GDPO-LISTENER: EXPRESSIVE INTERACTIVE HEAD GENERATION VIA AUTO-REGRESSIVE FLOW MATCHING AND GROUP REWARD-DECOUPLED POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

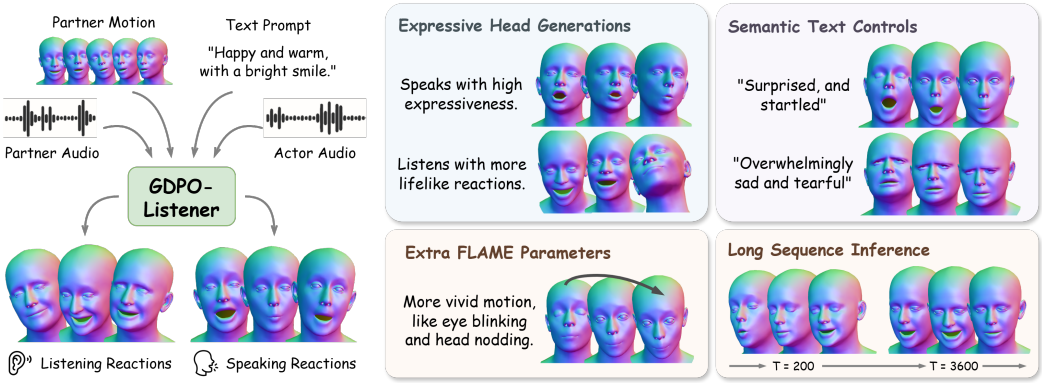

Figure 1: **GDPO-Listener**. Our framework generates expressive speaking and listening head reactions from multimodal dyadic conversational inputs. By utilizing an expanded FLAME parameter space, it supports naturalistic motions, such as eye blinking and head nodding. Furthermore, it enables explicit semantic text control to ensure contextually appropriate responses and maintains stable dynamics during long-sequence inference.

ABSTRACT

Generating realistic 3D head motion for dyadic interactions is a significant challenge in virtual human synthesis. While recent methods achieve impressive results with speaking heads, they frequently suffer from the 'Regression-to-the-Mean' problem in listener motions, collapsing into static faces, and lack the parameter space for complex nonverbal motions. In this paper, we propose GDPO-Listener, a novel framework that achieves highly expressive speaking and listening motion generation. First, we introduce an Auto-Regressive Flow Matching architecture enabling stable supervised learning. Second, to overcome kinematic stillness, we apply the Group reward-Decoupled Policy Optimization (GDPO). By isolating reward normalization across distinct FLAME parameter groups, GDPO explicitly incentivizes high variance expressive generations. Finally, we enable explicit semantic text control for customizable responses. Extensive evaluations across the Seamless Interaction and DualTalk datasets demonstrate superior performance compared to existing baselines on long-term kinematic variance, visual expressivity and semantic controllability.

1 INTRODUCTION

Conditional speaking & listening head generation is essential for building socially intelligent systems capable of establishing interpersonal trust through nonverbal behaviors Gratch et al. (2006). While traditional audio-driven models have achieved accurate speaker lip synchronization, they frequently neglect the listener's nuanced nonverbal feedback. Recent dyadic methods attempt to model

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

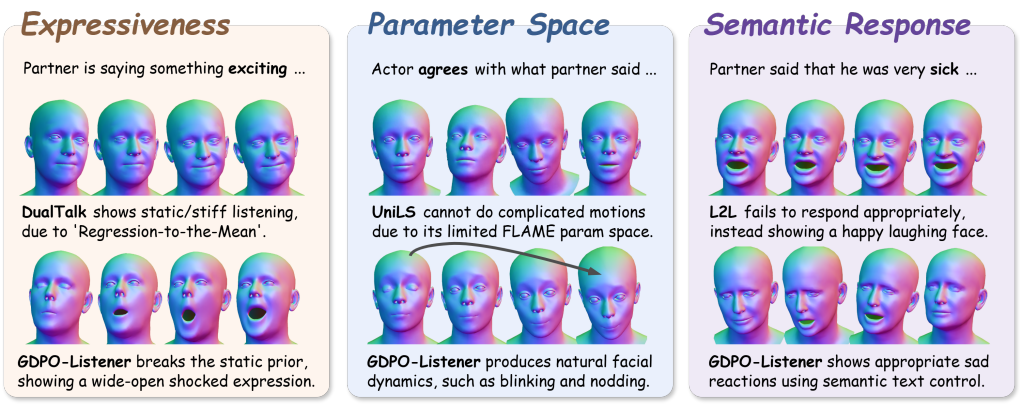

Figure 2: **Necessity of Our Method**. We resolve three critical gaps in previous baselines. (Left) Older methods suffer from static mean-collapse, but we synthesize high-variance expressive reactions. (Middle) Baselines cannot produce complex motions, whereas we enable natural eye blinks and head nods. (Right) Prior models often have no semantic guidance, but our text control ensures contextually appropriate responses.

these coordinated reactions, yet they typically restrict movements to short temporal windows and rely on simplified motion tokenization rather than high-fidelity expressions Tran et al. (2024); Ng et al. (2022); Liu et al. (2024); Fan et al. (2022); Richard et al. (2021); Xing et al. (2023). Despite recent advancements, generating highly expressive and naturalistic motions during dyadic interactions still remains challenging due to the following difficulties:

'Regression-to-the-Mean' Problem in Listener Head Learning. Speaking head generation is a highly deterministic, one-to-one mapping problem strongly controlled by the speaker's audio. For instance, if the input audio is the word 'ouch', the avatar must open its mouth widely, likely accompanied by a painful expression. Because the physical ground truth is strictly bound to the audio signal, supervised learning paradigms easily and accurately solve this problem. In contrast, listening head generation is non-deterministic Zhou et al. (2022). When the actor's own audio is silent, its motion is only weakly correlated with the partner's audio and motion, yielding an inherently one-to-many correspondence Ng et al. (2022); Wang et al. (2025); Ki et al. (2026). For example, if the speaker asks, 'Did your paper get accepted?', the listener might exhibit an excited, exaggerated nod (expressing 'yes'), or a sad, slow head shake (expressing 'no'). Both are entirely valid, natural responses to the exact same conversational context. Current listener generation methods all fundamentally treat this as a standard supervised learning task, whether they rely on pure regression (DIM, DualTalk Tran et al. (2024); Peng et al. (2025)), diffusion (DiffPoseTalk, INFP, CustomListener Sun et al. (2024); Zhu et al. (2025); Liu et al. (2024)), or auto-regressive models (ARTalk Chu et al. (2025a), UniLS Chu et al. (2025b)). They use either MSE loss in continuous spaces or Cross Entropy loss in discrete spaces, but the objective is always to minimize the error against a single ground-truth trajectory. When a network is forced to map a single input to a highly diverse set of valid outputs, it mathematically optimizes for the average of all possible futures. Consequently, the model collapses into predicting safe, low-variance, 'mean' motions. This 'Regression-to-the-Mean' physically manifests as a stiff, static, and unnatural listening avatar. To bypass this regression, recent state-of-the-art methods like UniLS Chu et al. (2025b) propose complex two-stage supervised learning pipelines: first learning an audio-free motion prior, followed by an audio-conditioned fine-tuning stage. However, this introduces critical flaws. First, the audio-free training stage is heavily biased toward the dominant speaking motions in the dataset, failing to learn listening priors. Second, the subsequent audio-conditioned fine-tuning stage inevitably suffers from catastrophic forgetting, degrading the expressiveness learned in the first stage. Instead of relying on multi-stage supervised workarounds, we treat it as an alignment problem. We introduce a post-training paradigm using Group reward-Decoupled Policy Optimization (GDPO) Liu et al. (2026). After standard supervised training, we post-train the model using sequence-level rewards designed to maximize motion diversity and expressive dynamics, avoiding average result and restoring natural expressiveness.

Lack of Expressiveness in Listener Head Generation. Current listening head models struggle to synthesize expressive, contextually appropriate responses. For example, they frequently produce

unreal motion artifacts (*e.g.*, unnatural head/jaw movements or high-frequency jitter). And they also suffer from severe contextual misalignment, where the avatar generates negative expressions (*e.g.*, sadness) in response to positive semantic cues (*e.g.*, hearing 'You passed the exam!'). This lack of natural expressiveness comes from four critical architectural and methodological weaknesses: *(a) Data Imbalance*. Natural dyadic datasets are dominated by neutral, low-energy listening states. By training uniformly across this imbalanced data, previous models statistically overfit to a static, low-amplitude prior, suppressing rare but critical expressive reactions (*e.g.*, enthusiastic nodding or shock). *(b) Restricted FLAME Parameter Spaces*. Existing methods typically predict only a limited subset of FLAME parameters (expression, jaw, and neck pose). Omitting eyelid articulation and global head rotation renders them physically incapable of generating essential behaviors like timely blinks, widened eyes, or large-amplitude head nods. *(c) Limited Explicit Semantic Guidance*. Audio features alone lack the high-level cognitive context required to determine how a listener should react, leading to inappropriate responses. While some recent methods (*e.g.*, UniLS Chu et al. (2025b)) attempt to inject diversity using latent style embeddings, these vectors are highly uninterpretable black boxes that require blind trial-and-error sampling or reference videos during inference. *(d) Absence of Expressiveness Control*. Real-world applications require control over the magnitude of a reaction. Previous generative baselines have a limited ability to modulate expression intensity over time, making it impossible to smoothly transition between a slight smile and an exaggerated laugh. In this work, we systematically address these four limitations to achieve highly expressive, intent-driven listener head generations.

To overcome these barriers, we present **GDPO-Listener**, a novel co-speech listener head generative framework designed to synthesize highly expressive and contextually appropriate listening behaviors. We formulate the base generation as an Auto-Regressive Flow Matching (AR-Flow) process, which seamlessly adapts for long sequences and multimodal prompting. It is further empowered by Classifier-Free Guidance (CFG) Ho & Salimans (2022) to support expression intensity control. Crucially, to overcome the 'Regression-to-the-Mean' problem and the inherent static bias of the training data, we propose a GDPO-based post-training scheme. It directly rewards dynamic motion to force the model out of its static, low-variance prior. By decoupling these rewards across the different numerical scales of the FLAME parameters, we ensure balanced expressiveness without dominant motions taking over. Next, to break the physical bottlenecks of previous baselines, we expand the predictive FLAME parameter space by adding eyelid articulation, eye pose, and global head rotation to our variational autoencoder (VAE) Kingma & Welling (2013). By further conditioning our model on explicit text prompts, we support interpretable, zero-shot semantic control over the listener's cognitive intent. Extensive evaluations on the Seamless Interaction Agrawal et al. (2025) and DualTalk datasets demonstrate that our framework achieves competitive metrics and superior motion quality compared to prior methods. In summary, our main contributions are:

(a) An Auto-Regressive Flow Matching interactive (speaker & listener) head generation framework that accepts multimodal prompts and explicit semantic or intensity controls.

(b) A Group reward-Decoupled Policy Optimization post-training reinforcement learning technique with rewards directly optimized for natural, highly expressive listener heads.

