REDYFLOW: TEMPORAL-PYRAMID RESIDUAL FLOW MATCHING FOR DYADIC CONVERSATIONAL AVATAR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dyadic conversational avatar generation requires speech-aligned articulation and coordinated behaviors that are not uniquely determined by audio. We introduce **ReDyFlow**, a two-stage framework that jointly generates the speaking and listening head-motion sequences of both participants, without requiring observed partner motion as input. Given paired audio, speaking activity, and reference motion styles, a role-aware deterministic predictor constructs paired motion anchors through multi-scale temporal modeling and interaction between preliminary predictions. An anchor-conditioned joint flow then stochastically refines both trajectories with a temporal-pyramid velocity network that captures multi-resolution context and bidirectional participant interactions. We further introduce rollout-based dynamics supervision to regularize the temporal statistics of integrated samples. Experiments on the **UniLS** benchmark show that ReDyFlow achieves better results on the majority of quantitative metrics, while qualitative comparisons and user preferences indicate more natural speaking motion and listening reactions.

1 INTRODUCTION

Conversational avatars must coordinate the facial and head motions of both participants, rather than animate two isolated speakers. Speech-driven methods have advanced facial articulation and head animation Fan et al. (2022); Xing et al. (2023); Stan et al. (2023), but conversations additionally involve changing speaking roles and non-verbal responses to an interlocutor. We study the joint generation of paired 3D head trajectories, including facial expressions, articulation, head pose, and eye movements, where both participants' speaking and listening motions are generated simultaneously rather than conditioning on observed partner motion.

This task combines strongly constrained and ambiguous dynamics. Articulation closely follows speech, whereas expressive gestures and listener reactions admit multiple plausible realizations. Deterministic regression provides consistent predictions but can average uncertain responses, producing subdued motion. Stochastic generation better represents such variation, but must still preserve predictable speech-driven content. Moreover, local facial movements and longer conversational responses span different temporal scales and are coupled across participants. When neither participant's motion trajectory is observed at inference time, interaction cues must be inferred jointly from conversational conditions rather than obtained from observed partner motion.

We introduce **ReDyFlow**, a two-stage framework combining role-aware deterministic prediction with anchor-conditioned joint flow matching. Given paired audio, speaking-activity indicators, and reference motion styles, a shared predictor first constructs motion anchors for both participants. It routes self and partner audio according to conversational roles, models temporal dynamics at multiple scales, and exchanges preliminary motion predictions between the two participants. The resulting anchors therefore capture both speech-driven motion and dyadic interaction while jointly predicting the speaking and listening trajectories of both participants without observed motion input.

A conditional flow then jointly refines both trajectories from an anchor-centered stochastic source distribution. The deterministic anchors remain explicit conditions throughout generation, allowing the flow to preserve predictable motion while modeling ambiguous expressive and interaction dynamics. Its temporal-pyramid velocity network combines multi-resolution temporal context with

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.

bidirectional interaction between the evolving participant states. The hierarchy is embedded within a single joint flow rather than implemented as separate scale-specific generators. We further employ rollout-based dynamics supervision, which differentiates through integrated trajectories and regularizes role-wise motion variation and velocity statistics beyond local flow-field fitting.

We evaluate ReDyFlow on the **UniLS** benchmark through quantitative, qualitative, and perceptual comparisons of speaking and listening motion. Our contributions are:

- We introduce a joint dyadic head-generation framework that simultaneously generates both participants' speaking and listening trajectories without requiring observed partner motion. A role-aware deterministic predictor constructs paired motion anchors through conversational audio routing, multi-scale temporal modeling, and interaction between preliminary motion predictions.
- We propose an anchor-conditioned joint flow with a structured stochastic source, temporal-pyramid velocity modeling, and bidirectional interaction between evolving participant states. Rollout-based dynamics supervision further regularizes the temporal statistics of integrated motion samples.
- Extensive experiments on UniLS demonstrate improvements over existing methods on the majority of quantitative metrics. Qualitative comparisons show more natural speaking motion, listening reactions, and temporal variation, while a video-based user study shows clear perceptual preference for ReDyFlow over the compared baselines.

2 Related Work

2.1 Speech-Driven Facial and Head Animation

VOCA Cudeiro et al. (2019) learns speech-conditioned vertex displacements, while MeshTalk Richard et al. (2021) separates audio-correlated and audio-uncorrelated facial dynamics. FaceFormer Fan et al. (2022) uses autoregressive temporal modeling, and SelfTalk Peng et al. (2023a) strengthens cross-modal consistency through self-supervision. UniTalker Fan et al. (2024) scales training across datasets with heterogeneous facial annotations through a unified multi-head architecture. Learned priors and controllable representations support discrete motion prediction Xing et al. (2023), personalization Thambiraja et al. (2023), and expressive control Peng et al. (2023b). Geometry-aware approaches further improve facial-motion modeling: OT-Talk Wang et al. (2025) combines Chebyshev graph convolutions with optimal-transport-based geometric supervision, while GrambaTalk Wang et al. (2026) integrates hierarchical graph representations and bidirectional Mamba for joint spatial–temporal modeling. Probabilistic methods Stan et al. (2023); Sun et al. (2024); Yang et al. (2024) address the one-to-many relationship between speech and motion. DiffPoseTalk conditions facial and head-motion synthesis on reference style, while DEEPTalk Kim et al. (2025) combines probabilistic emotion embeddings with a temporally hierarchical discrete motion prior for expressive facial animation. These approaches primarily animate individual speakers rather than jointly model two conversational trajectories.

2.2 Listener and Dyadic Conversational Generation

Listener-generation methods synthesize non-verbal feedback from speaker cues Ng et al. (2022); Zhou et al. (2022), with extensions to emotion-conditioned responses Song et al. (2023), transcript-based gesture prediction Ng et al. (2023), and textual control Liu et al. (2024). Their primary task is listener response, rather than joint generation across changing speaking roles.

Unified conversational models differ in their access to partner motion. DualTalk Peng et al. (2025) conditions target motion on paired audio and observed interlocutor motion, while TIMAR Chen et al. (2025) retains user head motion within interleaved audio–visual turn modeling. INFP Zhu et al. (2025), UniLS Chu et al. (2026), and MANGO Zhu et al. (2026) learn talking-and-listening behaviors from paired audio without requiring synchronized partner-motion observations. UniLS adapts an audio-free autoregressive motion prior using separate self- and partner-audio conditioning, whereas MANGO combines dual-audio diffusion with image-level supervision. Talking Together Shan et al. (2026) additionally models relative spatial configuration and mutual gaze from mixed conversational audio.

Learn2Chat Huang et al. (2026) modulates pretrained monologic motion priors through stochastic interaction latents predicted from paired speech. ReDyFlow instead constructs role-aware paired anchors that already include predicted-motion interaction, then jointly evolves both trajectories through a motion-space flow. Our reference-style conditioning is distinct from observing a synchronized partner trajectory.

2.3 Prior-Guided and Multi-Scale Motion Generation

Flow matching Lipman et al. (2023) learns time-dependent vector fields for sample generation through numerical integration. Conversational applications include flow-based talking-head video synthesis Chen & Liu (2025) and facial-motion generation conditioned on evolving conversational states Chen et al. (2026). SubtleTalk Ding et al. (2026) combines a deterministic prior with residual flow matching for single-speaker facial and head dynamics. Wei et al. Wei et al. (2026) use coarse motion priors and cross-person interaction for residual-flow-based skeletal-motion forecasting from observed history. Related interactive generation also includes InterGen Liang et al. (2024), which exchanges evolving participant features through mutual attention. ReDyFlow builds on prior-guided generation to jointly refine conversational head motions around speech- and role-conditioned paired anchors.

Multi-scale modeling is explored through ARTalk's motion codebook Chu et al. (2025), pyramidal video-generation stages Jin et al. (2025), and Laplacian image-residual generation Zhao et al. (2026). In contrast to target-space pyramid decompositions, our temporal hierarchy operates within one velocity network and predicts a single full-resolution paired-motion field.

Dynamic supervision also complements generative objectives. MDM Tevet et al. (2023) enables geometric and velocity losses through clean-motion prediction, while SubtleTalk Ding et al. (2026) regularizes reconstructed motion with velocity, smoothness, and standard-deviation losses. Our supervision explicitly differentiates through flow rollouts and constrains role-wise temporal statistics of the integrated paired trajectories, complementing local flow-field fitting with supervision of sampled motion dynamics.

3 Method

3.1 Problem Formulation and Overview

Given paired audio $\boldsymbol{A}_i$, speaking activity $\boldsymbol{q}_i$, and reference motion clips $\boldsymbol{S}_i$, we jointly generate both participants' head motions under conditions $\mathcal{C} = \{\boldsymbol{A}_i, \boldsymbol{q}_i, \boldsymbol{S}_i\}_{i=1}^{2}$. Each frame uses the 108-dimensional UniLS FLAME representation $\boldsymbol{m}_{i,t} = [\boldsymbol{e}_{i,t}, \boldsymbol{p}_{i,t}, j_{i,t}, \boldsymbol{o}_{i,t}]$, with 100 expression, three head-pose, one jaw, and four eye parameters. Both stages operate on $T = 100$ frames and share channel-wise normalization $\mathcal{N}(\boldsymbol{M}) = (\boldsymbol{M} - \boldsymbol{\mu}) \oslash \boldsymbol{\sigma}$ computed from training data. We denote normalized targets by $\boldsymbol{Y} = \mathcal{N}(\boldsymbol{M})$.

