ERTALK: AUTOREGRESSIVE SPEECH-DRIVEN 3D FACE ANIMATION GUIDED BY EXPRESSIVE REFERENCE MOTION

Anonymous authors
Paper under double-blind review

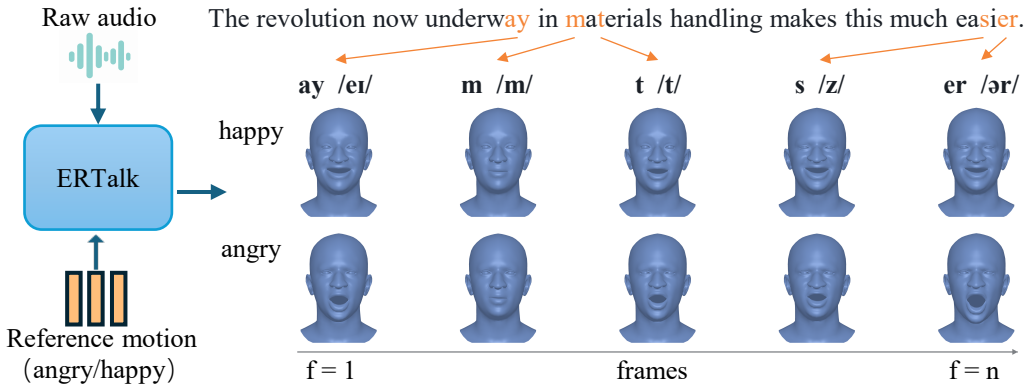

Figure 1: Overview of ERTalk. By selecting reference motion, users can control the emotional expression of speech-driven digital humans, as illustrated by happy and angry facial animations generated from the same audio.

ABSTRACT

Speech-driven 3D facial animation aims to generate natural and expressive facial motion with accurate lip synchronization. Discrete emotion labels alone provide limited information about fine-grained facial deformations. We introduce ERTalk, a reference-conditioned autoregressive model that uses speech and reference motion to generate synchronized lip movements and reference-guided expressions in a learned deformation space. Multi-State Reference Conditioning (MRC) augments the global reference prefix with layer-wise access to frame-level expression states, while differentiable decoding enables vertex and velocity supervision over all decoded vertices in valid frames. Our pretrained SMPL-X-based representation encodes facial expression separately from the target identity geometry. This representation allows the generated motion to be transferred across identities without rerunning the motion generator. Extensive evaluations demonstrate that our method outperforms existing approaches in both lip-sync accuracy and emotion control.

1 INTRODUCTION

3D facial animation supports realistic digital humans in film production, games, and virtual interaction. Speech-driven methods learn speech–motion relationships to automate animation traditionally requiring motion capture and manual editing (Karras et al., 2017; Cudeiro et al., 2019). Accurate lip synchronization is essential, but spoken content alone does not determine how a face should express emotion. A digital human should articulate the target speech naturally while conveying the expression intended by its user. These requirements call for facial-motion generation that combines speech synchronization with expressive control.

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.

Existing approaches offer different sources of emotional information. ProbTalk3D uses emotion and intensity labels (Wu et al., 2024), LSF-Animation extracts implicit emotion cues from speech (Lu et al., 2025), and ARTalk derives speaking style from reference motion (Chu et al., 2025). A discrete emotion label identifies an overall category, however, rather than a particular facial realization. Two happy expressions may differ in eyebrow elevation, cheek deformation, lip-corner motion, and intensity. Speech-derived emotion likewise need not match the expression a user wishes to create. A selected reference motion provides a concrete example of the desired expression, enabling finer control without requiring the user to specify a predefined emotion category.

Balancing accurate lip synchronization with fine-grained emotion control remains challenging. Reference motion contains both expression and articulation associated with its own speech, so directly transferring it can conflict with the pronunciation required by the target audio. Moreover, a compact global descriptor summarizes the reference but does not explicitly expose its individual expression states during generation. The central questions are how to make these detailed states available to the generator and how to incorporate them while preserving target-speech articulation, without requiring frame-wise correspondence between the reference and target.

We propose **ERTalk**, a reference-guided autoregressive framework that predicts discrete motion tokens from speech, reference motion, and previously generated tokens. Its central component, **Multi-State Reference Conditioning (MRC)**, supplements the retained global reference prefix with contextualized frame-level reference memory. Gated cross-attention in each Transformer layer lets the current speech–motion state attend to relevant reference features. The global prefix supplies an overall reference condition, while the memory provides access to finer expression states throughout generation. This combination supports reference-guided expression without prescribing a temporal alignment to the target utterance.

