necessitates both the capture of fine-grained facial details and head poses, as well as a high-level interpretation of emotional context.

The visual features are then projected and aggregated into image tokens $\mathbf{F}^{\text{vision}}$ through a multi-layer perceptron projector MLP^{vis} to map them into the input space of the LLM, given by

$$\mathbf{F}^{\text{vision}} = \{\text{MLP}^{\text{vis}}([\mathbf{v}_t^{\text{dino}}; \mathbf{v}_t^{\text{siglip}}])\}_{t=1}^{T}, \tag{3}$$

where $[\cdot;\cdot]$ denotes concatenation.

The text content $\{\mathbf{L}_t\}_1^T$ is tokenized into text tokens $\mathbf{F}^{\text{text}}$ through the LLaMA [39] tokenizer,

$$\mathbf{F}^{\text{text}} = \{\text{LLaMATokenizer}(\mathbf{L}_t)\}_{t=1}^{T}. \tag{4}$$

Subsequently, these text tokens are concatenated with image tokens to serve as the unified multimodal input.

3.3.2 Action De-Tokenizer. To align continuous facial motions with the discrete output space of the LLM, we implement a quantization strategy that maps continuous parameters into discrete tokens. Inspired by the work RT-2 [1], we discretize each dimension of the facial actions into 256 bins. To determine the optimal quantization boundaries, we first sort the action values from the training corpus, and truncate the top and bottom 1% to eliminate statistical outliers. The bins are then uniformly distributed over the remaining effective value range. This strategy effectively filters out noise and ensures that the limited representational capacity (i.e., the 256 bins) is concentrated within the valid motion interval, thereby significantly enhancing the modeling granularity for subtle micro-expressions and head poses. Finally, the discrete action tokens $\mathbf{F}^{\text{action}}$ predicted by the LLM are computed by

$$\mathbf{F}^{\text{action}} = \text{LLM}([\mathbf{F}^{\text{vision}}; \mathbf{F}^{\text{text}}]). \tag{5}$$

$\mathbf{F}^{\text{action}}$ is mapped back to the continuous action $\mathbf{A}_{1:T}$ to synthesize the listener's response via

$$\mathbf{A}_{1:T} = \text{ActionDe-Tokenizer}(\mathbf{F}^{\text{action}}). \tag{6}$$

3.3.3 Supervised Fine-Tuning. The first stage employs supervised fine-tuning to establish a reliable VLA mapping from multimodal inputs to facial actions. Given a dataset $\mathcal{D}_{\text{SFT}} = \{(\mathbf{I}_{1:T}, \mathbf{L}_{1:T}, \mathbf{A}^*_{1:T})^i\}_{i=1}^N$ containing aligned sequences of visual inputs $\mathbf{I}_{1:T}$, text inputs $\mathbf{L}_{1:T}$, and ground-truth facial actions $\mathbf{A}^*_{1:T} = \{\mathbf{a}^{\text{exp}^*}_{1:T}, \mathbf{a}^{\text{pose}^*}_{1:T}\}$, we use cross-entropy loss $\mathcal{L}_{\text{cross}}$ to optimize the VLA parameters θ

$$\begin{aligned}\mathcal{L}_{\text{pre}} = \frac{\lambda_{\text{exp}}}{T}\sum_{t=1}^{T}\mathcal{L}_{\text{cross}}(\mathbf{a}_t^{\text{exp}^*}, \mathbf{a}_t^{\text{exp}}) + \\ \frac{\lambda_{\text{pose}}}{T}\sum_{t=1}^{T}\mathcal{L}_{\text{cross}}(\mathbf{a}_t^{\text{pose}^*}, \mathbf{a}_t^{\text{pose}})),\end{aligned} \tag{7}$$

where λ_{exp} and λ_{pose} balance the contribution of expression and pose terms. To ensure temporal coherence and identity consistency, we incorporate additional regularization terms as

$$\mathcal{L}_{\text{temp}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T-1}\sum_{t=1}^{T-1}(\|\mathbf{a}_{t+1}^{\text{exp}} - \mathbf{a}_t^{\text{exp}}\|_2^2 + \|\mathbf{a}_{t+1}^{\text{pose}} - \mathbf{a}_t^{\text{pose}}\|_2^2). \tag{8}$$

We yield the overall loss function

$$\theta^*_{\text{SFT}} = \arg\min_{\theta}\left[\mathcal{L}_{\text{pre}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}\right], \tag{9}$$

where λ_{temp} is the trade-off parameter. The model trained through Eq. (9) serves as a strong initial policy $\pi_{\theta^*_{\text{SFT}}}$ that produces visually coherent and identity-consistent facial responses, providing a solid foundation for subsequent human-feedback optimization.