As shown in Fig. 1, a deterministic predictor first constructs paired anchors $\boldsymbol{B} = D_\theta(\mathcal{C})$, followed by joint flow-based refinement:

$$\hat{\boldsymbol{M}}_i = \mathcal{N}^{-1}(\boldsymbol{B}_i + \hat{\boldsymbol{R}}_i), \qquad i \in \{1, 2\}. \tag{1}$$

Here $\hat{\boldsymbol{R}}_i$ denotes refinement in normalized motion space. The anchors already incorporate role-dependent motion and predicted partner interaction, and remain explicit conditions during Stage 2. Neither stage requires synchronized ground-truth partner motion as an interaction input.

3.2 Role-Aware Deterministic Motion Anchors

Condition representations. A shared frozen HuBERT backbone encodes both audio streams. Two learned softmax mixtures of its intermediate layers produce articulation- and response-oriented features, aligned to motion frames. Five waveform descriptors—log-RMS, relative log-RMS, its temporal change, zero-crossing rate, and relative difference energy—additionally condition the response pathway. A role encoder combines joint speaking states, activity transitions, and turn-duration features, including a flag for an initially observed state with unknown onset. Durations are computed from full-source masks before window extraction, rather than reset at window boundaries.

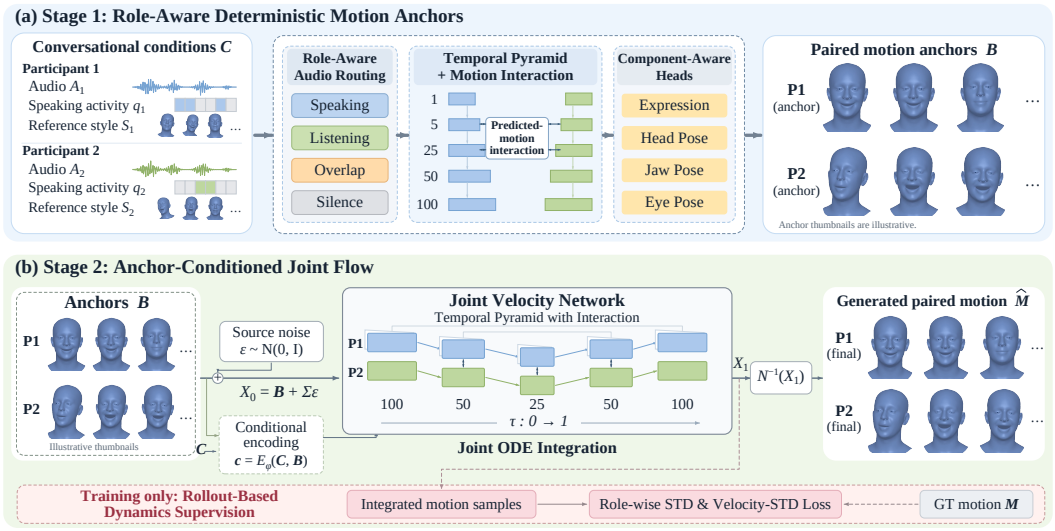

Figure 1: **Overview of ReDyFlow.** **(a)** Role-aware audio routing, temporal-pyramid modeling, and predicted-motion interaction construct paired motion anchors $\boldsymbol{B}$ from audio, speaking activity, and reference styles. **(b)** Conditioned on these inputs and anchors, a joint flow evolves both participants from $\boldsymbol{X}_0 = \boldsymbol{B} + \boldsymbol{\Sigma}\boldsymbol{\epsilon}$ through a temporal-pyramid velocity network with bidirectional interaction. The deterministic predictor remains frozen during Stage 2. The dashed training-only branch supervises role-wise motion and velocity statistics of integrated samples. Anchor thumbnails are illustrative.

A style encoder combines attention-pooled temporal features with the mean, standard deviation, and velocity standard deviation of normalized reference motion. We denote the role and style embeddings by $\boldsymbol{r}_i$ and $\boldsymbol{z}_i$.

Speech-activity-grounded routing. For participant i and partner $j = 3 - i$, lightweight branches construct features $\boldsymbol{h}_i^r$ for $r \in \mathcal{R} = \{\mathrm{spk}, \mathrm{lst}, \mathrm{ovl}, \mathrm{sil}\}$. Speaking and listening branches use self and partner audio, respectively, together with role and style. The silence branch uses role and style, while the overlap branch fuses the other branch features with role context.

With smoothed activity masks $\bar{q}_i$, the four routing priors are $\bar{q}_i(1-\bar{q}_j)$, $(1-\bar{q}_i)\bar{q}_j$, $\bar{q}_i\bar{q}_j$, and $(1-\bar{q}_i)(1-\bar{q}_j)$. Stacking these as $\boldsymbol{p}_{i,t}^0$, we obtain

$$\begin{aligned} \boldsymbol{\pi}_{i,t} &= \mathrm{softmax}\big(\log \max(\boldsymbol{p}_{i,t}^0, \epsilon) + \eta \tanh g(\boldsymbol{r}_{i,t})\big)\,, \\ \boldsymbol{h}_{i,t} &= \sum_{r \in \mathcal{R}} \pi_{i,t}^r \boldsymbol{h}_{i,t}^r. \end{aligned} \tag{2}$$

The output of g is zero-initialized; its bounded correction preserves the influence of the activity priors. Subsequent temporal modeling is shared across roles.

Temporal pyramid and predicted-motion interaction. At token counts $\mathcal{S}_1 = \{1, 5, 25, 50, 100\}$, average pooling $\mathcal{P}_s$, style modulation, and temporal blocks yield

$$\boldsymbol{H}_i^s = \mathcal{T}_s(\mathrm{FiLM}_s(\mathcal{P}_s(\boldsymbol{h}_i + \boldsymbol{r}_i), \boldsymbol{z}_i))\,. \tag{3}$$

The blocks combine self-attention and temporal convolutions. Parallel heads decode coarse motion from each role branch at $s \in \{5, 25\}$; pooled, renormalized routing weights combine them into preliminary motions $\boldsymbol{C}_i^s$.

Bidirectional cross-attention relates coarse motion components, temporal differences, and role context across participants, producing interaction features $\boldsymbol{u}_i^s$ and gated corrections $\boldsymbol{d}_i^s$. Activity-dependent priors emphasize listening over self-only speaking. The corrections enter the pyramid through

$$\widetilde{\boldsymbol{H}}_i^s = \mathrm{LN}(\boldsymbol{H}_i^s + \boldsymbol{u}_i^s + \boldsymbol{W}_s \boldsymbol{d}_i^s)\,, \quad s \in \{5, 25\}. \tag{4}$$

Other scales remain unchanged. Coarse motions are detached as interaction inputs but their heads receive auxiliary supervision. Top-down fusion propagates interaction context through the pyramid from coarse to fine, yielding features $\boldsymbol{F}_i^s$. Motion-level cues are thus derived from preliminary predictions rather than supplied partner trajectories.

Component-aware decoding. Expression, pose, and eyes use scale subsets $\mathcal{S}_e = \{1, 5, 25, 50, 100\}$, $\mathcal{S}_p = \{1, 5, 25, 50\}$, and $\mathcal{S}_o = \{5, 25, 50, 100\}$:

$$\boldsymbol{f}_{i,t}^k = \mathrm{LN}\left(\sum_{s \in \mathcal{S}_k} \alpha_{i,t}^{k,s} \boldsymbol{W}_{k,s} \mathcal{U}_T(\boldsymbol{F}_i^s)_t \right). \tag{5}$$

Here $\mathcal{U}_T$ interpolates features to full resolution. Scale weights are predicted from separately pooled speaking and non-speaking context within each window, then blended over time using smoothed activity masks. Expression and pose use separate temporal decoding heads. The jaw head combines fine-scale features at $\{25, 50, 100\}$ with expression, role, style, and an activity-gated self-audio pathway. The eye head additionally uses predicted head pose and its temporal difference to predict common and differential binocular parameters, $\boldsymbol{g} = (\boldsymbol{o}^L + \boldsymbol{o}^R)/2$ and $\boldsymbol{v} = (\boldsymbol{o}^L - \boldsymbol{o}^R)/2$ in raw eye space. The inverse mapping $\boldsymbol{o}^L = \boldsymbol{g} + \boldsymbol{v}$ and $\boldsymbol{o}^R = \boldsymbol{g} - \boldsymbol{v}$ recovers both eyes before normalization. This parameterization does not impose an explicit mutual-gaze constraint. Component outputs are concatenated into $\boldsymbol{B}_i$.

3.3 Anchor-Conditioned Joint Flow Matching

We freeze D_θ and learn a joint flow in normalized motion space. Given $\boldsymbol{B} \in \mathbb{R}^{2 \times T \times 108}$, we sample

$$\begin{aligned} \boldsymbol{X}_0 &= \boldsymbol{B} + \boldsymbol{\Sigma}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \\ \boldsymbol{X}_\tau &= (1-\tau)\boldsymbol{X}_0 + \tau \boldsymbol{Y}, \qquad \tau \sim \mathcal{U}[0,1]. \end{aligned} \tag{6}$$

Fixed component-wise noise scales $\boldsymbol{\Sigma}$ are shared across participants and frames. Participants share flow time τ, distinct from physical time, but use independently sampled noise entries. Following flow matching Lipman et al. (2023), we minimize

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathcal{D},\tau,\boldsymbol{\epsilon}} \left[\| v_\phi(\boldsymbol{X}_\tau, \tau; \mathcal{C}, \boldsymbol{B}) - (\boldsymbol{Y} - \boldsymbol{X}_0) \|_{W,\boldsymbol{q}}^2 \right]. \tag{7}$$

The error averages channels within each component and frames within each participant's speaking/non-speaking subset, then normalizes the weighted sum over valid participant–component–role groups.