Token classification alone does not account for the geometric differences between predicted motion codes. Through differentiable motion decoding, vertex and velocity losses constrain all decoded vertices in valid frames, encouraging accurate facial deformation and coherent motion while the model learns reference-conditioned generation. Using SMPL-X (Pavlakos et al., 2019), we pretrain an implicit space of continuous expression vectors. Our deformation decoder combines these vectors with separately supplied target neutral geometry to produce mesh-wide displacements, allowing the same generated trajectory to drive different target identities.

Our main contributions are as follows:

1. **Reference-guided autoregressive deformation generation.** We formulate speech-driven facial animation as motion-token prediction conditioned on target speech and selected reference motion. The implicit expression representation separates the motion trajectory from the target geometry supplied at decoding. The decoded expression trajectory can be reused with different target geometries without rerunning the generator.
2. **Multi-State Reference Conditioning.** MRC augments the global reference prefix with frame-level memory and layer-wise gated cross-attention. It exposes fine-grained reference states to the evolving generation context without frame-wise reference–target alignment.
3. **Geometry-aware motion supervision.** We connect discrete token prediction to continuous facial motion through differentiable decoding. Vertex and velocity losses over all valid vertices and frames provide geometric and temporal constraints beyond token classification.

2 Related Work

We review speech-driven animation, generative motion models, and emotion and style conditioning, focusing on explicit 3D motion for animatable avatars.

2.1 Speech-Driven 3D Facial Animation

Early systems such as JALI derive animation controls from phonetic information (Edwards et al., 2016). Data-driven methods instead learn articulation and speaking variation from synchronized speech and motion. Karras et al. jointly model speech, facial motion, and emotion (Karras et al., 2017), while VOCA transfers speaker-conditioned motion across geometries (Cudeiro et al., 2019).

Recent methods improve speech representations and supervision through self-supervised speech features, audiovisual consistency, lip-synchronization teacher models, multilingual training, and heterogeneous annotations (Haque & Yumak, 2023; Peng et al., 2023a; Zhuang et al., 2025; Sung-Bin et al., 2024; Fan et al., 2024); ScanTalk supports unregistered scans with different topologies (Nocentini et al., 2024). FLAME and SMPL-X parameterize shape, expression, and pose (Li et al., 2017; Pavlakos et al., 2019). DECA and EMOCA reconstruct animatable geometry from images, with EMOCA emphasizing emotion-sensitive reconstruction (Feng et al., 2021; Daněček et al., 2022). These representations connect reconstruction to generation by specifying the facial surface on which motion is expressed.

2.2 Autoregressive and Diffusion-Based 3D Facial Motion Generation

Autoregressive models predict motion from speech and previously generated states. FaceFormer uses causal attention to regress continuous vertex motion (Fan et al., 2022). MeshTalk separates audio-correlated and audio-uncorrelated factors in a categorical latent space (Richard et al., 2021). CodeTalker uses vector quantization (van den Oord et al., 2017) to learn a discrete facial-motion prior that alleviates oversmoothing (Xing et al., 2023), while ARTalk predicts multi-scale motion codes conditioned on speech, motion history, and reference style (Chu et al., 2025). Discrete prediction thus constrains the motion space while retaining temporal dependencies.

Diffusion models generate motion by progressively denoising a noisy representation, capturing the one-to-many speech–motion relationship. FaceDiffuser and DiffSpeaker operate on vertices or blendshapes (Stan et al., 2023; Ma et al., 2024); DiffPoseTalk jointly generates expression and head pose (Sun et al., 2024); and FaceTalk uses a neural head-model latent space (Aneja et al., 2024; Giebenhain et al., 2023). Where continuous regression predicts geometry directly, discrete priors constrain predictions to learned motion codes; diffusion instead refines noisy motion through successive denoising steps.

2.3 Emotion- and Style-Conditioned 3D Facial Animation

EmoTalk disentangles emotion and linguistic content (Peng et al., 2023b), whereas EMOTE combines lip-reading and emotion supervision with content–emotion swapping (Daněček et al., 2023). ProbTalk3D uses emotion and intensity labels (Wu et al., 2024); DEEPTalk learns a shared probabilistic speech–motion emotion space (Kim et al., 2025b); and LSF-Animation derives emotion from speech and identity from neutral geometry (Lu et al., 2025). ExpTalk aligns expressions with speech emotion through adaptive disentanglement and diffusion refinement (Qu et al., 2025). DEITalk combines emotion-label control with speech-derived intensity (Shen et al., 2024), while MEDTalk accepts emotion labels, text, or reference expression images (Liu et al., 2025).