3.4 Human-Feedback Reinforcement Learning

3.4.1 Human Feedback Dataset Construction. To support human-preference alignment for listener facial expression generation, we develop a human feedback data construction pipeline that converts multi-dimensional human ratings into high-confidence preference pairs for downstream optimization. Specifically, we first use the initial policy $\pi_{\theta^*_{\text{SFT}}}$ to sample $N = 4$ candidate listener actions $\{\mathbf{A}^j_{1:T}\}_{j=1}^N$ for each speaker multimodal input $S_{1:T}$. This yields a set of candidate trajectories $\{(\mathbf{S}_{1:T}, \mathbf{A}^j_{1:T})\}_{j=1}^N$. To provide an in-context human reference within each group, we additionally include the ground-truth action sequence $\mathbf{A}^*_{1:T}$, forming a candidate group

$$\mathcal{G}_{cand} = \{(\mathbf{S}_{1:T}, \mathbf{A}^1_{1:T}), \ldots, (\mathbf{S}_{1:T}, \mathbf{A}^N_{1:T}), (\mathbf{S}_{1:T}, \mathbf{A}^*_{1:T})\}. \tag{10}$$

We then render the trajectories $(\mathbf{S}_{1:T}, \mathbf{A}^j_{1:T})$ into the speaker-listener interaction videos τ^j, so that annotators can assess the listener's facial behavior in the full dyadic context rather than from isolated motion signals alone.

To obtain reliable human preference annotations, we adopt a rigorous annotation protocol that operationalizes human preference using four dimensions: Empathy (Emp.), Appropriateness (App.), Engagement (Eng.), and Naturalness (Nat.). We ask $M = 30$ annotators, all of whom are undergraduate students or above and have been carefully instructed on the annotation protocol, to evaluate the rendered speaker-listener videos. Each candidate group is independently evaluated by all M annotators to obtain human feedback that is more representative of general human preference and less sensitive to individual variation. During annotation, the candidate videos within each group are presented in random order, and their source identities are hidden from the annotators. Annotators are allowed to replay the videos when necessary before assigning scores. These dimensions are motivated by prior work on empathy in interactive agents [42], socially appropriate affective responses in context [27], engagement as a key factor in conversational interaction quality [31], and naturalness as an important perceptual criterion for facial expression animation [10]. Together, these dimensions capture whether a listener response is affectively attuned to the speaker, socially appropriate to the conversational context, interactionally involving, and perceptually plausible. For more detailed information on each dimension, please refer to the *supplementary materials*.

Annotators evaluate each candidate video τ^j along these four dimensions. For the m-th annotator, the preference score of each candidate video τ^j is computed by

$$\begin{aligned}r^m(\tau^j) = &\alpha_{\text{emp}}\text{Empathy}^m(\tau^j) + \alpha_{\text{app}}\text{Appropriateness}^m(\tau^j) \\ &+ \alpha_{\text{eng}}\text{Engagement}^m(\tau^j) + \alpha_{\text{nat}}\text{Naturalness}^m(\tau^j).\end{aligned} \tag{11}$$

The weighting coefficients $\alpha_{\text{emp}}, \alpha_{\text{app}}, \alpha_{\text{eng}}, \alpha_{\text{nat}} \in [0, 1]$ balance the relative importance, with $\sum \alpha = 1$ ensuring normalized rating. We set the four weights equally in our experiments. Since each candidate group is evaluated by all 30 annotators, we obtain the final preference score of each candidate video by averaging the scores across annotators

$$\bar{r}(\tau^j) = \frac{1}{M}\sum_{m=1}^{M} r^m(\tau^j). \tag{12}$$

In this way, the final score reflects a more broadly shared and stable human preference rather than an individual subjective judgment. We obtain a preference score for each trajectory within every candidate group

$$\mathcal{G} = \{(\mathbf{S}_{1:T}, \mathbf{A}^1_{1:T}, \bar{r}^1), \ldots, (\mathbf{S}_{1:T}, \mathbf{A}^N_{1:T}, \bar{r}^N), (\mathbf{S}_{1:T}, \mathbf{A}^*_{1:T}, \bar{r}^*)\}. \tag{13}$$

which serves as the final scored candidate group for subsequent preference pair construction.