Since $\boldsymbol{B}$ is fixed, translating to $\boldsymbol{R}_\tau = \boldsymbol{X}_\tau - \boldsymbol{B}$ gives a residual path from $\boldsymbol{\Sigma}\boldsymbol{\epsilon}$ to $\boldsymbol{Y} - \boldsymbol{B}$ with the same velocity. The anchors therefore serve as both source centers and explicit conditions. This coordinate change does not semantically separate speech-driven and interaction-dependent motion: all components remain refinable.

3.4 Temporal-Pyramid Joint Velocity Network

Generation conditions. Stage 2 reuses frozen HuBERT states but learns its own layer mixtures, audio adapters, and style encoder. Each participant's context combines self articulation/response features, partner response features, its anchor projection, joint speaking activity, full-source turn context, and reference mean, standard deviation, and velocity standard deviation. Availability embeddings identify missing conditions during motion-only pretraining. We denote the encoded context by $\boldsymbol{c} = E_\phi(\mathcal{C}, \boldsymbol{B})$.

Multi-resolution velocity prediction. Projected motion states and temporal positional embeddings enter an encoder–decoder with $100 \to 50 \to 25 \to 50 \to 100$ tokens. Downsampling uses average pooling; upsampling uses linear interpolation and projected concatenation with encoder features. Each block combines condition- and flow-time-modulated self-attention, depthwise temporal convolution, and a feed-forward network, with context resized to its resolution.

At the 50- and 25-token levels, including the 50-token decoder stage, bidirectional attention exchanges evolving participant states:

$$\begin{aligned} \boldsymbol{m}_i^s &= \mathrm{Attn}_s(\bar{\boldsymbol{H}}_i^s, \bar{\boldsymbol{H}}_j^s, \bar{\boldsymbol{H}}_j^s), \\ \boldsymbol{H}_i^s &\leftarrow \boldsymbol{H}_i^s + \mathrm{sigmoid}(G_s(\boldsymbol{c}_i^s + \boldsymbol{e}_\tau)) \odot \boldsymbol{m}_i^s, \end{aligned} \tag{8}$$

where $\bar{\boldsymbol{H}} = \mathrm{LN}(\boldsymbol{H})$ and $\boldsymbol{e}_\tau$ is the flow-time embedding. Both directions share parameters and update simultaneously from the same pre-update states, without participant-ID embeddings. Component-specific, condition-modulated heads predict expression, pose, jaw, and eye velocities, concatenated into one field $v_\phi \in \mathbb{R}^{2\times T\times 108}$. The pyramid thus operates inside a single joint flow, rather than generating separate residual targets at different scales.

3.5 Training Objectives and Rollout Dynamics

Deterministic-anchor supervision. We optimize $\mathcal{L}_{\text{anchor}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{stats}}$ with component and role weights. Reconstruction uses smooth-L_1 for expression and squared error for pose, jaw, and eyes in normalized space. Temporal terms supervise first/second differences and lag-$\{4, 8, 16\}$ displacements within unchanged speaking/non-speaking intervals. Auxiliary terms supervise multi-scale and branch/coarse predictions, common/differential binocular parameters, derived directions, and their temporal changes, and penalize interaction corrections. Statistical terms match role-wise normalized motion standard deviations, with additional raw-space pose/jaw terms. These losses require no FLAME mesh decoding.

Local endpoint supervision. A weak smooth-L_1 loss supervises the endpoint estimate from a single velocity evaluation:

$$\begin{aligned} \widetilde{\boldsymbol{Y}}_\tau &= \boldsymbol{X}_\tau + (1-\tau) v_\phi(\boldsymbol{X}_\tau, \tau; \mathcal{C}, \boldsymbol{B}), \\ \mathcal{L}_{\text{ep}} &= \mathbb{E}\Big[\rho_{W,\boldsymbol{q}}(\widetilde{\boldsymbol{Y}}_\tau - \boldsymbol{Y})\Big], \end{aligned} \tag{9}$$

using the same component/role aggregation as the flow loss.

Dynamics of integrated samples. During refinement, fresh source noise is integrated through the flow with gradients to obtain $\hat{\boldsymbol{Y}}^{\text{roll}}$, unlike the local estimate above. We denormalize it as $\hat{\boldsymbol{M}}^{\text{roll}} = \mathcal{N}^{-1}(\hat{\boldsymbol{Y}}^{\text{roll}})$ and define $\mathcal{D}_m(\boldsymbol{U}, \boldsymbol{V}) = \|\,\mathrm{Std}_m(\boldsymbol{U}) - \mathrm{Std}_m(\boldsymbol{V})\|_1$, where Std_m is the stabilized, channel-wise masked sample standard deviation. For role masks m_i^r and component dimensions d_k, motion variation is supervised by

$$\mathcal{L}_{\text{std}} = \mathrm{Avg}_{r,i}^{\text{valid}} \left[\sum_k \frac{1}{d_k} \mathcal{D}_{m_i^r}(\hat{\boldsymbol{M}}_i^{\text{roll},k}, \boldsymbol{M}_i^k) \right]. \tag{10}$$

Roles here are speaking and non-speaking, the latter including mutual silence. The operator averages valid participants within each role, then available roles and training pairs, requiring eight frames per role. Velocity statistics retain only true adjacent same-role pairs, $a_{i,t}^r = m_{i,t}^r m_{i,t+1}^r$:

$$\mathcal{L}_{\text{vstd}} = \mathrm{Avg}_{r,i}^{\text{valid}} \left[\frac{1}{108} \mathcal{D}_{a_i^r}(\Delta\hat{\boldsymbol{M}}_i^{\text{roll}}, \Delta\boldsymbol{M}_i) \right]. \tag{11}$$

Velocity contributions with fewer than eight valid differences are zeroed without discarding valid motion-variation terms. Disjoint role intervals are never concatenated to form differences. These losses constrain raw-parameter dynamics rather than directly optimizing vertex-space benchmark metrics. The combined Stage 2 objective is

$$\begin{aligned} \mathcal{L}_{\text{flow}} = \mathcal{L}_{\text{FM}} &+ \lambda_{\text{ep}} \mathcal{L}_{\text{ep}} \\ &+ \chi_{\text{roll}} \lambda_{\text{dyn}} \left(\mathcal{L}_{\text{std}} + \lambda_{\text{vstd}} \mathcal{L}_{\text{vstd}} \right), \end{aligned} \tag{12}$$

where χ_{roll} activates rollout supervision on selected refinement updates. Its statistics are evaluated after multiple predicted-state updates, rather than only on the single-step endpoint estimate.

3.6 Optimization and Inference

Stage 1 first trains without coarse-motion interaction, then trains only the interaction and feature-injection modules, followed by joint fine-tuning with HuBERT frozen. Stage 2 pretrains on single-person `TalkMotionOnly` data using zero anchors, missing-condition embeddings, and disabled dyadic interaction. Paired training then uses frozen Stage 1 anchors, followed by refinement with

Table 1: Quantitative comparison and ablation study on the UniLS Seamless Interaction test split. All metrics are lower-is-better. Baselines reported by UniLS Chu et al. (2026) appear above ReDyFlow (full), and ablation variants appear below. The best and second-best results in each column, ranked across all rows including ablations, are shown in **bold** and underlined, respectively.

	Speaking					Listening				
Method	LVE↓	MHD↓	FDD↓	PDD↓	JDD↓	FDD↓	PDD↓	JDD↓	F-FID↓	P-FID↓
DiffPoseTalk	9.48	2.96	32.66	7.89	1.40	–	–	–	–	–
ARTalk	7.46	2.12	31.64	7.66	1.19	–	–	–	–	–
ARTalk*	6.79	2.02	27.41	8.55	0.81	30.62	9.52	1.53	10.779	0.072
DualTalk	6.35	1.95	37.46	9.70	1.02	43.58	10.71	2.02	13.143	0.079
UniLS	5.83	1.89	**18.41**	4.67	0.71	**17.12**	4.75	**0.98**	4.304	0.038
ReDyFlow (full)	5.68	1.83	20.63	**4.55**	**0.66**	18.64	**4.58**	1.13	**0.229**	**0.0046**
w/o roll-out	5.78	1.91	21.84	4.56	0.71	20.87	4.96	1.15	0.290	0.005
w/o deterministic stage	7.49	2.18	28.86	5.74	1.01	26.54	5.75	1.40	1.270	0.037
w/o probabilistic stage	**4.90**	**1.62**	33.70	7.59	0.72	37.61	8.74	1.54	0.880	0.027
w/o interaction	5.89	1.92	26.78	5.66	0.82	22.47	5.21	1.25	1.020	0.032
w/o extra motion-only data	5.85	1.88	23.02	4.58	0.67	21.02	5.08	1.15	0.410	0.007
w/o temporal pyramid	6.17	1.97	25.48	5.61	0.87	23.94	6.11	1.40	0.730	0.009

rollout supervision. Motion-only pretraining initializes the same flow network, without adding an inference-time model.