2 Related Work

Our work builds on dyadic conversational head generation Peng et al. (2025); Siniukov et al. (2025); Chu et al. (2025b) and reinforcement learning for continuous generative models Wallace et al. (2024); Liu et al. (2025a). Existing approaches primarily rely on supervised learning, sometimes augmented with text, emotion, or style controls to reduce output ambiguity Siniukov et al. (2025); Teotia et al. (2026); Chen et al. (2026). However, weakly conditioned listener reactions remain inherently one-to-many, and trajectory-matching objectives can favor conservative, low-expressive outputs. Reinforcement learning is widely used for LLM post-training Ouyang et al. (2022); Shao et al. (2024) and has recently been explored in talking head video generation Tan et al. (2026), but it remains largely unexplored for dyadic listening head generation. We combine Auto-Regressive Flow Matching with Group reward-Decoupled Policy Optimization (GDPO) to align generated reactions with expressive and kinematic objectives while accounting for FLAME Li et al. (2017) parameter groups. This shifts the focus from reproducing individual reference trajectories to rewarding expressive, structurally coherent responses. A comprehensive literature review is provided in Appendix B.

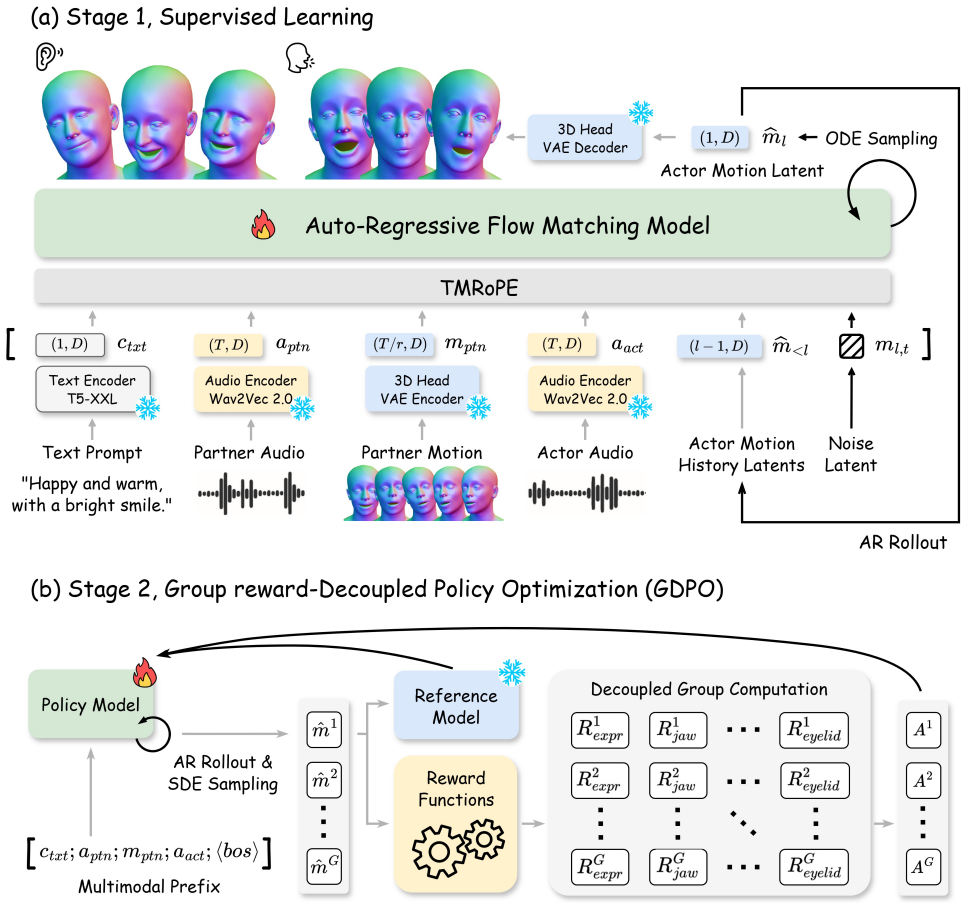

Figure 3: **GDPO-Listener Architecture**. Our framework has two training stages. *(a) Supervised Learning*. Multimodal inputs are encoded as prefix conditions, and an Auto-Regressive Flow Matching model iteratively predicts actor motion latents from noise and history via ODE sampling. *(b) Reinforcement Learning*. We then post-train the policy model via GDPO. We compute decoupled rewards for distinct FLAME parameters under SDE sampling to explicitly optimize expressiveness.

3 METHODS

An overview of our **GDPO-Listener** framework is presented in Fig. 3. We design a two-stage training mechanism to synthesize highly expressive listener head reactions. The following sections provide further details: Sec. 3.1 introduces our continuous facial kinematics representations. Sec. 3.2 explains our Auto-Regressive Flow Matching base model. Finally, Sec. 3.3 details Group reward-Decoupled Policy Optimization (GDPO), our reinforcement learning stage applied exclusively to listener frames to optimize expressive rewards.

3.1 PRELIMINARIES

We parameterize facial kinematics using the FLAME 3D morphable model (3DMM) Li et al. (2017). Previous dyadic methods restrict this space to low-dimensional representations, with separate expression and pose (jaw and neck) parameters. To capture the full spectrum of nonverbal cues, we expand the motion space by adding eyelid closure, eye pose, and global head rotation. This enables expressive behaviors and realistic interactions, such as eye blinking and intense head nodding.

Let $M \in \mathbb{R}^{L \times d}$ denote a continuous motion sequence. Recent listener generation models Tran et al. (2024); Ng et al. (2022), compress M using discrete quantization (*e.g.*, VQ-VAE). However, forcing continuous kinematics into discrete codebooks introduces quantization bounds, physically manifesting as high-frequency temporal jitter. To preserve smooth human motion, we employ a

transformer-based VAE to learn a continuous latent space. For stable optimization, we first normalize the FLAME parameters channel-wise to obtain M'. The VAE encoder $\mathcal{E}_{\text{enc}}$ then maps M' into a temporally compressed distribution $m \in \mathbb{R}^{L/r \times D_v}$, where r is the temporal downsampling rate.

3.2 Stage 1: Supervised Learning

In this stage, we employ a supervised auto-regressive flow matching formulation to establish a high-fidelity base policy for both speaker and listener head generation, ensuring strict adherence to multi-modal constraints. We first extract embeddings for the text prompt, the partner's audio and motion, and the actor's audio. These are projected into dimension D to form the conditioning prefix:

$$\begin{aligned} c_{\text{txt}} &= (\mathcal{P}_{\text{txt}} \circ \mathcal{T}_{\text{enc}})(\texttt{text_prompt}) \\ [a_{\text{act}}; a_{\text{ptn}}] &= (\mathcal{P}_{\text{audio}} \circ \mathcal{A}_{\text{enc}})([\texttt{actor_audio}; \texttt{partner_audio}]) \\ m_{\text{ptn}} &= (\mathcal{P}_{\text{motion}} \circ \mathcal{E}_{\text{enc}})(\texttt{partner_motion}) \end{aligned} \tag{1}$$

where $c_{\text{txt}} \in \mathbb{R}^{1\times D}$, $a_{\text{act}}, a_{\text{ptn}} \in \mathbb{R}^{L\times D}$, and $m_{\text{ptn}} \in \mathbb{R}^{L/r\times D}$. Here, L is the sequence length, r is the temporal downsampling rate, $\mathcal{P}_*$ are trainable linear projectors, and the upstream encoders ($\mathcal{T}_{\text{enc}}, \mathcal{A}_{\text{enc}}, \mathcal{E}_{\text{enc}}$) remain frozen.

Time-aligned Multimodal RoPE (TMRoPE). Precise audio-visual synchronization is critical for realistic speaking and listening head generation. And it is even more challenging due to various temporal resolutions of multimodal inputs (*e.g.*, high-frequency audio versus downsampled motion latents). To resolve this temporal ambiguity, we use TMRoPE, inspired by multimodal alignment in Qwen2.5-Omni Xu et al. (2025). We map all input streams into a shared temporal coordinate system by assigning an explicit timestamp index t to each token:

$$\begin{cases} t^{\text{text}} = 0 & \\ t_i^{\text{audio}} = i, & i \in \{0, 1, \ldots, L-1\} \\ t_j^{\text{motion}} = j \cdot r + \lfloor r/2 \rfloor, & j \in \{0, 1, \ldots, \lfloor L/r \rfloor - 1\} \end{cases} \tag{2}$$

where L is the audio sequence length and r is the motion downsampling rate. The offset $\lfloor r/2 \rfloor$ mathematically anchors each low-frequency motion token to the exact center of its corresponding high-frequency audio window. Within the transformer blocks, we apply Rotary Position Embeddings (RoPE) Su et al. (2024) with aforementioned modality-specific indices to self-attention input, equipping the model with strict, physically grounded cross-modal alignment.

Auto-Regressive Flow Matching (AR-Flow). We model the generation of the l-th motion latent token m_l via continuous flow matching. Let $t \in [0, 1]$ denote the flow timestep, where $m_{l,0} \sim \mathcal{N}(0, I)$ is Gaussian noise and $m_{l,1} = m_l$ is the l-th ground truth latent. The intermediate state is defined via linear interpolation $m_{l,t} = t \cdot m_l + (1-t)m_{l,0}$. The AR-Flow network $\mathcal{F}$ predicts the optimal velocity vector v_θ pointing toward the target:

$$v_\theta(m_{l,t}) = \mathcal{F}([c_{\text{txt}}; a_{\text{ptn}}; m_{\text{ptn}}; a_{\text{act}}; \hat{m}_{<l}; m_{l,t}], t) \tag{3}$$

The model is optimized using the Rectified Flow objective, minimizing:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, m_l, m_{l,0}} \left[\| v_\theta(m_{l,t}) - (m_l - m_{l,0}) \|^2 \right] \tag{4}$$

During inference time, we sample the latent token by solving the Ordinary Differential Equation (ODE) using Euler sampler from $t = 0$ to $t = 1$.

$$\hat{m}_{l,t+\Delta t} = \hat{m}_{l,t} + \Delta t \cdot v_\theta(\hat{m}_{l,t}) \tag{5}$$

Once the l-th motion latent $\hat{m}_l$ is fully estimated, it will be appended to the actor's motion history, such that $\hat{m}_{<(l+1)} = [\hat{m}_{<l}; \hat{m}_l]$. This updated history then conditions the AR-Flow network $\mathcal{F}$ for the subsequent latent prediction, enabling auto-regressive generation. For inference-time expressiveness control, we use Classifier-Free Guidance (CFG). We extrapolate the inference velocity with an unconditional prior $v_\emptyset(\hat{m}_{l,t})$ and the conditional prior $v_{\text{cond}}(\hat{m}_{l,t})$:

$$\hat{v}_\theta(\hat{m}_{l,t}) = v_\emptyset(\hat{m}_{l,t}) + \omega_l \cdot (v_{\text{cond}}(\hat{m}_{l,t}) - v_\emptyset(\hat{m}_{l,t})) \tag{6}$$

where $\omega_l \in \{\omega_{\text{speak}}, \omega_{\text{listen}}\}$ based on whether the actor is speaking or listening at frame l. This decoupling enables independent tuning of active lip-sync fidelity and reactive listener variance without retraining.