3.4.2 Preference Pair Construction. Although annotators provide multi-dimensional ratings, our downstream optimization is based on pairwise preference learning rather than absolute score regression. We therefore use the aggregated scores in Eq. 11 to rank the candidates within each group and convert the ranked candidates into binary preference pairs for subsequent optimization. Specifically, we designate the highest-ranked trajectory as the preferred response ($\mathbf{A}^w_{1:T}$) and the lowest-ranked trajectory as the dispreferred response ($\mathbf{A}^l_{1:T}$). This selection strategy yields a large preference margin and thus provides a higher-confidence supervision signal for preference alignment than comparisons between closely ranked candidates. Intuitively, the preferred response $\mathbf{A}^w_{1:T}$ exhibits stronger emotional attunement, better contextual appropriateness, higher interactional engagement, and greater perceptual naturalness than the dispreferred response $\mathbf{A}^l_{1:T}$. Subsequently, we construct the final preference dataset

$$\mathcal{D}_{\text{pref}} = \{(\mathbf{S}_{1:T}, \mathbf{A}^w_{1:T}, \mathbf{A}^l_{1:T})^i\}_{i=1}^K. \tag{14}$$

for Direct Preference Optimization (DPO) [33] training, where $\mathbf{A}^w_{1:T}$ and $\mathbf{A}^l_{1:T}$ denote the preferred (Pre) and dispreferred (Dispre) actions, respectively. This dataset serves as the supervision signal for subsequent direct preference optimization.

Table 1: Quantitative comparison on the L2L-Trevor dataset.

Method	L2 ↓	FD ↓	Variation	Diversity	P-FD ↓	L2 Affect(10^2) ↓
GT			0.1254	2.5427		
Random	0.6852	33.5481	0.1458	2.7813	34.5579	12.4864
NN	0.6047	28.4186	0.0914	2.1486	26.7514	10.6842
LM-Listener [29]	0.4345	17.6299	0.1189	2.9374	19.1583	6.3992
DIM [40]	0.3213	12.6872	0.1515	2.0347	24.5681	5.4751
DiffListener [16]	0.4000	15.7500	0.1000	2.9600	17.7500	-
MMLHG [19]	**0.2910**	10.0949	0.0704	2.2960	11.3908	2.6575
Ours (SFT)	0.3015	**9.1473**	0.0814	2.1756	11.2975	2.5724
Ours (SFT+RL)	0.3129	10.2385	0.0919	2.3754	**10.8247**	**2.4842**

Table 2: Quantitative comparison on the RealTalk dataset.

Method	L2 ↓	FD ↓	Variation	Diversity	P-FD ↓	L2 Affect(10^2) ↓
GT			0.0478	1.7250		
Random	0.3749	17.2541	0.0549	1.9452	18.4436	24.1739
NN	0.3147	16.0572	0.0496	1.6241	15.9657	20.1478
LM-Listener [29]	0.2416	10.8423	0.0402	1.6148	10.5483	12.2730
MMLHG [19]	0.1021	3.7914	0.0234	1.2643	3.8145	6.0427
Ours (SFT)	**0.0824**	**3.2425**	0.0371	0.9142	3.8036	4.5207
Ours (SFT+RL)	0.0973	3.5842	0.0421	1.2457	**3.7914**	**4.3531**

3.4.3 Direct Preference Optimization. We optimize the policy π_θ using Direct Preference Optimization (DPO). We initialize the policy π_θ with the supervised fine-tuned model $\pi_{\text{ref}} = \pi_{\theta^*_{\text{SFT}}}$ and freeze a copy of π_{ref} to serve as the reference policy. The loss function is defined as

$$\mathcal{L}_{\text{policy}} = -\mathbb{E}\Bigg[\log\sigma\Bigg(\beta\log\frac{\pi_\theta(\mathbf{A}^w_{1:T} \mid \mathrm{S}_{1:T})}{\pi_{\text{ref}}(\mathbf{A}^w_{1:T} \mid \mathrm{S}_{1:T})} - \beta\log\frac{\pi_\theta(\mathbf{A}^l_{1:T} \mid \mathrm{S}_{1:T})}{\pi_{\text{ref}}(\mathbf{A}^l_{1:T} \mid \mathrm{S}_{1:T})}\Bigg)\Bigg], \tag{15}$$