DiffPoseTalk and ARTalk encode reference style (Sun et al., 2024; Chu et al., 2025); StyleSpeaker combines reference-motion and audio-derived features (Yang et al., 2025). Chu et al. retrieve personalized examples using speech and motion history (Chu et al., 2026). Unlike methods using input reference clips, MemoryTalker modulates learned shared motion memory with audio-derived style, requiring only audio at inference (Kim et al., 2025a).

3 Method

Using a pretrained deformation space, ERTalk learns a discrete motion prior and then trains a speech-conditioned autoregressive generator with reference memory and geometric supervision (Figure 2).

3.1 Motion Representation and Discrete Prior

Given speech A and a fixed-length reference $R = (r_1, \ldots, r_l)$, we generate $D = (d_1, \ldots, d_T)$ with $r_j, d_t \in \mathbb{R}^{1024}$. Building on SMPL-X (Pavlakos et al., 2019), we pretrain an implicit expression space and its deformation decoder, representing facial expression separately from target identity geometry. Our frozen deformation decoder $\mathcal{G}_\phi$ combines these coefficients with the target neutral

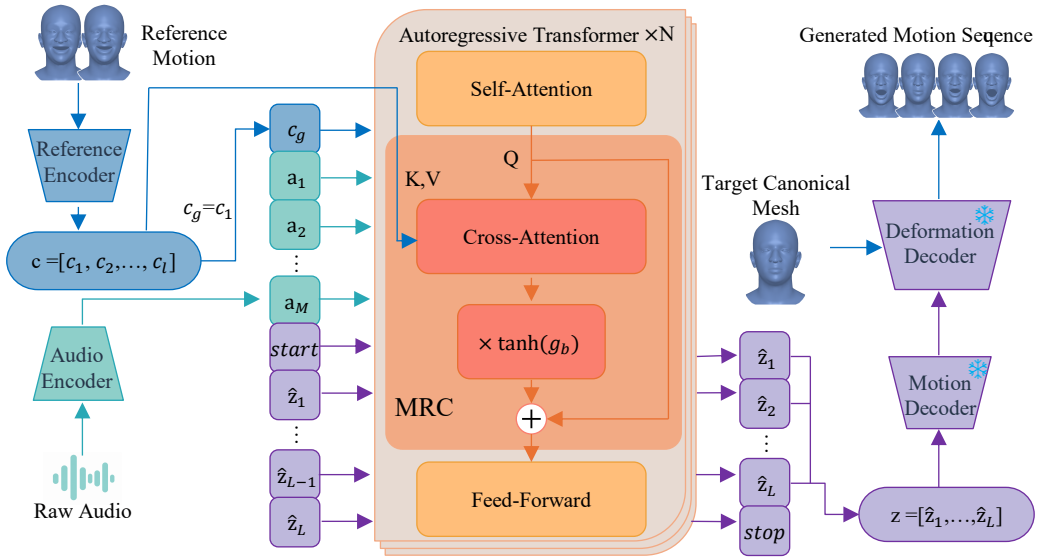

Figure 2: ERTalk pipeline. The audio encoder extracts speech features, and the reference encoder provides a global prefix and frame-wise memory. Conditioned on the prefix and speech features, the autoregressive Transformer predicts motion tokens from START and the generated history, with MRC injecting reference information through gated cross-attention at each block. STOP ends autoregressive sampling. The frozen motion decoder converts the generated motion tokens into deformation coefficients, which the frozen deformation decoder combines with the target canonical mesh to produce the facial motion sequence.

mesh V^{β}_{can} to produce per-vertex displacements:

$$O^{\beta}_{t} = \mathcal{G}_{\phi}(V^{\beta}_{\text{can}}, d_t), \qquad V^{\beta}_{t} = V^{\beta}_{\text{can}} + O^{\beta}_{t}. \tag{1}$$

Here, β denotes identity shape; the same coefficient trajectory can drive different target meshes without rerunning the generator.

A discrete motion autoencoder (dVAE) quantizes downsampled features (van den Oord et al., 2017) using a codebook $\mathcal{E} = \{e_i\}_{i=0}^{Q-1}$ with $Q = 8192$ and $e_i \in \mathbb{R}^{256}$. Twofold temporal compression yields $\mathbf{z} = (z_1, \ldots, z_L)$, where $L = T/2$ for even T. Pretraining combines coefficient and decoded-vertex reconstruction, vertex-displacement, and commitment losses. The motion decoder D_{vq} reconstructs coefficients for $\mathcal{G}_{\phi}$; both autoregressive training stages share the same frozen dVAE and codebook.