3.3 STAGE 2: REINFORCEMENT LEARNING

Supervised baselines suffer from the aforementioned 'Regression-to-the-Mean' problem, producing under expressive listener motions. To break this deterministic collapse and drive more expressive listener-head generation, we use Group reward-Decoupled Policy Optimization Liu et al. (2026) as a post-training stage for our autoregressive flow-matching framework.

From ODE to SDE. Following Flow-GRPO Liu et al. (2025a), we use stochastic Gaussian transitions for policy optimization. We augment the AR-Flow dynamics with exploration noise and apply Euler–Maruyama discretization:

$$\hat{m}_{l,t+\Delta t} = \hat{m}_{l,t} + \Delta t\, v_\theta(\hat{m}_{l,t}) + \sigma\sqrt{\Delta t}\,\epsilon, \tag{7}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and σ controls the noise scale. Our rollout transition omits the score-dependent drift correction used in Flow-GRPO, providing a simplified exploration sampler without enforcing marginal equivalence with the inference ODE. We use fewer integration steps during RL rollouts for training efficiency and retain the ODE sampler with full integration steps at inference.

Group reward-Decoupled Policy Optimization (GDPO). Applying group-based online RL methods to listener head generation presents unique challenges due to the multi-dimensional nature of the motion space. We optimize over N-dimensional distinct FLAME parameters, indexed by $i \in \{\texttt{expr}, \texttt{jaw}, \texttt{neck}, \texttt{eyelid}, \texttt{eyepose}, \texttt{rot}\}$. To maintain the distinctions among diverse reward combinations and accurately capture their relative scales, we adapt the GDPO. We first execute the AR rollout and SDE sampling to generate a full sequence. This yields a trajectory set $\{\hat{m}_l^j\}_{l=1}^{L/r}$, and L/r denotes the latent length. For each candidate trajectory, we compute sequence-level rewards R_i^j by comparing the generated sequence against the ground truth $\{m_l^j\}_{l=1}^{L/r}$. To prevent large-magnitude rewards from overshadowing penalties, we compute normalized advantages independently for each sample $j \in \{1, 2, \ldots, G\}$ within the group:

$$A_i^j = \frac{R_i^j - \text{mean}(\{R_i^k\}_{k=1}^G)}{\text{std}(\{R_i^k\}_{k=1}^G)}, \quad R_i^j = R_i^j(\{\hat{m}_l^j\}_{l=1}^{L/r}, \{m_l^j\}_{l=1}^{L/r}) \tag{8}$$

The final scalar advantage for the j-th rollout is then obtained via a normalized weighted sum across the N FLAME parameter spaces:

$$A^j = \frac{\sum_{i=1}^N \lambda_i A_i^j}{\sum_{i=1}^N \lambda_i + \epsilon_0} \tag{9}$$

where λ_i provides explicit control over the relative importance of specific FLAME parameters, and ϵ_0 ensures numerical stability during computation. For B conditioning windows with G rollouts each sampled from the behavior policy, we minimize

$$\mathcal{L}_{\text{GDPO}}(\theta) = \frac{1}{BG}\sum_{j=1}^{BG}\left[-\min\left(\tilde{\rho}^j A^j, \text{clip}(\tilde{\rho}^j, 1-\epsilon, 1+\epsilon)A^j\right) + \beta\mathcal{K}^j(\theta)\right]. \tag{10}$$

Here, $\tilde{\rho}^j = \exp(s_\theta^j - s_{\theta_{\text{old}}}^j)$ is a tempered trajectory ratio computed from normalized log-scores, A^j is the detached advantage, and ϵ controls clipping. $\mathcal{K}^j$ is a normalized conditional Gaussian KL penalty against the reference policy, evaluated at cached behavior-policy states. We cache rollouts without gradients and optimize through chunked replay. More details can be found in Appendix F.

Reward Function. To encourage expressive speaker and listener dynamics while preserving motion fidelity, we design a multi-component reward function. We decode $\hat{m}$ with the VAE decoder to obtain $\hat{M}$ in FLAME space, then compute per-group rewards on $\hat{M}$. For a given FLAME parameter feature i, the reward R_i synthesizes four distinct objectives: a positional variance penalty $\sum_{l=1}^L |\sigma(\hat{M}_l[i]) - \sigma(M_l[i])|$ to match structural motion amplitude, a velocity variance penalty $\sum_{l=1}^L |\sigma(\nabla\hat{M}_l[i]) - \sigma(\nabla M_l[i])|$ to capture high-frequency human dynamics, a mean-drift penalty $|\mu(\hat{M}[i]) - \mu(M[i])|$ to prevent global drift away from the segment's typical pose, and a standard MSE loss $\sum_{l=1}^L ||\hat{M}_l[i] - M_l[i]||^2$ to ensure semantic grounding. These individual penalties are aggregated via a weighted sum to compute the final reward for each FLAME feature term. Importantly, the reward primarily emphasizes dynamics-aware statistics (variance and velocity variance) rather than per-frame alignment. Combined with group-decoupled normalization, it encourages diverse, high-variance rollouts while using the mean and MSE anchors only to stabilize realism and avoid degenerate drift.

Methods	Speaking					Listening		
	LVE ↓	MHD ↓	FDD ↓	PDD ↓	JDD ↓	FDD ↓	PDD ↓	JDD ↓
DiffposeTalk (Sun et al. (2024))	9.48	2.96	32.66	7.89	1.40	-	-	-
L2L (Ng et al. (2022))	3.10	1.25	53.03	3.52	3.32	25.95	2.18	1.14
ARTalk (Chu et al. (2025a))	7.46	2.12	31.64	7.66	1.19	-	-	-
DualTalk (Peng et al. (2025))	6.53	1.95	37.46	9.70	**1.02**	43.58	10.71	2.02
UniLS (Chu et al. (2025b))	4.09	1.44	34.97	5.39	4.53	22.23	2.85	1.75
GDPO-Listener	**2.95**	**1.06**	**22.45**	**2.71**	1.26	**18.85**	**1.90**	**0.87**

Table 1: Quantitative Analysis on Seamless Interaction Dataset.

Methods	Speaking					Listening		
	LVE ↓	MHD ↓	FDD ↓	PDD ↓	JDD ↓	FDD ↓	PDD ↓	JDD ↓
L2L (Ng et al. (2022))	3.40	1.52	55.10	5.15	3.38	44.77	3.77	2.39
DualTalk (Peng et al. (2025))	**2.86**	1.34	64.84	5.04	3.70	55.94	5.08	2.79
UniLS (Chu et al. (2025b))	3.98	1.57	37.23	5.99	4.26	30.56	5.58	2.77
GDPO-Listener	3.08	**1.25**	**30.87**	**4.44**	**1.67**	**22.56**	**3.33**	**1.48**

Table 2: Quantitative Analysis on Seamless Interaction Expressive Subset.

4 EXPERIMENTS

We evaluate GDPO-Listener through quantitative comparisons (Section 4.1), qualitative results (Section 4.2), and ablation studies (Section 4.3). Experimental settings, dataset collection, and evaluation metrics are detailed in Appendix F.1, E.1, and C.4, respectively. Additional experimental analyses and visualizations are provided in Appendix C and D.

4.1 QUANTITATIVE RESULTS

Seamless Interaction Dataset. GDPO-Listener achieves the lowest geometric errors and the best performance across all listening metrics on the full test set in Tab. 1. To rigorously evaluate the 'Regression-to-the-Mean' problem, we further test on a curated expressive subset featuring high-variance ground-truth motions in Tab. 2. UniLS is executed with zero style embedding and zero history conditioning, as both take ground truth target motion as input. These results show that our model sustains expressive, dynamically faithful speaker and listener motions. Additional evaluation metrics are reported in Appendix C.5.

DualTalk Dataset. We verify generalization on the DualTalk dataset in Tab. 3. GDPO-Listener has competitive performances across those dynamic metrics. This demonstrates our expressivity improvements successfully generalize to novel data domains.

Zero-shot on Vico Dataset. Furthermore, we conducted a zero-shot cross-dataset evaluation on the VICO Dataset in Tab. 4. Because the VICO test set consists entirely of unseen identities, our model's performance conclusively demonstrates robust generalization ability.

User Studies. We evaluate perceptual quality with 42 Prolific participants. A chi-square power analysis supports the adequacy of this sample size, yielding 98.7% power at $\alpha = 0.05$ for the smallest observed effect (Naturalness; Cohen's w = 0.7). Participants prefer GDPO-Listener across all four criteria in Tab. 5. Further details of the user study can be found in Appendix C.7.

Efficiency Analysis. As shown in Tab. 6, GDPO-Listener requires the fewest parameters and GFLOPs among the compared methods while supporting real-time generation through sliding-window inference. † indicates that GFLOPs and parameters exclude frozen multimodal encoders.

4.2 QUALITATIVE RESULTS

Fig. 4 compares GDPO-Listener with baseline methods. During speaking, our method produces clearer articulation and more expressive facial motion. During listening, it generates varied, contextually appropriate reactions, such as frowning or laughter, while baselines exhibit limited motion and expression changes. Additional visualizations are provided in Appendix D.

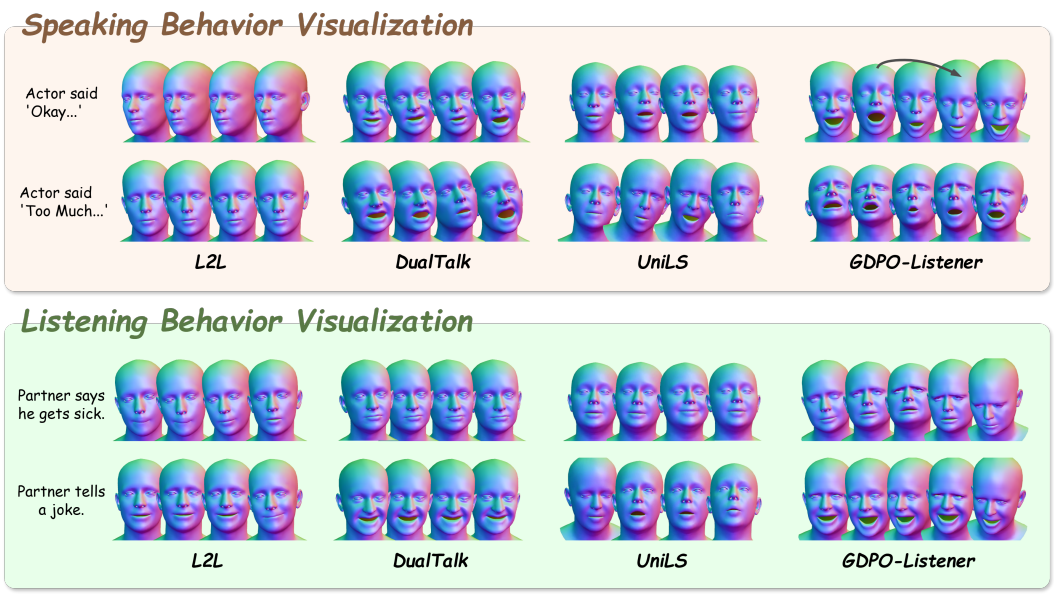

Figure 4: **Qualitative Comparisons**. Other methods have low-expressive speaking and static listening, our method shows better lip sync and highly expressive reactions.