where σ denotes the sigmoid function, and β is a hyperparameter controlling the strength of the KL-divergence constraint. In this way, the model is encouraged to generate listener facial expressions that are more consistent with human preferences while remaining close to the supervised initialization.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets. We evaluate our method on two interaction datasets: *L2L-Trevor* [28], and *RealTalk* [8]. L2L-Trevor is a single-listener dataset originally introduced in the work [28]. RealTalk dataset is a more extensive corpus involving diverse speaker-listener pairs, and comprises 692 dialogue videos, totaling 115 hours of conversational interaction. These datasets collectively enable thorough evaluation of facial expression generation across different interaction scenarios. More details on the datasets preprocessing procedures and implementation details are provided in the *Supplementary Material.*

4.1.2 Baseline Methods. Following the evaluation protocol in [19], we primarily compare our method with six baselines: Random, Nearest Neighbor (NN), LM-Listener [29], DiffListener [16], DIM [40], and MMLHG [19]. Random and NN serve as simple non-learning baselines. LM-Listener is a language-model-based listener generation method. DiffListener is a diffusion-based approach for listener facial behavior generation. DIM adopts an encoder-decoder framework trained with contrastive learning to model speaker-listener interactions. MMLHG is the current state-of-the-art framework that synthesizes listener reactions by fusing conversational content with the speaker's facial FLAME parameters.

4.1.3 Evaluation Metrics. We employ comprehensive metrics to assess different aspects of generated facial expressions. The *L2 distance* quantifies reconstruction accuracy: $\text{L2} = \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{A}_t - \mathbf{A}^*_t\|_2^2$. The *Fréchet Distance (FD)* measures distribution similarity between generated and real facial motions. The *Paired FD (P-FD)* evaluates motion quality while preserving temporal alignment: $\text{P-FD} = \frac{1}{N}\sum_{i=1}^{N}\text{FD}(\mathbf{A}^*_i, \mathbf{A}_i)$. The *L2 Affect for synchrony (L2 Affect)* quantifies the accuracy of the predicted listener emotional affect across the sequence: $\text{L2 Affect} = \frac{1}{N}\sum_{i=1}^{N}(\bar{v}_i - \bar{v}_i^{GT})^2$, where v denotes the facial emotional affect for the i-th frame, as estimated by the pretrained EMOCA [6] model. The Variation and Diversity quantify the richness and dynamic properties of the generated facial actions.

4.2 Quantitative Evaluation

Table 1 and Table 2 present the comprehensive quantitative results across the two datasets, respectively. The corresponding L2-affect value of DiffListener is denoted by "–" because this metric is not

Figure 3: Qualitative comparison of generated listener expressions on the RealTalk dataset. Our method produces more appropriate and emotionally aligned responses compared to baseline methods.

reported in the original work. In Table 2, we re-implement LM-Listener and MMLHG [19] on the RealTalk [8] dataset to ensure a fair comparison. We do not further reproduce DIM and DiffListener on RealTalk, because these two methods were not originally tested on this dataset, while LM-Listener and MMLHG are the most relevant baselines for comparison with our method. In particular, MMLHG is the more recent and stronger multimodal baseline among existing methods.

Overall, our method demonstrates superior performance compared to existing methods. Specifically, our SFT model (i.e., our VLA model trained via supervised fine-tuning) excels in motion fidelity, achieving the lowest FD scores of 9.1473 (vs. MMLHG's 10.0949 DIM's 12.6872, and DiffListener's 15.7500) on L2L-Trevor and 3.2425 (vs. MMLHG's 3.7914) on RealTalk. Furthermore, the SFT model attains an L2 error of 0.0824 on RealTalk, significantly outperforming the values of 0.1021 and 0.2416 obtained by MMLHG and LM-Listener methods, respectively, which confirms its precision in capturing authentic facial dynamics. On the L2L-Trevor dataset, the additional baselines DiffListener and DIM achieve performance between LM-Listener and MMLHG, further showing that our method remains superior to a diverse set of existing approaches.