3.2 Reference-Conditioned Autoregressive Generation

Speech features. Frozen HuBERT-Large (Hsu et al., 2021) encodes 16 kHz audio. Its features are resampled to 25 Hz and temporally downsampled to $A' = (a_1, \ldots, a_M)$, with $M = T/2$ for even-length windows, matching the motion-token rate. Audio and motion tokens use separate learned positional embeddings.

Multi-State Reference Conditioning. A 1×1 convolution and a six-layer, four-head encoder produce reference states of dimension $d = 1024$, scaled by a condition gain of 1.5:

$$\mathbf{c} = 1.5E_r(R) = (c_1, \ldots, c_l) \in \mathbb{R}^{l \times d}, \qquad c_g = c_1. \tag{2}$$

The first contextualized state provides the global prefix; all states form memory without temporal positional encoding, accessed by content. A causal Transformer (Vaswani et al., 2017) receives $[c_g, A']$, START, and motion history, with access to the full speech prefix but no future motion tokens. Its $N = 24$ blocks use 16-head self- and cross-attention and feed-forward dimension 4096. After

self-attention, each block's states $X^{(b)}$ query memory through gated cross-attention (Alayrac et al., 2022):

$$\widetilde{X}^{(b)} = X^{(b)} + \tanh(g_b)\ \mathrm{MHA}\Big(\mathrm{LN}(X^{(b)}), \mathbf{c}, \mathbf{c}\Big). \tag{3}$$

MHA denotes multi-head attention and LN layer normalization. The zero-initialized scalar gate g_b scales the residual before the feed-forward network, allowing context-dependent access to the fixed reference memory.

Next-token generation. START initiates generation and STOP terminates it. With $z_0 = \texttt{START}$ and $z_{L+1} = \texttt{STOP}$,

$$p_\theta(\mathbf{z}, \texttt{STOP} \mid A, R) = \prod_{n=1}^{L+1} p_\theta(z_n \mid z_{0:n-1}, A', c_g, \mathbf{c}). \tag{4}$$

Each sampled motion token becomes the next motion input. We use teacher forcing with shifted targets $[z_1, \ldots, z_L, \texttt{STOP}]$ and cross-entropy $\mathcal{L}_{\text{autoregressive}}$; prefix positions receive no motion-token loss.

3.3 Geometry-Aware Motion Supervision

Gumbel-Softmax (Jang et al., 2017) maps motion-code logits ℓ_n^{mot} to assignments $\pi_n = \mathrm{GumbelSoftmax}(\ell_n^{\text{mot}}; \tau_g)$ at temperature τ_g, enabling differentiable decoding:

$$\widetilde{e}_n = \sum_{i=0}^{Q-1} \pi_{n,i} e_i, \qquad \widetilde{D} = D_{\text{vq}}(\widetilde{e}_1, \ldots, \widetilde{e}_L). \tag{5}$$

Predicted and target coefficients are decoded with the same neutral geometry, excluding START and STOP. Let Ω_V contain all decoder vertices in valid, nonpadding frames, and $\Omega_\Delta = \{(t,v) : (t,v), (t+1,v) \in \Omega_V\}$ contain valid adjacent-frame pairs. With $\Delta V_{t,v} = V_{t+1,v} - V_{t,v}$ at 25 Hz and $\|u\|_\epsilon = \sqrt{\sum_{k=1}^{3} u_k^2 + 10^{-8}}$, the geometric losses are

$$\mathcal{L}_{\text{vertex}} = \frac{1}{|\Omega_V|} \sum_{(t,v)\in\Omega_V} \|\widetilde{V}_{t,v} - V_{t,v}\|_\epsilon, \tag{6}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{|\Omega_\Delta|} \sum_{(t,v)\in\Omega_\Delta} \|\Delta\widetilde{V}_{t,v} - \Delta V_{t,v}\|_\epsilon. \tag{7}$$

Together with next-token cross-entropy, the objective is

$$\mathcal{L} = \mathcal{L}_{\text{autoregressive}} + 1.0\mathcal{L}_{\text{vertex}} + 1.5\mathcal{L}_{\text{vel}}. \tag{8}$$

Gradients pass through both frozen decoders to the generator and reference pathway.

3.4 Training and Inference

During training, ground-truth motion tokens supervise next-token prediction through cross-entropy, while vertex and velocity losses constrain the decoded motion (Eq. 8). We optimize all autoregressive model parameters, including the reference pathway, with the dVAE and deformation decoder frozen. At inference, the model generates motion tokens sequentially from speech, reference motion, and the generated history, using classifier-free reference guidance (Ho & Salimans, 2022) and top-p sampling (Holtzman et al., 2020). STOP terminates generation; the motion-token sequence is adjusted to the target length $L = T/2$ and decoded into T frames of facial motion.