We use 25-FPS motion, 16-kHz audio, 768-dimensional features, and eight main temporal-attention heads. Flow block depths are $(1, 1, 2, 1, 1)$; expression/pose/jaw/eye source scales are $(1.0, 1.0, 0.35, 0.7)$ in normalized space. Loss weights are $\lambda_{\text{ep}} = 0.02$, $\lambda_{\text{dyn}} = 0.10$, and $\lambda_{\text{vstd}} = 0.10$. Four-step training rollouts use one pair per GPU every 16 updates. The flow network uses no dropout, so rollouts and inference evaluate the same conditioned vector field.

Inference starts from $\boldsymbol{X}^{(0)} = \boldsymbol{B} + \boldsymbol{\Sigma}\boldsymbol{\epsilon}$ and jointly updates both participants with $N = 10$ Euler steps:

$$\boldsymbol{X}^{(n+1)} = \boldsymbol{X}^{(n)} + \frac{1}{N} v_\phi\Big(\boldsymbol{X}^{(n)}, \frac{n}{N}; \mathcal{C}, \boldsymbol{B}\Big). \tag{13}$$

The output is $\mathcal{N}^{-1}(\boldsymbol{X}^{(N)})$, with refinement $\hat{\boldsymbol{R}} = \boldsymbol{X}^{(N)} - \boldsymbol{B}$. Full sequences use non-overlapping 100-frame windows with fixed reference styles and full-source turn features. The final window is padded and trimmed; outputs are concatenated without previous-generated-motion conditioning or cross-window blending.

4 Experiments

4.1 Experimental Setup

We follow UniLS Chu et al. (2026) for the Seamless Interaction splits, preprocessing, baselines, and evaluation metrics. For style-conditioned inference, reference motion clips are selected using the same protocol as UniLS. Baseline results are taken from UniLS rather than independent retraining. ReDyFlow jointly generates both participants and is evaluated separately on speaking and listening frames using the metrics in Table 1, all lower-is-better.

4.2 Quantitative Comparison and Ablation

Comparison with baselines. ReDyFlow achieves the best results among complete methods on seven of ten metrics (Table 1), covering speaking LVE/MHD/PDD/JDD and listening PDD/F-FID/P-FID. Compared with UniLS, speaking LVE decreases from 5.83 to 5.68 and listening PDD from 4.75 to 4.58, while F-FID/P-FID improve substantially from 4.304/0.038 to 0.229/0.0046. UniLS remains better on speaking and listening FDD and listening JDD, indicating that ReDyFlow improves most reconstruction and distributional measures without uniformly dominating all dynamics metrics.

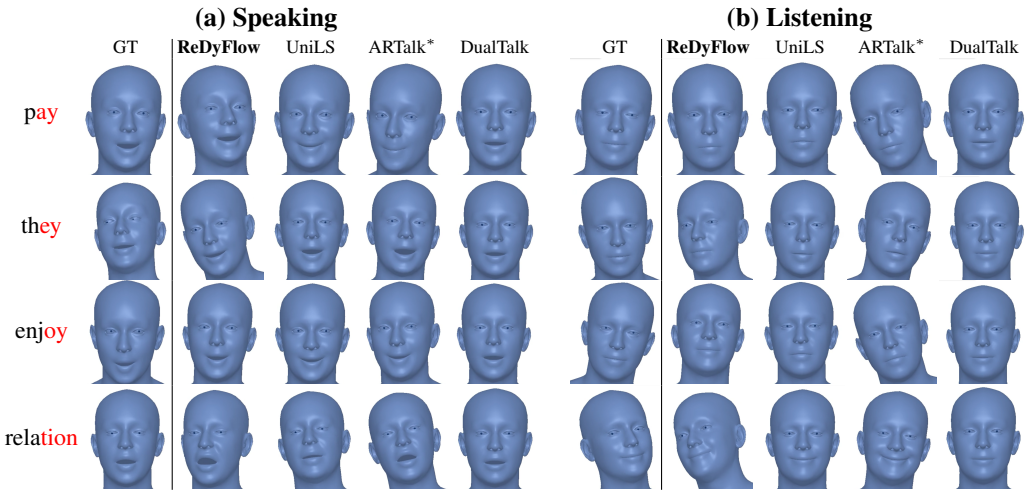

Figure 2: **Word-level comparison.** Speaking (left) and listening (right) at matched timestamps with identical rendering. Listening examples show concurrent responses to the interlocutor's indicated word.

Two stages and roll-out supervision. Removing either generation stage leads to clear degradation. Without Stage 1 anchors, all ten metrics worsen; using Stage 1 alone achieves the lowest LVE/MHD but substantially weaker motion dynamics and distribution matching. For example, adding Stage 2 reduces speaking/listening FDD from 33.70/37.61 to 20.63/18.64. Removing roll-out supervision also consistently degrades all ten metrics, including listening F-FID from 0.229 to 0.290, supporting its use for regularizing integrated motion dynamics.

Interaction and temporal hierarchy. Removing either cross-participant interaction or temporal-pyramid modeling from both stages worsens all metrics. Without interaction, listening F-FID/P-FID increase to 1.020/0.032, while removing the temporal hierarchy raises speaking/listening FDD to 25.48/23.94. These results support joint interaction modeling and multi-resolution temporal reasoning throughout the framework.

Additional motion-only data. Excluding the UniLS `TalkMotionOnly` subset while retaining motion-prior pretraining also degrades all ten metrics; for example, listening F-FID/P-FID increase from 0.229/0.0046 to 0.410/0.007. This confirms that additional motion-only data provide complementary benefits beyond prior learning from the paired training split.

4.3 Qualitative Evaluation and User Study

Visual comparisons. In the word-level examples (Fig. 2), ReDyFlow produces speaking lip shapes closest to ground truth among the compared methods, together with more natural facial expressions and head poses. Its listening reactions also most closely resemble the reference states in these examples. However, head pose, expression, and listener responses are not uniquely determined by audio, so resemblance to a single ground-truth realization should not be the sole criterion for judging naturalness. The timeline example (Fig. 3) further shows that ReDyFlow's changes in facial expression and head orientation most closely follow those of ground truth in the illustrated conversation, whereas some baselines exhibit nearly unchanged states across successive timestamps. Additional image comparisons are provided in Appendix F. Since static snapshots only partially convey motion quality, we encourage viewing the supplementary videos for a fuller comparison of speech–motion synchronization, temporal continuity, and reaction timing. Our user study likewise assesses preferences through video viewing rather than static-image comparisons.

User study. We conduct an A/B video preference study with 27 participants, comparing ReDyFlow against DualTalk, ARTalk*, UniLS, and ground truth. For each comparison, participants view 12 paired videos and provide judgments of lip synchronization, expression naturalness,

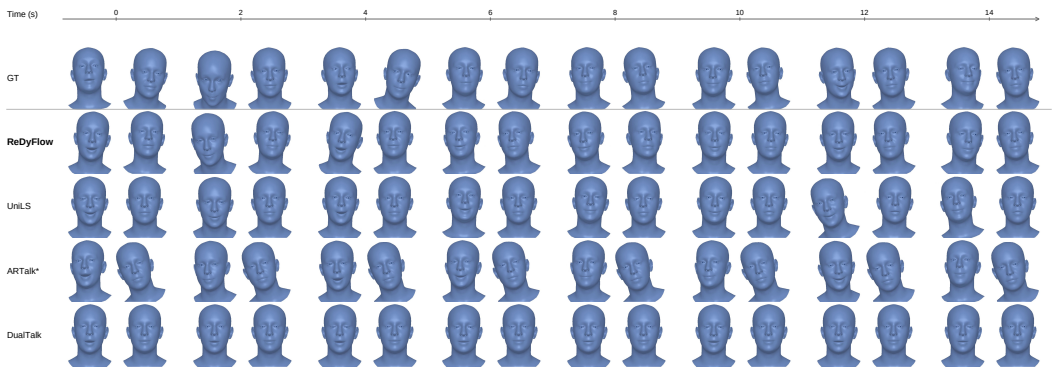

Figure 3: **Temporal comparison.** Methods are shown by row and timestamps by column, sampled every two seconds. Each cell retains both participants' original ordering, with a fixed crop across methods.

Table 2: User preference for ReDyFlow (%). Percentages are computed over non-tie responses after participants viewed all 12 video pairs for each comparison. Higher is better.

Comparison	**Sync↑**	**Exp↑**	**React↑**	**Pose↑**
vs. DualTalk	77.78%	85.19%	76.92%	92.59%
vs. ARTalk*	62.96%	65.38%	70.37%	66.67%
vs. UniLS	55.56%	59.26%	59.26%	62.96%
vs. Ground truth	37.04%	25.93%	57.69%	37.04%

reaction naturalness, and head-pose naturalness. This results in $12 \times 4 = 48$ video pairs viewed by each participant. Responses indicating that the two methods are equally good are excluded. Full protocol details are provided in Appendix D.

As shown in Table 2, ReDyFlow is consistently preferred to three generated baselines. The strongest preference is observed against DualTalk, with ReDyFlow receiving 77.78%/85.19%/76.92%/92.59% of valid votes for synchronization, expression, reaction, and pose, respectively. ReDyFlow is also preferred over ARTalk* across all four criteria, while the advantage over UniLS is smaller but remains consistent, with all preference rates above 50%. Compared with ground truth, ReDyFlow receives fewer votes for synchronization, expression, and head pose, indicating a remaining perceptual gap. However, its listening reactions receive 57.69% of the valid preference votes, suggesting that the generated responses can remain perceptually plausible.