Most importantly, the integration of the reinforcement learning stage (i.e., SFT+RL) drives substantial improvements in semantic and emotional alignment. As shown in Table 2, our SFT+RL model achieves the best L2 Affect score of 4.3531, surpassing both the SFT baseline (4.5207) and MMLHG (6.0427) by a wide margin. Similarly, on the L2L-Trevor dataset shown in Table 1, our SFT+RL model records the lowest P-FD of 10.8247, improving upon the SFT model's 11.2975 and outperforming LM-Listener, DiffListener, DIM, and MMLHG (with corresponding P-FD scores of 19.1583, 17.7500, 24.5681, and 11.3908, respectively). While the RL model exhibits a marginal increase in FD (e.g., 3.5842 on RealTalk and 10.2385 on L2L-Trevor) indicating larger reconstruction errors, its superior performance in affective metrics validates that our reinforcement learning strategy effectively prioritizes social appropriateness and emotional consistency over simple geometric reconstruction.

4.3 Qualitative Evaluation

We visualize the listener responses generated by our method and baselines alongside the ground truth in Figure 3. We compare our method with LM-Listener [29] and MMLHG [19] in the qualitative evaluation, as these two methods are the most relevant baselines to our setting and have also been reproduced on the RealTalk benchmark for a fair comparison. As observed in the left column, when the speaker expresses joy while discussing "biggest surprises", the LM-Listener [29] fails to react, producing a neutral and apathetic face. MMLHG [19] captures some positive sentiment, while our method also generates a more natural and intense smile, demonstrating superior emotional synchrony with the speaker's state.

The middle column in Figure 3 highlights the advantage of our preference alignment strategy. The speaker discusses a serious topic ("weren't sober"), necessitating a somber or attentive listener reaction. MMLHG exhibits a "hallucinated" positive emotion, generating an inappropriate smile that violates social norms. In contrast, our model correctly interprets the context and synthesizes a serious, attentive expression. As seen in the right column, while our model aligns well with the speaker's emotion, LM-Listener fails to produce emotionally consistent expressions. These results confirm that our method not only mimics facial motions but also understands social appropriateness, avoiding the "generic positivity" bias often found in SFT-based models.

4.4 User Study

To evaluate the perceptual quality and social appropriateness of the generated listener responses, we conduct a comprehensive user study involving 25 participants (aged 19-31, all university students or graduates). Each participant is presented with a series of videos. Each contains a speaker's video along with four corresponding listener responses generated by the two baseline models (LM-Listener and MMLHG), our SFT model, and our SFT+RL model. Participants are asked to rate the responses across four dimensions of

Table 3: User study results (Mean Opinion Score, 1-5 scale). Our method achieves the highest ratings across all aspects.

Method	App.	Emp.	Eng.	Nat.
LM-Listener [29]	2.7	3.1	3.4	2.9
MMLHG [19]	3.0	3.3	3.5	3.1
Ours (SFT)	3.2	3.4	3.7	3.3
Ours (SFT+RL)	**4.5**	**4.1**	**4.2**	**4.5**

Table 4: Ablation study on the RealTalk dataset.

Method	L2 ↓	FD ↓	P-FD ↓	L2 Affect(10^2) ↓
Full (Ours)	0.0973	3.5842	**3.7914**	**4.3531**
Random-Prefer	0.3142	12.3549	12.0354	15.2463
SFT-Preferred	0.1132	3.6791	3.9165	5.519
SFT-Only	**0.0824**	**3.2425**	3.8036	4.5207

Appropriateness (App.), Empathy (Emp.), Engagement (Eng.), and Naturalness (Nat.).

Ratings were assigned using a 5-point Likert scale, ranging from 1 (lowest quality) to 5 (highest quality). Upon the completion of the annotation process, we calculated the Mean Opinion Score (MOS) for each model across all test samples to derive the final evaluation results. The average score for a specific dimension $k \in \{\text{App.}, \text{Emp.}, \text{Eng.}, \text{Nat.}\}$ is calculated as

$$\text{MOS} = \frac{1}{N}\sum_{i=1}^{N} \text{score}_k(\tau^i). \tag{16}$$

where N denotes the total number of evaluated video samples and τ denotes the evaluation video. More details of the user study protocol are provided in the *Supplementary Material*.