Semantic Text Control. Audio alone may provide ambiguous cues for listener reactions. Through multimodal prefix conditioning, GDPO-Listener allows users to guide facial behavior with text prompts. As shown in Fig. 5, prompts such as 'happy' or 'sad' produce distinct expressions, enabling explicit control over the intended reaction. Additional examples are in Appendix D.1.

Long Sequence Inference. GDPO-Listener supports extended generation through sliding window autoregressive inference, conditioning each window on the latest latent. As shown in Fig. 5, our method sustains dynamic and expressive reactions over long sequences.

Expressiveness Control. We control expressiveness through Classifier-Free Guidance (CFG) scaling. As shown in Fig. 5, the default scale of 1.0 matches ground-truth dynamics for evaluation, while higher scales produce more vivid speaking and listening reactions.

4.3 Ablation Studies

We ablate our framework's key components in Tab. 7. Crucially, removing GDPO fine-tuning (*without GDPO*) severely degrades all dynamic metrics. This demonstrates the benefit of GDPO post-training for improving motion dynamics in the evaluated supervised AR-Flow model. Omitting

Methods	All Listening			Expressive Listening		
	FDD ↓	PDD ↓	JDD ↓	FDD ↓	PDD ↓	JDD ↓
DualTalk	34.36	3.55	1.95	40.57	3.68	2.43
UniLS	33.98	5.24	2.88	45.47	5.46	3.51
GDPO-Listener	**25.26**	**2.76**	**1.33**	**37.65**	**2.88**	**1.91**

Table 3: Quantitative Analysis on DualTalk Dataset.

Methods	Listening		
	FDD ↓	PDD ↓	JDD ↓
DualTalk	28.46	1.58	0.79
UniLS	20.14	2.07	1.69
GDPO-Listener	**13.98**	**1.27**	**0.49**

Table 4: Zero-shot on Vico Dataset.

Methods	Lip Sync	Naturalness	Liveliness	Diversity
L2L	0.8%	3.7%	2.9%	2.4%
DualTalk	30.4%	26.2%	20.4%	21.2%
UniLS	13.5%	15.6%	16.7%	13.8%
GDPO-Listener	**55.3%**	**54.5%**	**60.1%**	**62.7%**

Table 5: User Study with 42 Prolific Participants.

Methods	Real-Time	FPS ↑	GFLOPs† ↓	Params† ↓
ARTalk	✓	2,227	252.6	144.5 M
DualTalk	✗	**8,901**	323.9	638.8 M
UniLS	✓	1,194	537.3	52.6 M
GDPO-Listener	✓	2,357	**31.4**	**22.2 M**

Table 6: Computational Efficiency Comparison.

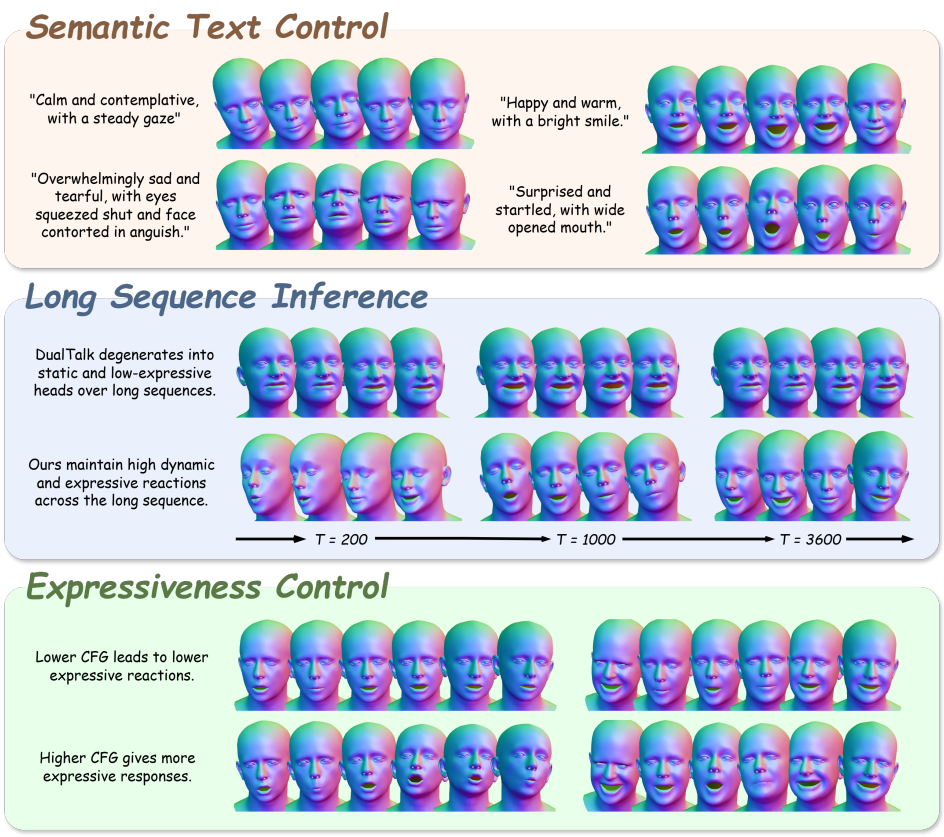

Figure 5: **Advanced Generation Capabilities**. (Top) Semantic text explicitly controls emotional states. (Middle) We sustain dynamic reactions during long sequences, avoiding baseline static decay. (Bottom) CFG scaling seamlessly modulates expressiveness intensity without retraining.

Methods	Speaking					Listening		
	LVE ↓	MHD ↓	FDD ↓	PDD ↓	JDD ↓	FDD ↓	PDD ↓	JDD ↓
GDPO-Listener	2.95	1.06	**22.45**	**2.71**	**1.26**	**18.85**	**1.90**	**0.87**
without text prompts	2.73	1.04	33.02	2.97	1.44	22.19	2.11	0.87
without RL	3.07	1.12	39.54	3.50	1.81	28.03	2.15	1.09
with fewer denoise steps	**2.61**	**0.98**	42.12	3.94	2.11	29.39	2.19	1.17

Table 7: Ablation Studies on Seamless Interaction Dataset.

the text prefix (*without text prompts*) similarly hurts expressiveness, as the model loses semantic guidance to resolve acoustic ambiguities. Reducing flow sampling steps from 10 to 4 (*with fewer denoise steps*) severely penalizes expressivity. More ablation studies can be found in Appendix C.

5 Conclusions

We presented GDPO-Listener, combining Auto-Regressive Flow Matching with Group reward-Decoupled Optimization post-training for expressive conversational head generation. Natural listening behavior is often muted, making expressive reactions difficult to learn from predominantly low-motion data. Our experiments show that explicitly rewarding motion dynamics during post-training yields stronger improvements than traditional supervised learning or expressive subset fine-tuning. Balancing dynamics and reconstruction rewards improves the trade-off between expressiveness and geometric fidelity, while group-wise normalization improves upon standard GRPO. Improvements extend to metrics outside the reward, and the user study supports the quality of the complete system. These findings demonstrate that reward-based post-training enhances speaker and listener expressiveness while preserving motion fidelity.

REFERENCES

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, Sahir Gomez, Hongyu Gong, Srivathsan Govindarajan, Brandon Han, Sen He, Denise Hernandez, Yordan Hristov, Rongjie Huang, Hirofumi Inaguma, Somya Jain, Raj Janardhan, Qingyao Jia, Christopher Klaiber, Dejan Kovachev, Moneish Kumar, Hang Li, Yilei Li, Pavel Litvin, Wei Liu, Guangyao Ma, Jing Ma, Martin Ma, Xutai Ma, Lucas Mantovani, Sagar Miglani, Sreyas Mohan, Louis-Philippe Morency, Evonne Ng, Kam-Woh Ng, Tu Anh Nguyen, Amia Oberai, Benjamin Peloquin, Juan Pino, Jovan Popovic, Omid Poursaeed, Fabian Prada, Alice Rakotoarison, Rakesh Ranjan, Alexander Richard, Christophe Ropers, Safiyyah Saleem, Vasu Sharma, Alex Shcherbyna, Jia Shen, Jie Shen, Anastasis Stathopoulos, Anna Sun, Paden Tomasello, Tuan Tran, Arina Turkatenko, Bo Wan, Chao Wang, Jeff Wang, Mary Williamson, Carleigh Wood, Tao Xiang, Yilin Yang, Julien Yao, Chen Zhang, Jiemin Zhang, Xinyue Zhang, Jason Zheng, Pavlo Zhyzheria, Jan Zikes, and Michael Zollhoefer. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset, 2025. URL https://arxiv.org/abs/2506.22554.

Bohong Chen and Haiyang Liu. Dystream: Streaming dyadic talking heads generation via flow matching-based autoregressive model. *arXiv preprint arXiv:2512.24408*, 2025.

Hejia Chen, Haoxian Zhang, Xu He, Xiaoqiang Liu, Pengfei Wan, Shoulong Zhang, and Shuai Li. Mindflow: Harmonizing cognitive semantics and acoustic dynamics for facial animation generation in dyadic conversations. *arXiv preprint arXiv:2606.27779*, 2026.

Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems*, 37:57642–57670, 2024.

Xuangeng Chu, Nabarun Goswami, Ziteng Cui, Hanqin Wang, and Tatsuya Harada. Artalk: Speech-driven 3d head animation via autoregressive model. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–9, 2025a.

Xuangeng Chu, Ruicong Liu, Yifei Huang, Yun Liu, Yichen Peng, and Bo Zheng. Unils: End-to-end audio-driven avatars for unified listening and speaking. *arXiv preprint arXiv:2512.09327*, 2025b.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18770–18780, 2022.

Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, R J van der Werf, and Louis-Philippe Morency. Virtual Rapport. In Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier (eds.), *Intelligent Virtual Agents*, pp. 14–27, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-37594-4.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Taekyung Ki, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, and Sung Ju Hwang. Avatar forcing: Real-time interactive head avatar generation for natural conversation. *arXiv preprint arXiv:2601.00664*, 2026.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. December 2013. URL http://arxiv.org/abs/1312.6114. arXiv: 1312.6114.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194:1–194:17, November 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130813. URL https://dl.acm.org/doi/10.1145/3130800.3130813.

Habin Lim, Jae-Ho Lee, Hah Min Lew, Ji-Su Kang, and Gyeong-Moon Park. Faceplex: Full-duplex joint speech-facial motion generation for conversational avatars. *arXiv preprint arXiv:2606.30145*, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8009–8019, 2025b.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

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

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*, pp. 20395–20405, 2022.