5 Conclusion

We presented **ReDyFlow**, a two-stage framework for joint conversational head generation that combines role-aware deterministic prediction with anchor-conditioned flow matching. Given paired audio, speaking activity, and reference motion styles, the model constructs dyadic motion anchors and jointly refines both participants through a temporal-pyramid flow, with rollout-based supervision regularizing sampled motion dynamics. Experiments on UniLS demonstrate improved reconstruction, distribution matching, and conversational motion quality across speaking and listening settings.

Limitations and future work. ReDyFlow currently operates on fixed-length segments and does not explicitly model long-term or multi-turn conversational context. Training supervision is based on tracked FLAME parameters rather than ground-truth 3D scans, and the current output contains motion parameters without appearance or texture. Style conditioning also relies on reference motion clips; future work could introduce a dedicated controllable style module. Finally, several dynamic metrics remain behind UniLS, suggesting that the continuous formulation may still produce overly smooth motion in some cases. Combining continuous generation with discrete motion representations, together with richer interaction cues and longer-term context, is a promising direction.

AI USE STATEMENT

The research ideas, methodology, experimental design, and interpretation of results are the authors' own. We used generative AI tools as assistants in supporting roles: (i) coding assistance for implementation, debugging, evaluation and visualization scripts, and dataset-format conversion; (ii) checking and refining the presentation of mathematical derivations developed by the authors; (iii) literature search and summarization, with all cited works verified by the authors; and (iv) language editing, Chinese–English translation, reference formatting, and preparation of schematic figures. All AI-assisted code and text were reviewed by the authors, and code was tested before use.

AI tools were not used to generate training or evaluation data, experimental results, or result figures. All quantitative results come from our executed evaluations or cited baseline reports, human-study judgments were provided by volunteer participants, and qualitative comparisons are frames rendered from the models' motion outputs. The authors take full responsibility for the content of this paper, including all AI-assisted material.

The authors take full responsibility for the final manuscript, including the correctness, originality, attribution, and integrity of all AI-assisted content and artifacts.

ETHICS STATEMENT

Our experiments use existing publicly available research data, including the UniLS-processed Seamless Interaction dataset Chu et al. (2026). We do not collect new audio, video, or facial-motion recordings for model training or benchmark evaluation. The user study is conducted online with 27 volunteers who view rendered comparison videos and provide A/B preference judgments. Participation is fully anonymous, and participants provide consent before taking part in the evaluation. We do not collect names, contact information, demographic information, or other personally identifying or sensitive data, nor do we record participants' appearance, voice, or movements. Only the final aggregate vote counts for each comparison criterion are retained and reported. The study protocol is described in Appendix D.

ReDyFlow is intended for research on conversational avatars. We therefore recommend clearly identifying synthesized content and obtaining authorization for identity-specific applications. Generated nonverbal behavior should not be treated as evidence of a real person's intentions or consent.

REPRODUCIBILITY STATEMENT

The model formulation and training objectives are described in Sec. 3, and the experimental setting and comparisons in Sec. 4. Appendix A documents preprocessing, architecture, optimization schedules, loss weights, checkpoint selection, random seeds, hardware, and end-to-end inference settings. Appendix A specifies the ablation implementations and distinguishes quantitative evaluation from rendering-stage smoothing. The human-study protocol and preference aggregation are detailed in Appendix D, with additional DualTalk comparisons in Appendix E. Supplementary videos provide continuous motion examples beyond the static qualitative figures.

REFERENCES

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.

Junjie Chen, Fei Wang, Zhihao Huang, Qing Zhou, Kun Li, Dan Guo, Linfeng Zhang, and Xun Yang. Towards seamless interaction: Causal turn-level modeling of interactive 3d conversational head dynamics. *arXiv preprint arXiv:2512.15340*, 2025.

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, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25142–25152, 2026.

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

Chenyang Ding, Shuai Tan, Qunfen Lin, Xinwei Jiang, Zijiao Zeng, and Ye Pan. SubtleTalk: Generating controllable weakly-correlated facial dynamics for 3d talking heads via residual flow matching. *arXiv preprint arXiv:2608.06408*, 2026.

Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model. In *European Conference on Computer Vision*, pp. 204–221. Springer, 2024.

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.

Zikai Huang, Siyue Chen, Xuemiao Xu, Haoxin Yang, Cheng Xu, Yihong Lin, and Shengfeng He. Learn2chat: Rethinking dyadic talking heads via interaction-modulated monologic priors. *arXiv preprint arXiv:2607.10313*, 2026.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, 2025.

Jisoo Kim, Jungbin Cho, Joonho Park, Soonmin Hwang, Da Eun Kim, Geon Kim, and Youngjae Yu. Deeptalk: Dynamic emotion embedding for probabilistic speech-driven 3d face animation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 4275–4283, 2025.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 2024. doi: 10.1007/s11263-024-02042-6.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

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*, pp. 10083–10093, 2023.

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

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20687–20697, 2023b.

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 Computer Vision and Pattern Recognition Conference*, pp. 21055–21064, 2025.

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.

Mengyi Shan, Shouchieh Chang, Ziqian Bai, Shichen Liu, Yinda Zhang, Luchuan Song, Rohit Pandey, Sean Fanello, and Zeng Huang. Talking together: Synthesizing co-located 3d conversations from audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2603.08674.

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

Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pp. 1–11, 2023.

Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yongjin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (ToG)*, 43(4):1–9, 2024.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20621–20631, 2023.

Xinmu Wang, Xiang Gao, Xiyun Song, Heather Yu, Zongfang Lin, Liang Peng, and Xianfeng Gu. Ot-talk: Animating 3d talking head with optimal transportation. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 1340–1349, 2025.

Xinmu Wang, Fumi Honda, Xiang Gao, Junqi Huang, Wei Chen, and Xianfeng David Gu. Grambatalk: generating 3d talking head with spatial graph and temporal mamba. *The Visual Computer*, 42(10):463, 2026.

Wei Wei, Yinyuan Zhao, and Ruixuan Yu. Residual flow matching with dynamic cross-interaction for 3d multi-person motion prediction. *arXiv preprint arXiv:2608.03379*, 2026.

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.

Karren D. Yang, Anurag Ranjan, Jen-Hao Rick Chang, Raviteja Vemulapalli, and Oncel Tuzel. Probabilistic speech-driven 3D facial motion synthesis: New benchmarks, methods, and applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27294–27303, 2024.

Zelin Zhao, Petr Molodyk, Haotian Xue, and Yongxin Chen. Laplacian multi-scale flow matching for generative modeling. *arXiv preprint arXiv:2602.19461*, 2026.

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.

Lei Zhu, Lijian Lin, Ye Zhu, Jiahao Wu, Xuehan Hou, Yu Li, Yunfei Liu, and Jie Chen. MANGO: Natural multi-speaker 3d talking head generation via 2d-lifted enhancement. *arXiv preprint arXiv:2601.01749*, 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.

A Implementation Details

This section specifies the eight-GPU NVIDIA V100 training configuration of ReDyFlow, including data preparation, network dimensions, optimization, and full-sequence inference. We additionally report the measured training costs and aggregate inference throughput.

A.1 Data Preparation and Conditioning

Motion representation and normalization. Motion is processed at 25 FPS with 16-kHz audio. Each training window contains 100 motion frames and 64,000 audio samples per participant. The compact 108-dimensional representation concatenates expression, head pose, jaw, and eye parameters with dimensions $(100, 3, 1, 4)$. From the original 112-dimensional motion code, the implementation retains slices `[:104]`, `[106:108]`, and `[109:111]` using zero-based indexing. Global head pose is mean-centered over the complete source sequence before windowing. Channel-wise normalization statistics are estimated from full paired training sequences, combining both participants and speaking/non-speaking states rather than fitting separate role-dependent normalizers. Each training pair is visited once for this calculation, avoiding repeated counting of overlapping training windows. Standard deviations are floored at 10^{-5}. The resulting statistics are fixed and reused in all Stage 1 and Stage 2 phases, including motion-only pretraining.

Window and reference sampling. Training windows are indexed with a stride of 50 frames. Their starts are randomly shifted by an integer offset between 0 and 50 and clamped to keep the window within the source sequence. The participant order is swapped with probability 0.5, consistently reordering audio, motion, activity masks, and reference clips. Each participant uses a 100-frame style reference sampled from its own source sequence. Training reference starts are sampled uniformly from the valid range. For validation and test references, the supplied data loader uses $s_{\text{ref}} = \min(\lfloor L/2 \rfloor, L - 100)$ for source length L; thus, the reference begins at the midpoint unless clamped by the end. The sampling protocol does not require the style and target intervals to be disjoint. Full-sequence inference keeps the chosen reference fixed across windows.

Activity and audio processing. Speaking activity is read from the dataset masks. Turn-duration and transition features are computed on full-source masks before extracting a window and follow the reordered participant keys. Consequently, window boundaries do not reset the turn context. Dataset-level audio normalization is disabled: the HuBERT wrapper standardizes each input waveform internally, while the five lightweight energy/spectral features are computed from the waveform before that standardization.