The quantitative results are summarized in Table 3. Our SFT+RL method achieves significantly higher scores across all categories. Notably, the RL stage yields substantial gains over the SFT baseline, boosting Appropriateness from 3.2 to 4.5 and Empathy from 3.4 to 4.1. Furthermore, our method outperforms the strongest baseline MMLHG by a wide margin (e.g., +1.5 in Appropriateness and +0.8 in Empathy). These results confirm that our reinforcement learning strategy, guided by human feedback, effectively aligns the generated facial expressions with human social preferences, ensuring they are not only natural but also socially adept and empathetic.

4.5 Ablation Study

We validate our design choices by comparing the full model against three variants: (1) SFT-Only, where the stage-one model is used directly; (2) Random-Prefer, where DPO is used with random preference labels; and (3) SFT-Preferred, which uses standard SFT on the labeled trajectories with human preference. These variants are designed to examine the necessity of preference optimization, the importance of accurate human feedback, and the advantage of human preference learning over straightforward supervision on preferred samples.

As shown in Table 4, comparing Full (Ours) with SFT-Only reveals a clear trade-off: while SFT-Only yields lower reconstruction errors (L2/FD) by strictly imitating the ground truth trajectories, our Full model achieves superior P-FD and L2 Affect scores.

This result shows that the preference optimization stage effectively shifts the learning objective from geometric reconstruction toward social and emotional alignment. The degraded performance of Random-Prefer further demonstrates that the gains of our method do not come from the DPO procedure alone, but rely on accurate human preference supervision. Besides, our method outperforms SFT-Preferred, demonstrating that the contrastive DPO objective, which distinguishes between preferred and dispreferred behaviors, is more effective than straightforward supervision on positive samples alone. Overall, these ablation results consistently confirm the effectiveness of our full design.

4.6 Discussion and Limitations

While our method achieves strong improvements in human preference alignment, it also reveals a trade-off with visual consistency. In particular, preference optimization improves emotional alignment and social appropriateness, but may slightly reduce geometric fidelity in some cases. A possible reason is that the preference pairs used for training are selected based on human judgments rather than guaranteed inclusion of the ground-truth response.

In addition, the current preference construction strategy does not fully exploit all available human annotations. To obtain more reliable supervision, we only retain the highest-ranked and lowest-ranked responses in each candidate group, leaving a considerable amount of potentially useful annotated data unused. More data-efficient preference learning strategies that exploit intermediate-ranked candidates or the full ranking structure may further improve performance.

5 Conclusion

In this paper, we presented a novel method of listener facial expression generation aligned with human preference for dyadic interactions. Our method can effectively incorporate human feedback to ensure that generated listener expressions are contextually and emotionally congruent with the speaker's cues. Our method frames the generation of listener facial expressions as an action learning problem, which can make the collection of unbiased human feedback focused more on expressive quality and social appropriateness. We introduce a vision-language-action model trained via supervised fine-tuning and refined by a human-feedback reinforcement learning strategy, which can dynamically produce appropriate facial responses. Extensive experiments on two dyadic benchmarks can demonstrate that our method outperforms state-of-the-art methods.

6 Acknowledgments

This work was supported by the Shenzhen Science and Technology Program under Grant No. JCYJ20241202130548062, the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006, and the Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No. 2023ZDZX1034.

References

[1] Anthony Brohan, Noah Brown, Justice Carbajal, and et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint arXiv:2307.15818* (2023).

[2] Yong Cao, Wen C. Tien, Petros Faloutsos, and Frédéric H. Pighin. 2005. Expressive speech-driven facial animation. *ACM Trans. Graph.* 24, 4 (2005), 1283–1302.

[3] Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2026. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. *ACM Comput. Surv.* 58, 2 (2026), 53:1–53:37.

[4] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 9 (2023), 10850–10869.

[5] David Curto, Albert Clapés, Javier Selva, Sorina Smeureanu, Julio Junior, Jacques CS, David Gallardo-Pujol, Georgina Guilera, David Leiva, Thomas B Moeslund, et al. 2021. Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions. In *Int. Conf. Comput. Vis.* 2177–2188.

[6] Radek Danecek, Michael J. Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 20279–20290.

[7] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.* 18749–18758.

[8] Scott Geng, Revant Teotia, Purva Tendulkar, Sachit Menon, and Carl Vondrick. 2023. Affective Faces for Goal-Driven Dyadic Communication. *arXiv preprint arXiv:2301.10939* (2023).