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 (ICCV)*, pp. 10083–10093, October 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1173–1182, 2021.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

Liam Schoneveld, Zhe Chen, Davide Davoli, Jiapeng Tang, Saimon Terazawa, Ko Nishino, and Matthias Nießner. Sheap: Self-supervised head geometry predictor learned via 2d gaussians. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14162–14172, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Maksim Siniukov, Di Chang, Minh Tran, Hongkun Gong, Ashutosh Chaubey, and Mohammad Soleymani. Ditailistener: Controllable high fidelity listener video generation with diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11991–12001, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-Jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4), 2024. doi: 10.1145/3658221.

Weiting Tan, Andy T Liu, Ming Tu, Xinghua Qu, Philipp Koehn, and Lu Lu. Flowportrait: Reinforcement learning for audio-driven portrait video generation. *arXiv preprint arXiv:2603.00159*, 2026.

Kartik Teotia, Helge Rhodin, Hyeongwoo Kim, Marc Habermann, and Christian Theobalt. Steer: Steerable dyadic head avatars. *arXiv preprint arXiv:2607.23840*, 2026.

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

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Yinuo Wang, Yanbo Fan, Xuan Wang, Guo Yu, and Fei Wang. Diffusion-based realistic listening head generation via hybrid motion modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15885–15895, 2025.

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025.

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. Talking head generation with probabilistic audio-to-visual diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7645–7655, 2023.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. *arXiv preprint arXiv:2406.07548*, 2024.

Deyu Zhou, Quan Sun, Yuang Peng, Kun Yan, Runpei Dong, Duomin Wang, Zheng Ge, Nan Duan, and Xiangyu Zhang. Taming teacher forcing for masked autoregressive video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7374–7384, 2025.

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*, pp. 124–142. Springer, 2022.

Yihan Zhou, Zikai Huang, Yuyang Yu, Xuemiao Xu, Cheng Xu, and Shengfeng He. Intersing: Explicit interaction dynamics for 3d duet singing animation and beyond. *arXiv preprint arXiv:2609.04903*, 2026.

Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. Infp: Audio-driven interactive head generation in dyadic conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10667–10677, 2025.

APPENDIX

TABLE OF CONTENTS

- STATEMENTS ..... A
  - AI USE STATEMENT A.1
  - ETHICS STATEMENT A.2
  - REPRODUCIBILITY STATEMENT A.3
- RELATED WORK ..... B
- EXPERIMENTAL RESULTS C
  - EFFECTIVENESS OF REINFORCEMENT LEARNING C.1
  - ABLATION STUDIES ON REWARD COMPONENTS C.2
  - ABLATION STUDIES ON FLAME GROUP ITEMS C.3
  - EVALUATION METRICS C.4
  - ADDITIONAL EVALUATION METRICS C.5
  - COMPARISON BETWEEN GDPO AND GRPO C.6
  - USER STUDIES C.7
- ADDITIONAL VISUALIZATIONS ..... D
  - EXTRA TEXT CONTROL VISUALIZATIONS D.1
  - PHOTOREALISTIC GAUSSIAN AVATAR INTEGRATION D.2
  - VIDEO DEMO D.3
- DATASET DETAILS ..... E
  - DATASET COLLECTION E.1
  - TEXT ANNOTATION DETAILS E.2
  - DATASET STATISTICS E.3
- IMPLEMENTATION DETAILS ..... F
  - EXPERIMENT SETTINGS F.1
  - ARFLOW & GDPO TRAINING DETAILS F.2
  - LONG SEQUENCE INFERENCE F.3
  - EXPRESSIVENESS CONTROL F.4
- LIMITATIONS & FUTURE WORK ..... G

A STATEMENTS

A.1 AI USE STATEMENT

We used generative AI tools to assist with writing and improve the grammar, clarity, and consistency of the manuscript. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content. Our use of vision-language models for data annotation is described in the Appendix E.2.

A.2 ETHICS STATEMENT

This work studies expressive dyadic head generation using the Seamless Interaction dataset. The proposed method is intended to support virtual avatars and human–computer interaction. However, generated facial behaviors may be misused to fabricate reactions or misrepresent an individual's emotions or intentions. Such outputs should be clearly identified as synthetic and should not be interpreted as evidence of a real person's psychological state. Furthermore, biases in the training data may affect the representation of expressions across demographic and cultural groups. We encourage responsible use that respects privacy and consent, and discourage deceptive impersonation or other harmful applications.

A.3 Reproducibility Statement

To facilitate reproducibility, we describe the model architecture and two-stage training procedure in the main paper section 3, with further implementation details in Appendix F. Dataset preparation, text annotation prompts, and dataset statistics are provided in Appendix E. Appendix C details the evaluation metrics and complementary ablation studies examining reinforcement learning, reward components, and FLAME parameter groups. Appendix D presents the user study interface and additional qualitative visualizations. Together, these materials document the training and evaluation procedures used in our experiments.

B Related Work

Dyadic Conversational Head Generation. Talking head models predict facial motion from strongly correlated audio, but listening head generation is fundamentally different. Listener reactions are weakly correlated with the speaker's audio, making the mapping highly stochastic. Learning expressive listener reactions under such weak conditioning remains challenging, and prior work has reported insufficient motion variation in supervised models. Early dyadic approaches used pure regression (DIM Tran et al. (2024), DualTalk Peng et al. (2025)) or diffusion (DiTaiListener Siniukov et al. (2025), INFP Zhu et al. (2025), FacePlex Lim et al. (2026), InterSing Zhou et al. (2026)). Relying on fixed-window forward passes, they require extra temporal smoothing and struggle with real-time, long-sequence generation. Pure Auto-Regressive (AR) models, like UniLS Chu et al. (2025b) or LM-Listener Ng et al. (2023), solve this by causally predicting frames, natively supporting infinite sequences. However, concurrent AR methods fail in their representation space. The field uses FLAME to parameterize continuous 3D facial kinematics. Recent AR models tokenize these movements into discrete codebooks via VQ-VAE Van Den Oord et al. (2017) or BSQ-VAE Zhao et al. (2024). Forcing continuous data into discrete spaces introduces severe quantization errors. In facial animation, this manifests as unnatural jitter, abruptly snapping between states. Just like DyStream Chen & Liu (2025), to maintain continuous human kinematics, we then employ Auto-Regressive Flow Matching, combining causal AR generation with the mathematically high-fidelity synthesis of continuous latent flows.

Resolving the 'Regression-to-the-Mean' Problem. Listener reactions admit multiple plausible responses to the same conversational context. Prior work has explored training and conditioning strategies to improve listener expressiveness under this ambiguity. Some works frame the issue as an exposure bias between teacher-forcing training and auto-regressive inference Zhou et al. (2025). They apply techniques like Scheduled Sampling, or in continuous generative spaces, Diffusion-Forcing and Self-Forcing. Others, such as UniLS Chu et al. (2025b), employ two-stage training: unconditional free-motion followed by conditional dyadic finetuning. Another direction is to strengthen the weak conditioning to force a specific output. DiTaiListener injects semantic text control; STEER Teotia et al. (2026) requires explicit gaze, head rhythm, and emotion controls; MindFlow Chen et al. (2026) uses 8 emotion categories; and UniLS and ARTalk utilize style embeddings derived from Ground Truth. However, these approaches still rely primarily on supervised learning from ground-truth trajectories, which may not sufficiently encourage expressive listener dynamics in a one-to-many generative task. Instead, we treat listening head generation as an alignment problem. By applying Group reward-Decoupled Policy Optimization (GDPO), we abandon the restrictive trajectory-matching approach. We optimize a specialized reward function, achieving highly expressive and high-fidelity listener reactions without structural degradation.

Reinforcement Learning for Listening Head Generation. While prior head generation methods rely on supervised trajectory matching Ng et al. (2022); Sun et al. (2024); Yu et al. (2023), RLHF Ouyang et al. (2022) and preference-based optimization (*e.g.*, PPO, DPO, GRPO) provide a compelling alternative to directly optimize perceptual or task-driven objectives Schulman et al. (2017); Rafailov et al. (2023); Shao et al. (2024). Importantly, these post-training paradigms have started to move beyond discrete-token LLMs to continuous generative models, including diffusion for images and videos Wu et al. (2025); Wallace et al. (2024); Liu et al. (2025b), and have very recently been explored in talking-head animation as well Tan et al. (2026). However, RL-style post-training remains largely unexplored in dyadic listening head generation, and directly applying preference optimization to 3D facial kinematics is non-trivial. Heterogeneous parameter subspaces with vastly different scales and dimensionalities can destabilize naive reward aggregation. In dyadic

Methods	Speaking					Listening		
	LVE ↓	MHD ↓	FDD ↓	PDD ↓	JDD ↓	FDD ↓	PDD ↓	JDD ↓
AR-Flow	3.07	1.12	39.54	3.50	1.81	28.03	2.15	1.09
AR-Flow with more denoising steps	3.06	1.14	30.90	3.13	1.37	21.07	2.37	0.91
AR-Flow + SFT on expressive subset	3.05	1.14	39.21	3.41	1.79	26.14	2.11	0.98
AR-Flow + GDPO (ours)	**2.95**	**1.06**	**22.45**	**2.71**	**1.26**	**18.85**	**1.90**	**0.87**

Table 8: Effectiveness of Reinforcement Learning on Seamless Interaction Dataset. Starting from the Stage 1 supervised AR-Flow model, we compare additional denoising steps, supervised fine-tuning (SFT) on the expressive subset, and Stage 2 reinforcement learning with GDPO.

Reward Components				Speaking					Listening		
$\sigma(M)$	$\sigma(\nabla M)$	μ	MSE	LVE ↓	MHD ↓	FDD ↓	PDD ↓	JDD ↓	FDD ↓	PDD ↓	JDD ↓
–	–	–	–	3.07	1.12	39.54	3.50	1.81	28.03	2.15	1.09
–	✓	–	–	3.23	1.17	23.02	2.90	**1.25**	19.64	**1.89**	0.81
–	–	–	✓	**2.56**	**0.96**	57.16	5.12	3.18	37.05	2.64	1.53
–	✓	–	✓	3.01	1.09	23.19	2.88	**1.25**	19.32	**1.89**	**0.80**
–	✓	✓	✓	3.00	1.06	23.51	2.85	1.26	19.04	1.91	0.81
✓	✓	–	✓	3.01	1.08	23.19	2.87	**1.25**	19.26	**1.89**	**0.80**
✓	✓	✓	✓	2.95	1.06	**22.45**	**2.71**	1.26	**18.85**	1.90	0.87

Table 9: Ablation Studies on Reward Components on Seamless Interaction Dataset.

listener head generation, where multiple reactions can be equally valid for the same context, MSE-style trajectory matching can penalize expressive yet appropriate responses, and even continuous adaptations such as Diffusion-DPO Wallace et al. (2024) may still favor conservative outputs when preference signals are trajectory-agnostic in such multimodal settings. Recent flow-model post-training (*e.g.*, Flow-GRPO Liu et al. (2025a)) show that exploration-aware rollouts are critical for stable group-relative policy updates, as deterministic sampling can yield insufficient intra-group diversity to estimate advantages reliably. Yet, none of these works address the unique challenges posed by 3D facial kinematics in dyadic settings, a gap our method is designed to address.