A.2 Network Configuration

Audio backbone and adapters. The audio backbone is initialized from `hubert-large-ll60k` and kept frozen in evaluation mode. Task-specific networks are initialized independently of earlier task checkpoints. Hidden states from layers $\{6, 12, 18, 24\}$ supply two learned layer mixtures per stage. Stage 1 initializes its articulation and response mixtures with a logit bias of 4 toward layers 6 and 24, respectively; Stage 2 initializes its mixtures uniformly. Stage 2 reuses the same backbone evaluations but learns separate audio adapters, style and role encoders, and anchor projections. The complete Stage 1 predictor remains frozen and in evaluation mode during Stage 2, including disabling its style dropout.

Deterministic predictor. Stage 1 smooths activity masks with a five-frame kernel and uses a bounded routing-logit correction with scale 0.5. Its scale blocks combine Transformer layers with kernel-3 temporal convolutions: dilations are $(1, 2)$ at 100 and 50 tokens, (1) at 25 tokens, and no temporal convolutions at 5 and 1 tokens. Coarse-motion interaction uses strengths 0.7 and 1.0 at the 5- and 25-token resolutions. Coarse predictions are detached when entering these interaction modules. Expression, pose, eye, and jaw decoding use scales $\{1, 5, 25, 50, 100\}$, $\{1, 5, 25, 50\}$, $\{5, 25, 50, 100\}$, and $\{25, 50, 100\}$, respectively. The style encoder combines attention-pooled temporal features with 324-dimensional reference statistics formed by the mean, standard deviation, and velocity standard deviation. Per-participant style dropout substitutes a learned null-style embedding rather than a zero-valued motion clip.

Table 3: Architecture configuration. Stage 1 depths list Transformer layers in ascending scale order; Stage 2 depths list flow blocks along the encoder–decoder.

Setting	Stage 1	Stage 2
Hidden dimension	768	768
Temporal-attention heads	8	8
Feed-forward dimension	1536	2048
Temporal token counts	$(1, 5, 25, 50, 100)$	$(100, 50, 25, 50, 100)$
Corresponding depths	$(0, 2, 2, 1, 1)$	$(1, 1, 2, 1, 1)$
Interaction resolutions	5 and 25	50, 25, and decoder 50
Interaction dimension / heads	128 / 4	768 / 8
Network dropout	0.05	0
Style embedding dimension	768	768
Training style dropout	0.10 per participant	None
Auxiliary adapter width	Prosody: 64; fusion: 128	Audio: 256
Output dimensions	$100 + 3 + 1 + 4$	$100 + 3 + 1 + 4$

Table 4: Executed training schedules and approximate costs on eight V100 GPUs. Parameter counts include only weights updated in the corresponding phase. Phase times are estimated from the recorded epoch counts and approximate per-epoch times.

Phase	Epochs	Batch/GPU	Trainable (M)	Time/epoch (min)	Estimated phase time
Stage 1: Canonical	10	32	190.08	35	5 h 50 min
Stage 1: Interaction	2	32	1.21	35	1 h 10 min
Stage 1: Joint	3	32	191.28	35	1 h 45 min
Stage 2: Prior	20	64	49.8	8	2 h 40 min
Stage 2: Audio-paired	60	32	65.4	30	30 h
Stage 2: Refine	20	32	65.4	30	10 h

Joint flow network. Stage 2 uses a single 100-frame ODE with an internal temporal encoder–decoder, as specified in Table 3. Temporal downsampling uses average pooling; upsampling uses linear interpolation and projected encoder–decoder skip concatenations. The style encoder maps the 324-dimensional reference statistics through a $324 \to 768 \to 768$ MLP. Cross-participant attention gates are initialized with bias -2. Component output heads use a 384-dimensional intermediate layer; condition-dependent affine modulation uses gain $1 + 0.1 \tanh(\gamma)$ and shift 0.1β. The final velocity projections are initialized with weight standard deviation 10^{-3} and zero bias. Source-noise standard deviations are $(1.0, 1.0, 0.35, 0.7)$ for expression, pose, jaw, and eyes in normalized space. There are no fixed participant-ID embeddings.

A.3 Distributed Optimization and Training Schedule

The implementation uses PyTorch distributed data parallelism with one process per V100 GPU and no gradient accumulation. Audio-backbone forward passes are internally split into microbatches of four waveforms; this does not change the optimizer batch size. Both stages use AdamW with $\beta = (0.9, 0.99)$, weight decay 0.01, and gradient-norm clipping at 1.0. Training uses FP16 automatic mixed precision and dynamic loss scaling initialized to 256, with activation checkpointing enabled and TF32 disabled. Loss statistics and the flow integration state are computed in FP32. Data loading uses two training workers and one validation worker per process, with spawned workers and process-local read-only LMDB access.

Executed training schedule and computational cost. Stage 1 is trained for 10, 2, and 3 epochs in its canonical, interaction, and joint phases, respectively. Stage 2 completes the configured 20 prior, 60 audio-paired, and 20 refine epochs without early stopping. Table 4 reports these executed schedules and their computational costs. Without gradient accumulation, 32 conversations per GPU give a global batch of 256 conversations, while 64 individual clips per GPU give a global prior batch of 512 clips. Each training clip contains 100 frames.

Trainable parameters and GPU memory. The phase-specific trainable counts exclude the frozen HuBERT backbone; Stage 2 counts additionally exclude the frozen Stage 1 predictor. They therefore do not describe the total number of parameters loaded for inference, and counts from successive phases must not be summed as distinct models. The reported peak training memory is approximately 9.68 GB per GPU for Stage 1, 3.9 GB per GPU for the Stage 2 non-audio prior, and 13.2 GB per GPU for each of the audio-paired and refine phases. The Stage 1 value is a stage-level peak, not a separate measurement for each of its phases. These are training-memory measurements rather than inference-memory requirements.

Summing the phase-time estimates gives approximately 8 h 45 min for Stage 1 and 42 h 40 min for Stage 2, or 51 h 25 min for both stages on the eight-GPU server. This total is derived from approximate per-epoch runtimes rather than an independently measured cumulative wall-clock duration.

Stage 1 optimization. Canonical training disables explicit coarse-motion interaction and trains the remaining task network at 10^{-4}. Interaction training updates only the interaction modules and feature-injection layers at 5×10^{-5}. Joint training updates the task network with rates 10^{-5} for canonical parameters and 2×10^{-5} for decoding, interaction, and newly introduced conditioning/layer-mixture parameters. The optimizer is reinitialized at each phase, starting from the best checkpoint of the preceding phase. Learning rates are updated once per epoch using cosine decay with phase-specific base rates and a lower learning-rate bound of 10^{-7}. There is no learning-rate warm-up in Stage 1. Separately, normalized dispersion and additional parameter-dynamics losses are multiplied in the canonical phase by $s(e) = \min(\max((e-2)/6, 0), 1)$, where e is the one-based epoch. This gives two epochs without these terms, followed by a six-epoch ramp. Interaction and joint training use their full weights immediately.

Stage 2 optimization. The same flow network is trained sequentially as prior, paired, and refine. Prior training uses `TalkMotionOnly`, a zero anchor, and explicit unavailable-condition embeddings. Dyadic interaction and the audio, style, role, turn, and anchor encoders are frozen, while availability embeddings and their normalization remain trainable. The prior learning rate is 8×10^{-5}. Paired training uses 2×10^{-5} for the pretrained flow core and 10^{-4} for the condition parameter group. Refinement uses 5×10^{-6} and 2×10^{-5}, respectively. The condition group includes the condition encoder, cross-participant attention and gates, and component-head modulation. The remaining flow parameters form the core. All flow parameters are trainable during paired training and refinement, while Stage 1 remains frozen. Each phase starts from its predecessor's best checkpoint with a fresh optimizer.

Stage 2 uses an epoch-wise warm-up/cosine schedule with `warmup_epochs=1` and a minimum learning-rate ratio of 0.05. In this implementation, a single warm-up epoch already uses the base rate; cosine decay begins at the base rate in epoch 2 and reaches 5% of that rate at the last epoch. Prior epochs are capped at $\min(2500, |\mathcal{L}_{\text{train}}|)$ distributed batches, whereas paired and refine phases use the full training loader. Thus, a prior epoch need not traverse every motion-only window.

A.4 Loss Configuration

Table 5 lists component weights and role multipliers. Here non-speaking includes both listening and mutual silence; it differs from the listening-only routing branch. Reconstruction uses Smooth-L_1 for expression and MSE for pose, jaw, and eyes. First- and second-order differences use Smooth-L_1, with $\beta = 0.05$, on valid same-role intervals.

Stage 1 auxiliary losses. Velocity and acceleration loss families each have overall weight 1; normalized dispersion has weight 0.05 and uses the reconstruction component weights. The overall multi-scale reconstruction weight is 0.20. For ascending scales $(1, 5, 25, 50, 100)$, the expression weights are $(0.05, 0.10, 0.25, 0.25, 0.35)$. The corresponding pose weights are $(0.20, 0.35, 0.30, 0.15, 0)$; jaw weights are $(0, 0, 0.10, 0.30, 0.60)$; and eye weights are $(0, 0.25, 0.35, 0.25, 0.15)$. These weights supervise pooled predictions and are distinct from learned feature-fusion weights.

Canonical coarse, branch coarse, refined coarse, and interaction-delta regularization weights are 0.20, 0.15, 0.30, and 0.005, respectively. Canonical training activates the first two, interaction training activates the latter two, and joint training uses all four with the first two multiplied by 0.25. Coarse scale weights are $0.4/0.6$ at 5/25 tokens. The speaking, listening, overlap, and silence

Table 5: Loss weights in component order. Temporal role multipliers apply to both first- and second-order difference losses. Rows describe their respective loss families, not a single product across all rows.