[9] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Adv. Neural Inform. Process. Syst.* 27 (2014).

[10] C. Martin Grewe, Tuo Liu, Christoph Kahl, Andrea Hildebrandt, and Stefan Zachow. 2021. Statistical Learning of Facial Expressions Improves Realism of Animated Avatar Faces. *Frontiers in Virtual Reality* Volume 2 - 2021 (2021). doi:10.3389/frvir.2021.619811

[11] Benjamin Havell, Paul L. Rosin, Saeid Sanei, Andrew J. Aubrey, A. David Marshall, and Yulia Hicks. 2012. Hybrid phoneme based clustering approach for audio driven facial animation. In *Int. Conf. on Acoust., Speech & Signal Process.* 2261–2264.

[12] Cornelia Herbert. 2021. How to Deal with Incongruence? The Role of Social Perception and Bodily Facial Feedback in Emotion Recognition in Human Agent Interaction–Evidence from Psychology as Potential and Challenge for Multimodal User-Centered Approaches. In *Int. Conf. on Practical Appl. of Agents & Multi-Agent Syst.* 28–39.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.* 6840–6851.

[14] Ailin Huang, Zhewei Huang, and Shuchang Zhou. 2022. Perceptual Conversational Head Generation with Regularized Driver and Enhanced Renderer. In *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*.

[15] Jiajian Huang and Zitong Yu. 2025. Multiple Appropriate Facial Reaction Generation Based on Multi-View Transformation of Speaker Video. In *ACM Int. Conf. Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 13992–13996. doi:10.1145/3746027.3762246

[16] Siyeol Jung and Taehwan Kim. 2025. DiffListener: Discrete Diffusion Model for Listener Generation. In *Int. Conf. on Acoust., Speech & Signal Process.* 1–5. doi:10.1109/ICASSP49660.2025.10890674

[17] Chen Kejia, Zhang Jiawen, Hu Jiacong, Luo Jian, Feng Zunlei, and Song Mingli. 2025. SHAPE: Self-Improved Visual Preference Alignment by Iteratively Generating Holistic Winner. *arXiv preprint arXiv:2503.04858* (2025).

[18] Taekyung Ki, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, and Sung Ju Hwang. 2026. Avatar Forcing: Real-Time Interactive Head Avatar Generation for Natural Conversation. *arXiv preprint arXiv:2601.00664* (2026).

[19] Peiwen Lai, Weizhi Zhong, Yipeng Qin, Xiaohang Ren, Baoyuan Wang, and Guanbin Li. 2025. LLM-driven multimodal and multi-Identity listening head generation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 10656–10666.

[20] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194:1–194:17.

[21] Cong Liang, Jiahe Wang, Haofan Zhang, Bing Tang, Junshan Huang, Shangfei Wang, and Xiaoping Chen. 2023. UniFaRN: Unified Transformer for Facial Reaction Generation. In *ACM Int. Conf. Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 9506–9510. doi:10.1145/3581783.3612854

[22] Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li. 2024. BATON: Aligning Text-to-Audio Model with Human Preference Feedback. In *Int. Joint Conf. on Artificial Intelligence*. 4542–4550.

[23] Xi Liu, Ying Guo, Cheng Zhen, Tong Li, Yingying Ao, and Pengfei Yan. 2024. CustomListener: Text-Guided Responsive Interaction for User-Friendly Listening Head Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2415–2424.

[24] Chetwin Low and Weimin Wang. 2025. TalkingMachines: Real-Time Audio-Driven FaceTime-Style Video via Autoregressive Diffusion Models. *arXiv preprint arXiv:2506.03099* (2025).

[25] Cheng Luo, Siyang Song, Weicheng Xie, Micol Spitale, Zongyuan Ge, Linlin Shen, and Hatice Gunes. 2025. ReactFace: Online Multiple Appropriate Facial Reaction Generation in Dyadic Interactions. *IEEE Trans. Vis. Comput. Graph.* 31, 9 (2025), 6190–6207.