C Experimental Results

C.1 Effectiveness of Reinforcement Learning

We examine whether alternatives to reinforcement learning can achieve comparable improvements in listener expressiveness. Starting from the same Stage 1 AR-Flow model, we evaluate additional denoising steps and supervised fine-tuning (SFT) on an expressive subset. The former increases inference computation, while the latter emphasizes expressive examples but retains the supervised learning objective. Neither explicitly optimizes generated motion dynamics through sequence-level rewards. As shown in Table 8, these alternatives yield partial improvements but consistently underperform GDPO across all reported metrics. This indicates that increasing sampling computation or emphasizing expressive training data alone is insufficient to match GDPO's gains in our experiments. These results demonstrate the effectiveness of reinforcement learning for expressive dyadic listening head generation.

C.2 Ablation Studies on Reward Components

We investigate the contributions of the four reward components by comparing the supervised baseline with configurations containing one to four terms in Table 9. The velocity-variance term $\sigma(\nabla M)$ alone substantially improves motion dynamics, identifying it as a key contributor to expressiveness. In contrast, MSE alone achieves the lowest LVE and MHD but substantially worsens dynamic metrics, highlighting the limitations of optimizing frame-level reconstruction alone. Combining reward components introduces trade-offs between geometric accuracy and motion dynamics. The full four-component reward achieves the lowest speaking FDD and PDD and listening FDD, while maintaining competitive performance on the remaining metrics. Although individual configurations perform

Methods	Speaking					
	FDD (*expr*)↓	JDD (*jaw*)↓	PDD (*neck*)↓	EDD (*eyelid*)↓	GDD (*eyepose*)↓	RDD (*rot*)↓
GDPO-Listener	**23.56**	**1.26**	**2.85**	**37.72**	**7.55**	1.79
w/o Corresponding Item	31.03	1.33	3.46	53.63	9.89	**1.74**
Methods	Listening					
	FDD (*expr*)↓	JDD (*jaw*)↓	PDD (*neck*)↓	EDD (*eyelid*)↓	GDD (*eyepose*)↓	RDD (*rot*)↓
GDPO-Listener	**19.05**	**0.81**	**1.92**	**43.63**	**8.52**	**1.23**
w/o Corresponding Item	22.85	0.98	2.41	60.62	10.85	1.24

Table 10: Ablation Studies on Group-Specific Reward Items. Each ablation removes its column's reward item in the Post-Training. All variants use the same reduced post-training budget.

Methods	FD ↓			P-FD ↓			MSE ↓			SID ↑			rPCC ↓		
	Exp	Jaw	Neck	Exp	Jaw	Neck	Exp	Jaw	Neck	Exp	Jaw	Neck	Exp	Jaw	Neck
Speaking															
DualTalk	93.94	7.49	8.28	94.92	7.58	8.44	19.43	2.89	3.15	0.93	1.19	0.87	0.20	0.21	0.55
UniLS	40.07	6.23	7.03	40.94	6.26	7.21	8.50	2.35	2.84	1.27	**1.93**†	1.05†	0.27	0.25	0.52
AR-Flow (w/o GDPO)	29.80	5.17	9.29	30.65	5.24	9.44	6.37	1.93	3.38	2.20	1.15	0.69	0.17	0.27	0.53
GDPO-Listener	**23.75**	**3.14**	**5.87**	**24.92**	**3.21**	**5.99**	**5.51**	**1.31**	**2.23**	**2.26**	1.61	**1.07**	**0.15**	**0.18**	**0.49**
Listening															
DualTalk	94.05	7.16	9.91	95.06	7.28	10.08	19.16	2.63	3.53	0.69	0.68	0.48	0.19	0.28	0.53
UniLS	39.17	6.14	6.99	39.82	6.19	7.14	8.03	2.17	2.53	0.81	**0.97**†	**0.59**†	0.25	0.28	0.54
AR-Flow (w/o GDPO)	29.38	6.22	8.89	29.97	6.27	9.00	6.05	2.16	3.08	1.18	0.53	0.40	0.17	0.32	0.49
GDPO-Listener	**24.85**	**3.65**	**6.06**	**26.06**	**3.73**	**6.16**	**5.46**	**1.38**	**2.13**	**1.45**	0.87	0.57	**0.15**	**0.25**	**0.48**

Table 11: Additional Quantitative Analysis on Seamless Interaction Dataset. Following DualTalk, for better readability, FD, P-FD and MSE for pose are multiplied by 100; FD, P-FD and MSE for jaw are multiplied by 1000 and MSE for expression is multiplied by 10.

better on certain metrics, the complete reward provides a balanced combination of geometric fidelity and expressive dynamics across speaking and listening.

C.3 Ablation Studies on FLAME Group Items

We conduct leave-one-out ablations to evaluate the contribution of each FLAME parameter group's reward during post-training. Across six experiments, we disable the reward for one group while retaining the other five and keeping the output representation unchanged. All experiments in this table use the same reduced number of post-training iterations relative to the main results. Table 10 reports each ablation's corresponding metric. Alongside FDD, JDD, and PDD, we apply the same dynamic-deviation formulation to eyelid, eye-pose, and global rotation parameters, denoted EDD, GDD, and RDD, respectively. Removing a group's reward worsens its corresponding metric in all listening evaluations and five of six speaking evaluations. These results show that group-specific rewards generally improve the dynamics of their respective parameters, supporting their inclusion for comprehensive control of facial and head motion.

C.4 Evaluation Metrics

We evaluate speaking and listening segments separately using Voice Activity Detection (VAD). Speaking segments require active actor speech, whereas listening segments require the partner to be speaking while the actor remains silent. For speaking, we assess lip synchronization using Lip Vertex Error (LVE), which averages the maximum Euclidean error among corresponding lip vertices per frame, and overall geometric fidelity using Mean Head Distance (MHD), which averages Euclidean errors across corresponding head vertices and frames. We further evaluate motion dynamics using Upper-face Dynamic Deviation (FDD), head-pose Dynamic Deviation (PDD), and jaw Dynamic Deviation (JDD), computed from the absolute differences between predicted and ground-truth standard deviations of frame-to-frame feature velocities. For listening, we prioritize these dynamic metrics because direct vertex errors can penalize plausible reactions that differ from the recorded response. The same dynamic deviation formulation is also applied to eyelid, eye-pose, and global rotation parameters, yielding EDD, GDD, and RDD, respectively.

C.5 Additional Evaluation Metrics

Additional evaluation metrics are shown in Table 11, which reports the complementary metrics adopted from prior work on the Seamless Interaction dataset. Fréchet Distance (FD) measures motion distribution alignment; Paired Fréchet Distance (P-FD) assesses the joint distribution of speaker and listener behavior; Mean Squared Error (MSE) measures feature-level reconstruction accuracy; SID evaluates motion diversity; Residual Pearson Correlation Coefficient (rPCC) measures discrepancies in inter-person motion correlation. GDPO-Listener outperforms DualTalk on every reported metric for expression, jaw and neck in both speaking and listening. UniLS is evaluated without style or history conditioning, since both take ground-truth motion of the target speaker as input, and its single global head rotation is scored in the neck block. † marks the SID entries where UniLS scores are higher than or close to ours: it predicts only one jaw dimension and zeroes the rest, and its FD there is about twice ours, so the diversity is not matched by accuracy. Most importantly, FD, SID and rPCC are not part of the GDPO reward, yet adding GDPO to the Stage-1 AR-Flow improves all three. The policy therefore does not merely exploit its own reward, rather, it improves distributional, diversity and coordination metrics that are not part of the rewards.

C.6 Comparison Between GDPO and GRPO

We compare GDPO with standard GRPO to evaluate the contribution of reward-decoupled optimization beyond reinforcement learning alone. Standard GRPO normalizes aggregated rewards, whereas GDPO normalizes each FLAME parameter group's reward independently before combining the resulting advantages. This distinction is important because FLAME parameter groups exhibit different reward scales and distributions, potentially allowing larger-scale components to dominate the aggregated signal. Figure 6 shows that GDPO achieves higher training rewards than GRPO. Rewards are expressed as negative errors, so higher values indicate better performance. Table 12 further shows consistently lower errors across all reported speaking and listening metrics on the Seamless Interaction dataset. These results demonstrate the benefit of reward-decoupled optimization for balancing heterogeneous FLAME parameter groups in dyadic head generation.

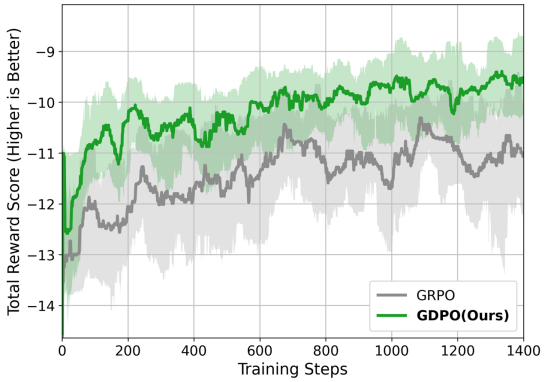

Figure 6: Training Reward Analysis of GDPO & GRPO.

Methods	Speaking			Listening		
	FDD ↓	PDD ↓	JDD ↓	FDD ↓	PDD ↓	JDD ↓
GDPO-Listener	**22.45**	**2.71**	**1.26**	**18.85**	**1.90**	**0.87**
GRPO	29.53	2.92	1.34	22.84	1.93	0.88

Table 12: Quantitative Analysis of GDPO & GRPO.

C.7 User Studies

We evaluated the perceptual quality of our generated speaking and listening head motions through a user study on Prolific. Instead of evaluating speaking and listening in isolation, the test videos feature continuous conversational interactions to better reflect real-world performance. A total of 42 native English speakers evaluated GDPO-Listener against L2L, DualTalk, and UniLS in 4-way forced-choice tasks. We assessed four metrics: *Lip Synchronization* (audio-mouth alignment), *Motion Naturalness* (human-like motion realism), *Liveliness* (dynamic, engaging behavior), and *Diversity* (temporal variety of facial expressions and head movements). Fig. 7 illustrates the study interface. In each task, participants watch a video featuring a reference partner alongside four anonymous generated avatars, presented in random order. Below the video, a forced-choice matrix requires users to select the best actor for each evaluation metric. As shown in Tab. 5, our method is consistently preferred over all baselines across all aspects. Our method is then shown to be superior at generating realistic, dynamic conversational interactions.