Weight	Expression	Pose	Jaw	Eyes
Stage 1 reconstruction	1.00	2.00	5.00	0.75
Reconstruction: speaking	1.25	1.00	2.00	1.00
Reconstruction: non-speaking	1.00	1.00	1.00	1.25
Velocity	0.20	0.40	0.40	0.40
Acceleration	0.03	0.05	0.06	0.05
Temporal: speaking	1.00	1.00	1.25	1.00
Temporal: non-speaking	1.00	1.00	0.50	1.25
Multi-scale reconstruction	1.00	2.00	5.00	0.50
Coarse reconstruction	1.00	2.00	3.00	1.00
Mid-term displacement	1.00	1.00	0.50	0.50
Stage 2 FM/endpoint	1.00	1.00	0.75	0.50
FM/endpoint: speaking	1.00	1.00	1.20	1.00
FM/endpoint: non-speaking	1.20	1.30	1.20	1.10

branch-loss weights are $(1.0, 0.65, 0.50, 0.25)$. Binocular common, differential, angular, velocity, and event terms have weights $(0.35, 0.08, 0.20, 0.25, 0.10)$, with a non-speaking multiplier of 1.35. Event weighting uses threshold 0.01 and boost 2.0. Additional raw pose/jaw standard-deviation terms use weights $1.0/0.2$, and displacement supervision at lags $\{4, 8, 16\}$ has weight 0.05. These additional parameter-dynamics terms use a non-speaking multiplier of 1.2.

Stage 2 objectives and rollouts. Flow matching uses MSE with the component/role weights in Table 5. The local endpoint Smooth-L_1 term has weight 0.02 and $\beta = 0.05$. During refinement, a four-step differentiable Euler rollout is performed on one conversation per GPU every 16 training batches, starting with batch index 0. Rollouts use fresh Gaussian source noise. Their raw-parameter dynamics loss has weight 0.10, with relative velocity-standard-deviation weight 0.10. Position variation is component-balanced, whereas velocity variation is averaged over all 108 channels. A role needs at least eight frames for its motion statistics and eight valid adjacent differences for its velocity contribution. Temporal differences are taken before masking, so separated role intervals are not concatenated. Both stages disable the optional expression-covariance objective. No training loss requires FLAME vertex decoding.

A.5 Validation and Checkpoint Selection

Training uses base random seed 42. Validation shards are disjoint and are not padded with repeated samples. The training loaders use fixed 100-frame validation windows, distinct from full-sequence benchmark evaluation. Stage 1 selects the best checkpoint in each phase using the total validation objective with the dynamic-loss scale fixed to 1, irrespective of its training ramp. Validation losses are computed per conversation and averaged over the dataset.

Stage 2 selects checkpoints exclusively by validation FM loss. Validation source noise and flow times are deterministically seeded from record indices with base seed 12345. Additionally, up to the first 64 validation records are sampled using seeds 12345 and 54321 and ten Euler steps. Their parameter-space statistics and the difference between the two samples are diagnostics, not checkpoint selection criteria or substitutes for benchmark metrics. The validation batch sizes per GPU are 8 conversations for Stage 1, 32 individual clips for the Stage 2 prior, and 4 conversations for Stage 2 paired/refine training. Checkpoints store the optimizer, loss scaler, random-number-generator states, normalization statistics or their model buffers, and the experiment configuration. Stage 2 also records a hash of the fixed Stage 1 checkpoint, which is checked when loading the generator.

Table 6: End-to-end inference throughput on the complete test set using eight NVIDIA V100 GPUs with batch size 1 per GPU. A batch contains one dyadic sequence and jointly outputs both participants. Timings include frozen pretrained HuBERT feature extraction and exclude rendering. Stage 2 uses ten Euler steps.

Method	Stage 1 (min)	Stage 2 (min)	Total (min)	Dyadic FPS
ReDyFlow (full)	6	9	15	1536.44

B Inference

B.1 Full-Sequence Inference

The inference API uses ten Euler steps for a single joint trajectory of both participants, not ten steps for each pyramid level. For style-conditioned inference, reference motion clips are selected strictly following the UniLS inference protocol, without additional manual selection or method-specific filtering. Full sequences are divided into non-overlapping 100-frame windows. The last window is padded and its output trimmed to the valid length. Full-source turn features and fixed style references are reused; windows are concatenated without previous-generated-motion conditioning, overlap blending, or per-window pose recentering. The provided default seed is 12345, with source noise seeded by seed $+ 1009\,\text{sequence_index} + \text{window_index}$. The API returns denormalized predictions and Stage 1 anchors, together with normalized residuals. It applies no guidance, residual-gain calibration, or temporal smoothing. Any separately enabled rendering postprocessing is outside this generation API.

B.2 Measured Inference Throughput

Measurement setup. We evaluate inference over the complete test set using eight NVIDIA V100 GPUs with a batch size of one dyadic sequence per GPU. Each batch contains one conversational pair and jointly generates the motion of both participants. The test set contains 3,457 dyadic sequences, each with 400 synchronized motion frames. Timing includes audio-feature extraction with the frozen pretrained HuBERT backbone and excludes FLAME rendering.

End-to-end timing. The two generation stages are executed sequentially. Stage 1 anchor prediction takes approximately 6 minutes over the complete test set, followed by approximately 9 minutes for Stage 2 refinement with ten Euler steps. The total end-to-end wall-clock time is therefore

$$t_{\text{total}} = t_{\text{Stage 1}} + t_{\text{Stage 2}} = (6 + 9) \times 60 = 900\ \text{s}. \tag{14}$$

All throughput values below are computed using this combined 15-minute runtime.

Aggregate throughput. Because each generated sample contains both participants, we use a *dyadic frame* to denote the pair of generated head motions at the same timestamp. The complete test set therefore contains

$$3457 \times 400 = 1{,}382{,}800$$

dyadic frames. The aggregate eight-GPU throughput of the complete generation pipeline is

$$\text{FPS}_{\text{dyadic}} = \frac{3457 \times 400}{900} \approx 1536.44. \tag{15}$$

Equivalently, if the two participants are counted separately, this corresponds to approximately 3072.89 individual head-motion frames per second.

The full system therefore processes approximately 1,536 synchronized dyadic frames per second across eight GPUs. Since every dyadic frame contains the generated motions of both participants, this is equivalent to approximately 3,073 individual head-motion frames per second. These measurements report aggregate test-set throughput rather than single-GPU performance, single-sequence latency, or causal streaming speed.

C Evaluation and Rendering Protocol

Evaluation. We follow the UniLS evaluation protocol Chu et al. (2026), assessing both participants on valid speaking and listening frames. Metrics use motion parameters and the corresponding FLAME geometry, not rendered RGB images. Listening F-FID and P-FID aggregate the 100-dimensional expression and 8-dimensional pose parameters, respectively, over valid listening frames across the test set. Baseline quantitative results are taken from UniLS on the same test split. Our metric evaluation uses motion predictions before rendering-stage temporal smoothing; this postprocessing is not applied when computing quantitative results.

Rendering and smoothing. We use the UniLS default rendering script and settings, displaying each participant at 512×512 pixels and 25 FPS with a neutral FLAME identity and fixed camera, lighting, and appearance. Default Savitzky–Golay smoothing is enabled for predicted motion, using a nine-frame window and a third-order polynomial. The same rendering and smoothing settings are applied to ReDyFlow and all baselines. This smoothing is confined to video preparation: user-study preferences are assessed on the rendered, smoothed videos, whereas quantitative metrics are computed without this rendering-stage postprocessing. Ground truth is rendered from tracked FLAME parameters rather than shown as original RGB footage.

Temporal alignment. Word-level comparisons use matched timestamps across methods, with listening examples aligned to the interlocutor's spoken word. The longer timeline samples both participants every two seconds and preserves their original left–right ordering. A fixed spatial crop is shared across methods and timestamps without per-frame recentering or pose alignment. Supplementary videos provide continuous motion beyond these sampled visualizations.

D User Study

Study design. Following the A/B preference protocol of UniLS Chu et al. (2026), we recruited 27 volunteers to compare ReDyFlow with DualTalk, ARTalk*, UniLS, and ground truth. The same 12 source conversations are used for all four comparisons, yielding 48 paired-video trials per volunteer. Each trial compares two versions of the same conversation, one from ReDyFlow and one from the comparison target. Both versions show the two interlocutors; A and B identify the compared versions rather than individual interlocutors.

Stimuli and presentation. Compared videos use matched temporal intervals and audio, with the same neutral FLAME identity, camera, and lighting. Ground truth is rendered from tracked FLAME parameters rather than shown as original RGB footage. The two versions are presented side by side with anonymous A/B labels and randomized method-to-A/B assignment. The assignment is retained only for response aggregation, while method names are hidden from participants.

Evaluation criteria. For each paired-video trial, participants evaluate four perceptual criteria. They may prefer A, prefer B, or indicate that the two versions are equally good. Different versions may therefore be preferred for different criteria. The questions are:

1. **Lip synchronization (Sync):** Which video better synchronizes the speaking participants' lip movements with the audio?
2. **Expression naturalness (Exp):** Which video exhibits more natural facial expressions throughout the conversation?
3. **Reaction naturalness (React):** Which video exhibits more natural listening reactions to the interlocutor's speech?
4. **Head-pose naturalness (Pose):** Which video exhibits more natural head movements and orientations throughout the conversation?