[26] Qirong Mao, Qiwei Wu, Na Liu, Yakui Ding, and Lijian Gao. 2025. Scattering-Conditioned Diffusion Models for Multiple Appropriate Facial Reaction Generation. In *ACM Int. Conf. Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 13985–13991. doi:10.1145/3746027.3762245

[27] Youssef Mohamed, Séverin Lemaignan, Arzu Güneysu, Patric Jensfelt, and Christian Smith. 2025. Context Matters: Understanding Socially Appropriate Affective Responses Via Sentence Embeddings. In *Social Robotics: 16th International Conference, ICSR + AI 2024, Odense, Denmark, October 23–26, 2024, Proceedings, Part I* (Odense, Denmark). Springer-Verlag, Berlin, Heidelberg, 78–91. doi:10.1007/978-981-96-3522-1_9

[28] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning to listen: Modeling non-deterministic dyadic facial motion. In *IEEE Conf. Comput. Vis. Pattern Recog.* 20395–20405.

[29] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. 2023. Can Language Models Learn to Listen?. In *Int. Conf. Comput. Vis.* 10049–10059.

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.* 2024 (2024).

[31] Arthur Pellet-Rostaing, Roxane Bertrand, Auriane Boudin, Stéphane Rauzy, and Philippe Blache. 2023. A multimodal approach for modeling engagement in conversation. *Frontiers in Computer Science* 5 (2023), 1062342.

[32] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. 2025. DualTalk: Dual-Speaker Interaction for 3D Talking Head Conversations. In *IEEE Conf. Comput. Vis. Pattern Recog.* 21055–21064.

[33] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Adv. Neural Inform. Process. Syst.*

[34] Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Rémi Munos. 2024. Multi-turn Reinforcement Learning with Preference Human Feedback. In *Adv. Neural Inform. Process. Syst.*, Vol. 37. 118953–118993.

[35] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference Ranking Optimization for Human Alignment. In *AAAI Conf. on Artificial Intelligence*. 18990–18998.

[36] Luchuan Song, Guojun Yin, Zhenchao Jin, Xiaoyi Dong, and Chenliang Xu. 2023. Emotional Listener Portrait: Realistic Listener Motion Simulation in Conversation. In *Int. Conf. Comput. Vis.* 20782–20792.

[37] Kim Sung-Bin, Lee Chae-Yeon, Gihun Son, Oh Hyun-Bin, Janghoon Ju, Suekyeong Nam, and Tae-Hyun Oh. 2024. MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset. In *Annual Conf. of the Int. Speech Commun. Assoc.*, Itshak Lapidot and Sharon Gannot (Eds.).

[38] Lucien Tisserand and Heike Baldauf-Quilliatre. 2024. Rejecting a robot's offer: An analysis of preference. *Discourse & Commun.* 18, 6 (2024), 931–941.

[39] Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).

[40] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. 2024. DIM: Dyadic Interaction Modeling for Social Behavior Generation. In *Eur. Conf. Comput. Vis.*, Vol. 15095. 484–503.

[41] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. In *IEEE Conf. Comput. Vis. Pattern Recog.* 12780–12790.

[42] Özge Nilay Yalçın. 2020. Empathy framework for embodied conversational agents. *Cognitive Systems Research* 59 (2020), 123–132.

[43] Haijie Yang, Zhenyu Zhang, Hao Tang, Jianjun Qian, and Jian Yang. 2024. ConsistentAvatar: Learning to Diffuse Fully Consistent Talking Head Avatar with Temporal Guidance. In *ACM Int. Conf. Multimedia*. 3964–3973.

[44] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Int. Conf. Comput. Vis.* 11975–11986.

[45] Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2024. SpeechAlign: Aligning Speech Generation to Human Preferences. In *Adv. Neural Inform. Process. Syst.*

[46] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 8652–8661.

[47] Sizhe Zhao, Chenyang Wang, Weiyu Zhao, Zonglin Li, Ming Li, and Shengping Zhang. 2025. REA-Listener: Real-Time Listening Head Generation with Dynamic Emotion Modeling and Flexible Modality Adaptation. In *ACM Int. Conf. Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 9733–9742. doi:10.1145/3746027.3755093

[48] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. 2022. Responsive Listening Head Generation: A Benchmark Dataset and Baseline. In *Eur. Conf. Comput. Vis.*, Vol. 13698. 124–142.

[49] Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. 2025. INFP: Audio-Driven Interactive Head Generation in Dyadic Conversations. In *IEEE Conf. Comput. Vis. Pattern Recog.* 10667–10677.