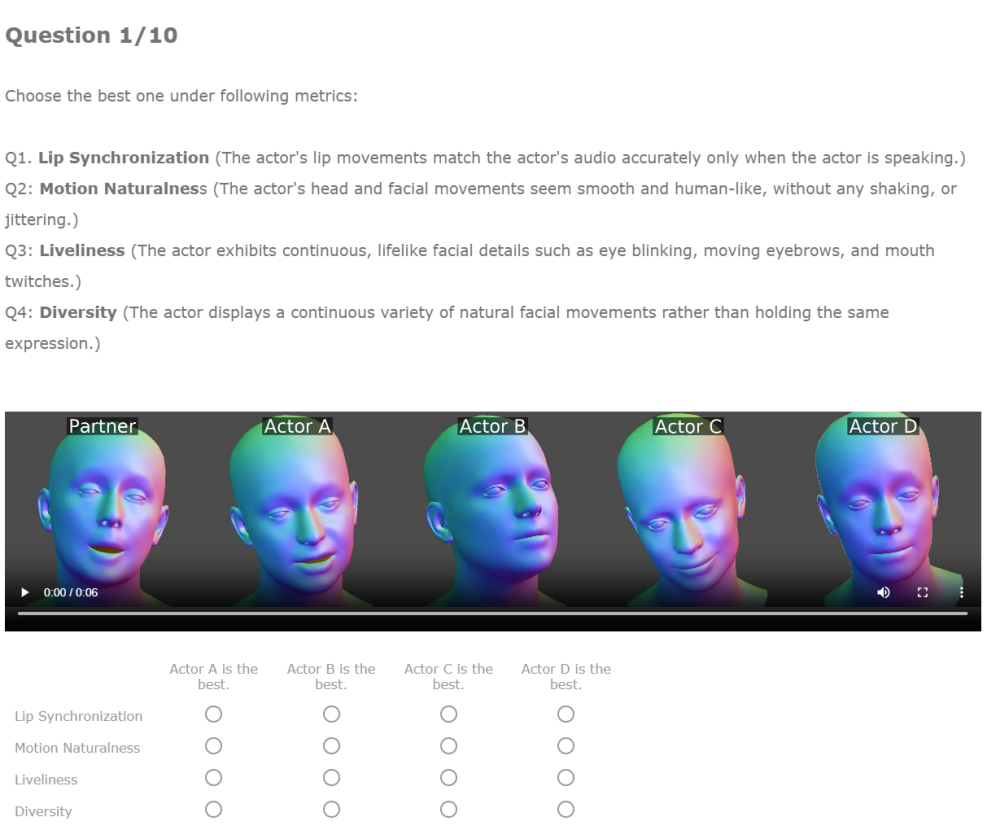

Figure 7: The Interface of Our User Study.

D Additional Visualizations

D.1 Extra Text Control Visualizations

Fig. 8 provides further qualitative results demonstrating our semantic text control. By conditioning on diverse prompts, GDPO-Listener accurately synthesizes the described emotional states and facial geometries. These visualizations highlight our model's capacity for fine-grained, decoupled control over conversational expressions.

D.2 Photorealistic Gaussian Avatar Integration

To visualize the generated motion beyond FLAME meshes, we integrate GDPO-Listener with GAGAvatar Chu & Harada (2024), which reconstructs animatable Gaussian head avatars from a single image. We use our predicted FLAME parameter sequences to drive these avatars and render photorealistic videos across different identities. Figure 9 presents qualitative comparisons with UniLS and DualTalk for both speaking and listening behaviors. In these examples, GDPO-Listener produces more expressive facial dynamics and varied head movements that remain evident in the rendered avatars, demonstrating its applicability to photorealistic conversational animation. Video comparisons are provided in the supplementary materials.

D.3 Video Demo

We strongly encourage readers to view the supplementary video to fully evaluate the temporal dynamics and visual quality of our method.

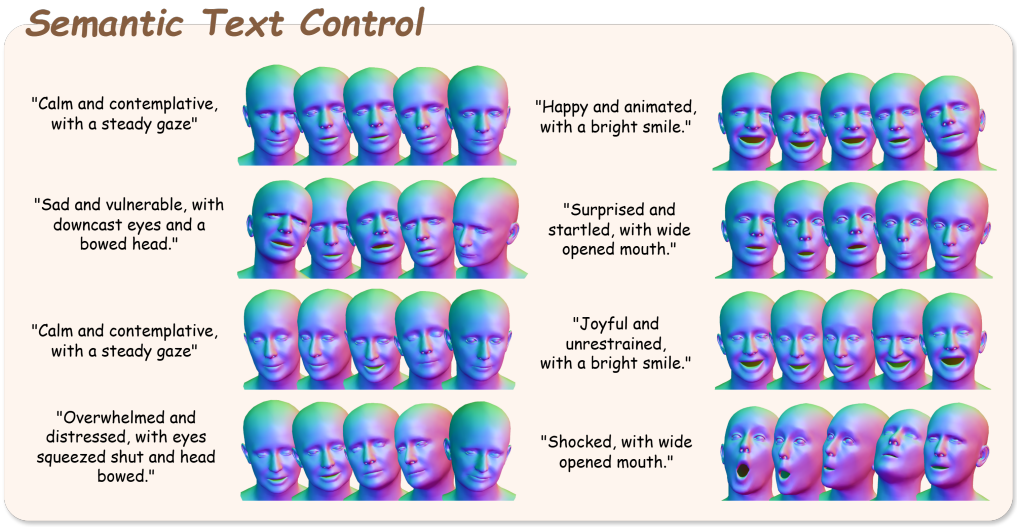

Figure 8: Additional Text Control Visualizations.

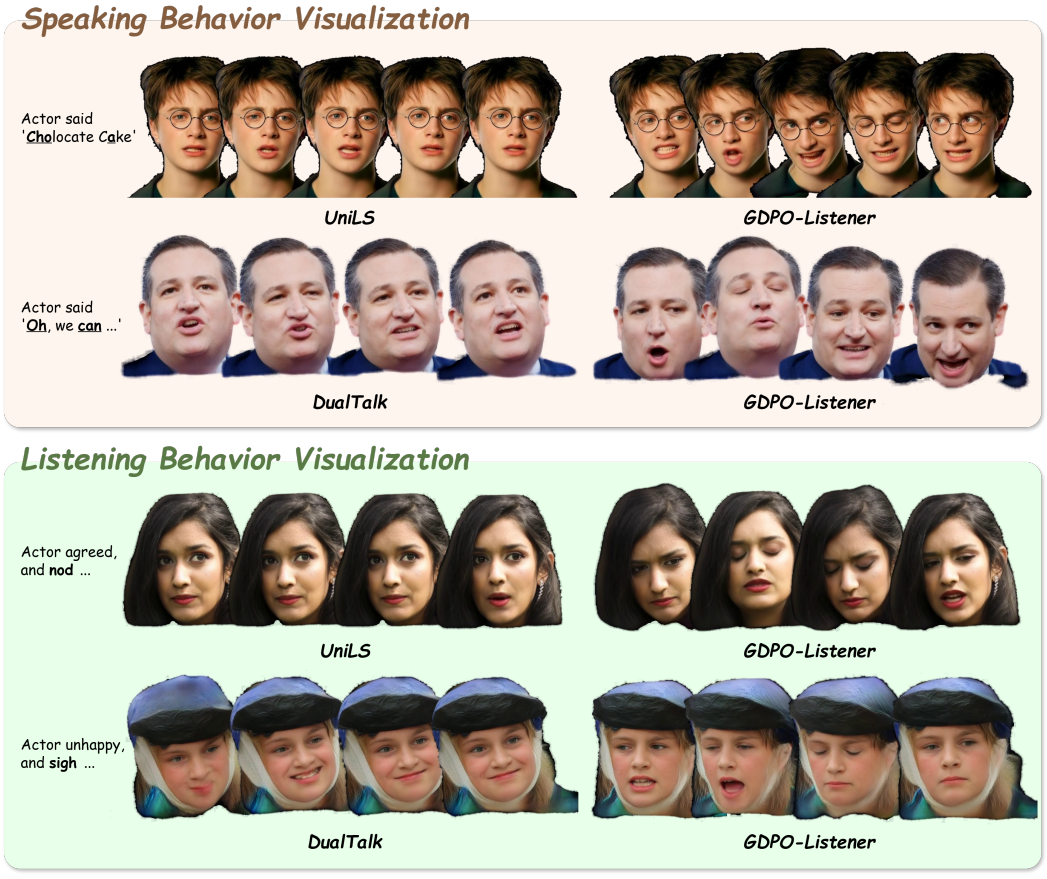

Figure 9: Photorealistic Gaussian Avatar Comparisons.

Qwen3-VL System Prompt
Describe the person's facial expression in a single concise sentence. 1. The specific Emotion: Use precise words like 'skeptical', 'amused', 'bored', 'attentive', 'confused', 'happy', 'angry', 'calm', or 'focused'. (Avoid generic 'neutral'). 2. The specific Facial Geometry and Gaze like 'narrowed eyes', 'raised eyebrows', 'head tilted', 'looking away'. **Rules:** 1. IGNORE mouth movements related to speech. 2. DO NOT use words like 'speaking', 'talking', or 'listening'. 3. Focus on the VIBE. Even if the face is mostly still, find the subtle emotion. **Examples:** 1. Skeptical and doubting, with one eyebrow raised and lips pursed. 2. Joyful and energetic, smiling broadly with crinkled eyes. 3. Bored and detached, looking away with a blank expression. 4. Attentive and focused, staring directly ahead with a slight nod.

Table 13: The explicit system prompt provided to Qwen3-VL to extract decoupled expressive text descriptions from our sampled 8-frame clips.

E Dataset Details

E.1 Dataset Collection

We utilize the Seamless Interactions dataset of dyadic conversational videos. For data processing, we apply MediaPipe for face detection, extract per-frame FLAME parameters using SHeaP Schoneveld et al. (2025), and apply Savitzky-Golay filtering Savitzky & Golay (1964) for temporal smoothing. We train on a 718 hours subset and evaluate on the full 31.5 hours test split, without introducing any external training data (unlike UniLS). To rigorously evaluate dynamic reactions, we curate an expressive test subset by filtering the test set based on the variance of ground-truth FLAME parameters. Additionally, we evaluate our framework on the DualTalk dataset.

E.2 Text Annotation Details

To enable our model with text control, we require high-quality, continuous descriptions of the actors' facial behaviors. We automate this annotation process using the Qwen3-VL vision-language model. For each 200-frame video clip in our dataset, we uniformly sample 8 frames to provide the model with sufficient temporal context while maintaining computational efficiency. We specifically design the prompt to decouple the actor's underlying emotional state and facial geometry from speech-related mouth movements. The model is instructed to output a single, concise sentence capturing specific emotions and gaze dynamics while strictly ignoring active articulation. The exact system prompt, including the critical constraints and few-shot examples provided to the model, is detailed in Table 13. For text-conditioned benchmark evaluation, prompts are generated from the target actor's reference video using the same Qwen3-VL annotation pipeline. The no-text ablation evaluates generation without these target-derived captions.