Each participant views $12 \times 4 = 48$ paired videos and provides four criterion-level judgments for each pair.

Preference aggregation. Responses are aggregated separately for each comparison target m and criterion k. After decoding the randomized A/B assignment, we discard responses in which the participant rates A and B as equally good. For each remaining valid response, let $y_{u,c,m,k} = 1$ when participant u prefers ReDyFlow for conversation c, and 0 when the comparison method is preferred. Denoting the set of valid non-tie responses by $\mathcal{V}_{m,k}$, we compute

$$\mathrm{Pref}_{m,k} = \frac{100}{|\mathcal{V}_{m,k}|} \sum_{(u,c)\in\mathcal{V}_{m,k}} y_{u,c,m,k}. \tag{16}$$

Thus, ties do not contribute to either the numerator or denominator. With 27 participants and 12 conversations, each comparison–criterion has at most $27 \times 12 = 324$ judgments before tie removal, while the effective denominator depends on the number of valid non-tie responses. For every comparison, including ground truth, the reported percentage denotes the fraction of valid votes favoring ReDyFlow.

E Additional Experiments on DualTalk

Experimental setting. We additionally evaluate our reference-conditioned DualTalk adaptation on the Test and out-of-distribution (OOD) splits, covering 1,066 and 754 individual sequences, respectively. We report final predictions over complete sequences using FD, P-FD, MSE, SID, and rPCC, following DualTalk's metric definitions and display scales Peng et al. (2025). Table 7 compares our results with DualTalk, UniLS, and Learn2Chat. DualTalk results are taken from its original paper; UniLS and Learn2Chat results are taken from the evaluation reported by Learn2Chat Huang et al. (2026).

Table 7: Additional comparison on DualTalk. Metric definitions and display scales follow DualTalk Peng et al. (2025); lower is better except for SID. E, J, and P denote expression, jaw, and head pose. L2C(D) and L2C(U) identify Learn2Chat using DEEPTalk and UniTalker backbones, respectively Huang et al. (2026).

	FD↓			P-FD↓			MSE↓			SID↑			rPCC↓		
Method	E	J	P	E	J	P	E	J	P	E	J	P	E	J	P
Test															
DualTalk	11.14	1.90	3.83	11.88	1.97	3.97	3.59	1.04	1.72	3.48	2.23	1.72	4.73	1.37	2.38
UniLS	25.48	3.14	7.96	26.45	3.20	8.20	6.22	1.48	3.38	2.79	2.61	1.08	7.33	2.26	4.41
L2C (D)	14.95	1.83	5.78	15.96	1.93	5.97	4.60	1.16	2.19	3.12	2.25	1.38	6.59	1.34	2.58
L2C (U)	15.58	1.84	5.46	16.58	1.95	5.62	4.62	1.16	2.18	3.10	2.25	1.49	6.99	1.38	2.49
Ours	8.33	1.29	2.99	8.78	1.33	3.07	2.36	0.67	1.15	3.72	2.59	1.90	3.93	1.03	1.96
OOD															
DualTalk	21.71	3.15	5.89	22.56	3.22	6.06	5.97	1.50	2.48	2.98	1.94	1.38	6.86	1.60	3.28
UniLS	26.18	2.92	7.79	27.11	2.98	8.02	6.40	1.42	3.38	2.84	2.62	1.05	7.68	1.92	5.03
L2C (D)	22.34	2.51	6.00	23.41	2.61	6.20	6.24	1.42	2.42	2.79	2.09	1.36	7.72	1.46	2.31
L2C (U)	22.55	2.50	5.62	23.60	2.59	5.79	6.16	1.40	2.22	2.79	2.07	1.48	7.88	1.46	2.22
Ours	9.23	1.55	3.30	9.68	1.59	3.37	2.60	0.77	1.24	3.67	2.53	1.78	4.42	1.11	1.92

Results. Our adaptation obtains numerically lower FD, P-FD, MSE, and rPCC than the reported baselines for all three components on both splits. Compared with DualTalk, expression FD/MSE decrease from 11.14/3.59 to 8.33/2.36 on Test and from 21.71/5.97 to 9.23/2.60 on OOD. Expression and pose SID are also higher, while UniLS retains the highest jaw SID: 2.61/2.62 on Test/OOD versus our 2.59/2.53. These results provide additional evidence of favorable reconstruction and distributional matching, without a uniform advantage across all reported metrics.

F Additional Qualitative Results

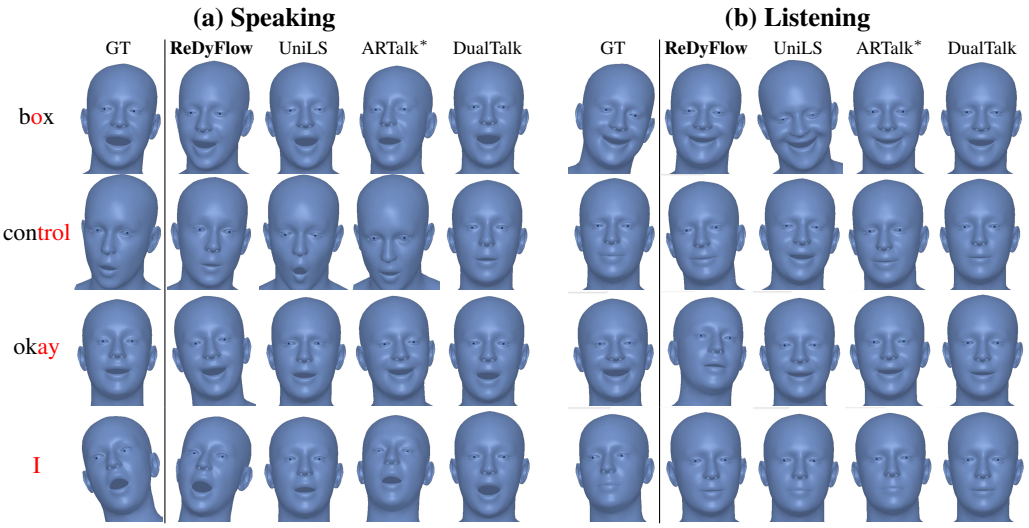

Figure 4: **Additional word-level comparisons.** Speaking (left) and listening (right) at matched timestamps with identical rendering. Listening examples show concurrent responses to the interlocutor's indicated word.

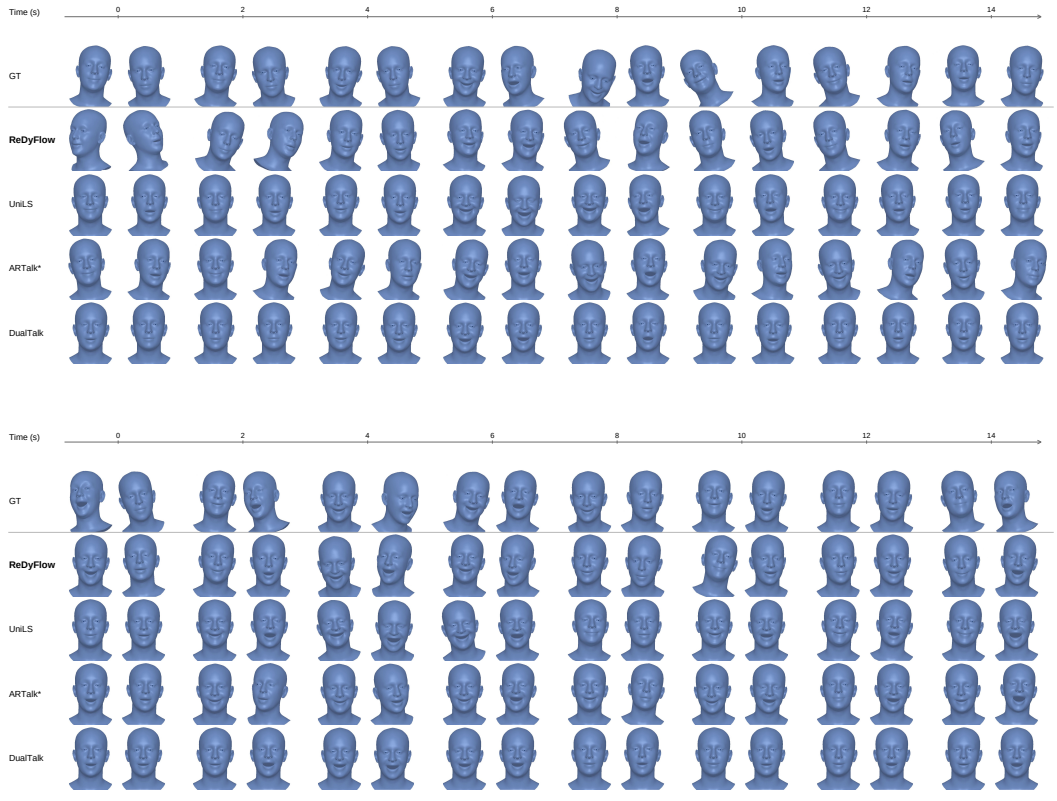

Figure 5: **Temporal comparison.** Methods are shown by row and timestamps by column, sampled every two seconds. Each cell retains both participants' original ordering, with a fixed crop across methods.