E.3 Dataset Statistics

The training data contain an uneven distribution of listener motion intensities. As shown in Fig. 10, listener head-pitch variance is concentrated at low values, with highly dynamic segments occurring less frequently. Consequently, the predominance of low-motion training examples can bias learning toward conservative listener responses. Under weak conversational conditioning, standard supervised objectives may provide insufficient incentive to capture the expressive behaviors found in less frequent training segments, resulting in stiff or low-variance motion. To alleviate this tendency, we apply RL post-training to explicitly reward listener dynamics. Stochastic rollouts enable exploration of alternative responses, while sequence-level rewards favor motion and velocity statistics that bet-

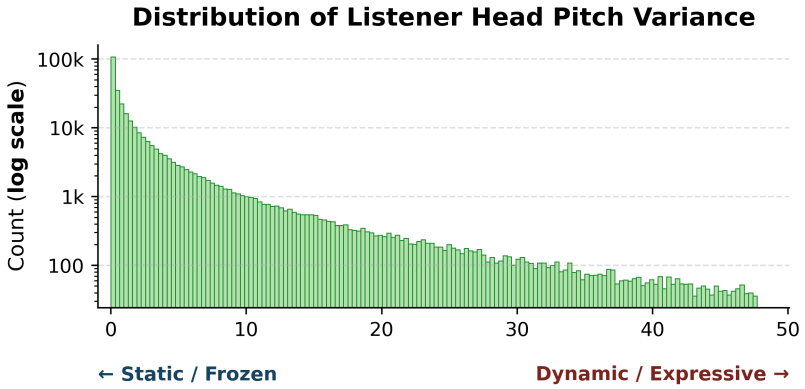

Figure 10: Unique Domain Bias of Static Listening Heads in Seamless Interaction Dataset

ter match the reference behavior. Our experiments show that this post-training improves listener expressiveness while maintaining motion fidelity.

F Implementation Details

F.1 Experiment settings

Our VAE ($d = 67$) compresses $L = 200$ frames of $N = 6$ FLAME parameter groups (`expr`, `jaw`, `neck`, `eyelid`, `eyepose`, `rot`). With downsampling $r = 8$, the latent length is 25. AR-Flow is a continuous Transformer (6 layers, 16 heads, $D = 512$) conditioned on frozen T5-XXL and Wav2Vec2.0-Large. We train on 8 L40s GPUs using AdamW. Stage 1 trains the base AR-Flow (lr=10^{-4}, batch size 32, 50 epochs). Stage 2 post-trains via GDPO (lr=5×10^{-6}, batch size 4, 2 epochs). For GDPO, we set group size $G = 4$, SDE variance $\sigma = 0.5$, KL penalty $\beta = 0.01$, replay chunk step to 2. We synchronize the behavior policy every 8 iterations for ratio computation, while keeping the reference policy fixed. To accelerate RL simulation, generation trajectories use 4 flow denoising steps during training, while 10 steps are used for inference. At inference, GDPO-Listener requires no external style embedding or ground-truth actor-motion prefix. Its autoregressive history consists of previously generated actor latents, alongside the available conversational conditioning.

F.2 ARFlow & GDPO Training Details

Stage 1: Auto-Regressive Flow Matching. The denoiser is causal and its velocity head reads only the terminal token, so one forward pass supervises exactly one latent position l of a sequence

$$S_l(x) = \left[\, c_{\text{txt}};\ a_{\text{ptn}};\ m_{\text{ptn}};\ a_{\text{act}};\ h_{<l};\ x \,\right], \qquad h_{<l} = \left[\, z_{\text{sos}};\ z_0;\ \ldots;\ z_{l-1} \,\right]. \tag{11}$$

Training all positions at once would make each of them read its predecessors' noisy copies instead of the clean history used at inference, so we supervise one randomly drawn position per pass, minimizing

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{l\sim\mathcal{U}\{0,\ldots,L_z-1\}}\, \mathbb{E}_{\tau\sim\mathcal{U}[0,1],\,\epsilon\sim\mathcal{N}(0,I)} \Big[\big\| v_\theta\big(S_l(x_\tau),\tau\big) - (z_l - \epsilon) \big\|_2^2 \Big], \quad x_\tau = (1-\tau)\epsilon + \tau z_l, \tag{12}$$

which is unbiased up to the constant $1/L_z$. The index l is shared by the batch and τ is drawn per sample, and the history is always the clean ground-truth prefix. For classifier-free guidance, each conditioning stream is dropped independently per sample with probability p_{drop}, its embeddings zeroed and its keys masked out, so one network covers every condition subset.

Stage 2: GDPO Post-Training. We maintain three networks: the trainable policy π_θ, a behavior policy $\pi_{\theta_{\text{old}}}$ periodically synchronized from π_θ, and the frozen Stage-1 reference policy π_{ref}. Dropout is disabled during rollout and replay. For each conditioning window, the behavior policy generates G trajectories using Eq. 7. We cache the sampled states and transitions, and compute rewards on the decoded motion without gradients. Let $H^j_{l,k}$ contain the conditioning inputs, cached

autoregressive history, current latent $x^j_{l,k}$, and flow time t_k. With K uniformly spaced integration steps and $\Delta t = 1/K$, the transition kernel of policy p is

$$q_p(\cdot \mid H^j_{l,k}) = \mathcal{N}\left(x^j_{l,k} + \Delta t\, v_p(H^j_{l,k}), \sigma^2 \Delta t\, I\right). \tag{13}$$

We define the normalized trajectory log-score as

$$s^j_p = -\frac{1}{KD_v} \sum_{l=0}^{L_z-1} \sum_{k=0}^{K-1} \frac{\left\| x^j_{l,k+1} - \left(x^j_{l,k} + \Delta t\, v_p(H^j_{l,k})\right)\right\|_2^2}{2\sigma^2 \Delta t}, \tag{14}$$

where L_z is the latent sequence length and D_v is the dimension of each latent vector. Policy-independent additive terms are omitted because they cancel in score differences. The initial noise distributions are also shared by all policies. Consequently, for a complete sampled trajectory τ^j conditioned on C^j,

$$\tilde{\rho}^j = \left(\frac{P_\theta(\tau^j \mid C^j)}{P_{\theta_{\text{old}}}(\tau^j \mid C^j)}\right)^{1/(KD_v)}. \tag{15}$$

This normalization produces a tempered ratio rather than the unscaled trajectory importance-sampling ratio. The reference penalty in Eq. 10 is defined by

$$\begin{aligned} \mathcal{K}^j(\theta) &= \frac{1}{KD_v} \sum_{l=0}^{L_z-1} \sum_{k=0}^{K-1} D_{\text{KL}}\left(q_\theta(\cdot \mid H^j_{l,k}) \,\|\, q_{\text{ref}}(\cdot \mid H^j_{l,k})\right) \\ &= \frac{\Delta t}{2\sigma^2 KD_v} \sum_{l=0}^{L_z-1} \sum_{k=0}^{K-1} \left\| v_\theta(H^j_{l,k}) - v_{\text{ref}}(H^j_{l,k})\right\|_2^2. \end{aligned} \tag{16}$$

This is a conditional transition regularizer evaluated at cached behavior-policy states. Reference velocities are computed without gradients and may be cached. During optimization, all cached states, autoregressive histories, advantages, and behavior-policy scores are held fixed. Define the active-branch indicator

$$\chi^j = \begin{cases} 1, & A^j \geq 0 \text{ and } \tilde{\rho}^j \leq 1+\epsilon, \\ 1, & A^j < 0 \text{ and } \tilde{\rho}^j \geq 1-\epsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

At clipping boundaries, this definition selects the unclipped branch's subgradient. The gradient of Eq. 10 is

$$\nabla_\theta \mathcal{L}_{\text{GDPO}} = \sum_{j=1}^{BG} c^j \nabla_\theta s^j_\theta + \frac{\beta}{BG} \sum_{j=1}^{BG} \nabla_\theta \mathcal{K}^j(\theta), \qquad c^j = -\frac{\tilde{\rho}^j A^j \chi^j}{BG}. \tag{18}$$

A first gradient-free replay computes the complete trajectory scores, ratios, and coefficients c^j. A second replay back-propagates

$$\sum_{j=1}^{BG} \text{stopgrad}(c^j) s^j_\theta + \frac{\beta}{BG} \sum_{j=1}^{BG} \mathcal{K}^j(\theta) \tag{19}$$

in chunks of n_c denoising transitions, accumulating the contributions of both terms over all chunks. No optimizer step is taken until all chunk gradients have been accumulated. This reproduces the gradient of Eq. 10 for the fixed cached batch while retaining at most n_c transition computation graphs at a time. Cached states and histories are detached, so neither the sampler nor the VAE decoder is differentiated.

Group rewards are winsorized within each group before normalization and the aggregated advantage of Eq. 9 is standardized once more over the batch, so the weights λ_i set relative priorities only, and $\lambda_i = 0$ realizes the leave-one-out ablations of Appendix C.3. Finally, an epoch is stopped early whenever the measured KL between $\pi_{\theta_{\text{old}}}$ and π_θ exceeds a threshold, and since the chunked backward is incompatible with the DDP reducer, gradients are all-reduced manually before clipping and the optimizer step.

F.3 Long Sequence Inference

We slide the training window along the sequence and seed each window with the last latent token of the previous one:

$$s_w = w\,(L_z - 1), \quad \hat{z}_{\,s_w+j} \;\sim\; \pi_\theta\big(\cdot \mid C^{(w)},\; \hat{z}_{\,s_w:\,s_w+j}\,\big), \quad j = 1, \ldots, L_z - 1, \tag{20}$$

where $C^{(w)}$ is the prefix of Eq. 11 sliced to frames $[\,rs_w,\; rs_w + L\,)$ with window-local TMRoPE indices, so the context never exceeds the trained length. The autoregression thus continues across boundaries instead of restarting from z_{sos}, and the windows are decoded separately and cross-faded on their overlap. Cost per window is constant, so the sequence length is unbounded.

F.4 Expressiveness Control

The guidance scale of Eq. 6 is a function of both the token and the flow time, which turns classifier-free guidance into an intensity slider:

$$\hat{v}_\theta = v_\emptyset + \omega(l,\tau)\,\big(v_{\text{cond}} - v_\emptyset\big), \qquad \omega(l,\tau) = \begin{cases} \omega_q^{\text{listen}}, & l \in \mathcal{L} \\ \omega_q^{\text{speak}}, & l \notin \mathcal{L} \end{cases} \quad \text{for } \tau \in [\tau_q, \tau_{q+1}), \tag{21}$$

where $\mathcal{L}$ is the set of listening tokens, obtained by down-sampling the VAD mask to the latent rate, and $\{[\tau_q, \tau_{q+1})\}$ is a partition of the flow interval. The conditional and unconditional velocities are evaluated in one batched forward pass; when all scales equal one the unconditional branch is skipped, which is the setting used for the automatic motion-metric benchmarks. Raising ω on the early segment, where the trajectory is still far from the data, amplifies the motion, while keeping $\omega = 1$ on the late segment preserves facial geometry; separating ω^{listen} from ω^{speak} lets reaction intensity be tuned without touching lip synchronization.

G Limitations & Future Work

GDPO-Listener focuses on facial and head motion within an expanded FLAME parameterization. Extending the framework to jointly model body posture and hand gestures would enable more comprehensive and coordinated nonverbal behavior. Also, our rewards target motion dynamics and geometric fidelity, while conversational intent and individual preferences are not explicitly assessed. Future work could explore multimodal LLM-based rewards or preference feedback collected from human raters to complement the current objectives and further improve the contextual appropriateness of generated reactions.