4 Experiments

Datasets. Autoregressive pretraining (0–180k) uses 34,869 non-public video clips: 18,499,579 frames at 25 FPS (205.551 hours). We adapt and evaluate on MEAD (Wang et al., 2020) at 25 FPS, with 15,016/3,491/3,269 training/validation/test clips from 31/7/7 identities, respectively, and no identity or clip overlap. ERTalk and its ablations are adapted only on the training partition. Quantitative comparisons and ablations use a fixed 1,620-clip test subset covering seven identities, eight emotions, and three intensity levels; paired qualitative comparisons and the user study also use held-out test data.

Implementation Details. We freeze the 62k-step dVAE/codebook and 230k-step deformation decoder. Autoregressive pretraining uses one NVIDIA A100 with effective batch size 100; continuation (180k–412k) uses one NVIDIA RTX 4090. Both use AdamW with base learning rates 10^{-4} and 5×10^{-5}, respectively. Full ERTalk and its ablations use continuation training/validation batch sizes of 32/16, without gradient accumulation, and a 5,000-update learning-rate warm-up followed by cosine decay. Even-length training windows cover 20–400/20–80 frames in the two stages, within the common valid audio–motion interval. The two stages take approximately 190.3 and 51.7 hours.

Pretraining derives a global prefix from 50 reference frames; continuation uses 20 frames for both the prefix and MRC memory. Training preferentially samples nonoverlapping same-video references; validation also uses same-video references. Quantitative evaluation and ablations use $l = 20$ consecutive reference frames following the target window in the same clip, without overlap.

All reported ERTalk experiments use 412k-step checkpoints with the global reference prefix enabled. Quantitative sampling uses temperature 0.8, top-p 0.8, and repetition penalty 2.0.

4.1 Quantitative Results

Baselines and evaluation. ARTalk (Chu et al., 2025) is retrained on our MEAD training partition; DEEPTalk (Kim et al., 2025b), EMOTE (Daněček et al., 2023), LSF-Animation (Lu et al., 2025), and ProbTalk3D (Wu et al., 2024) use official MEAD-trained checkpoints, native conditioning interfaces, and unmodified predictions. All methods use the designated test subset.

We compare predicted and reconstructed ground-truth (GT) geometry using official FLAME–SMPL-X vertex correspondences (Li et al., 2017; Pavlakos et al., 2019), with one sequence per clip and fixed clip-level seeds. Full and ablated ERTalk share target windows, GT, semantic regions, and seeds.

We report four metrics. Mean Vertex Error (MVE) measures overall accuracy by averaging Euclidean errors over corresponding full-face vertices and valid frames. Lip Vertex Error (LVE) (Xing et al., 2023) and Emotional Vertex Error (EVE) (Peng et al., 2023b) measure lip and upper-face accuracy, respectively, by averaging each region's per-frame maximum Euclidean error over valid frames. Average Facial Dynamics Deviation (AFDD) measures motion variation using an absolute-deviation variant of upper-face dynamics deviation (FDD) (Xing et al., 2023): it averages absolute differences between predicted and GT coordinate-wise temporal standard deviations over upper-face vertices and coordinate axes. Lower scores are better.

Table 1: Comparison on 1,620 MEAD test clips: errors (lower is better), with best results in bold.

Method	**MVE** ↓	**LVE** ↓	**EVE** ↓	**AFDD** ↓
ARTalk	0.85534	0.78905	1.08064	0.40967
DEEPTalk	1.07834	0.90894	1.21668	0.86340
EMOTE	0.99142	0.93406	1.08670	0.80885
LSF-Animation	1.14102	1.40368	0.99186	0.52170
ProbTalk3D	1.13831	1.25847	0.87408	0.48849
ERTalk	**0.52846**	**0.53777**	**0.47619**	**0.38181**

ERTalk achieves the lowest errors across the four reported metrics (Table 1). Figure 3(a) shows lower ERTalk errors around the mouth and upper face; several baselines show errors concentrated in the lower face.

4.2 Qualitative Results

Figure 4 compares ERTalk with five baselines and reconstructed ground truth on held-out test examples of anger, contempt, disgust, happiness, and surprise. Under the same rendering settings and at corresponding time points, ERTalk agrees closely with the target in mouth opening, lip contours, and local expression cues.

The examples show local expression variations that emotion-category labels alone do not specify: the contempt example exhibits mouth asymmetry and uneven lip corners; anger and surprise exhibit large mouth openings with distinct brow configurations; happiness shows raised lip corners; and disgust

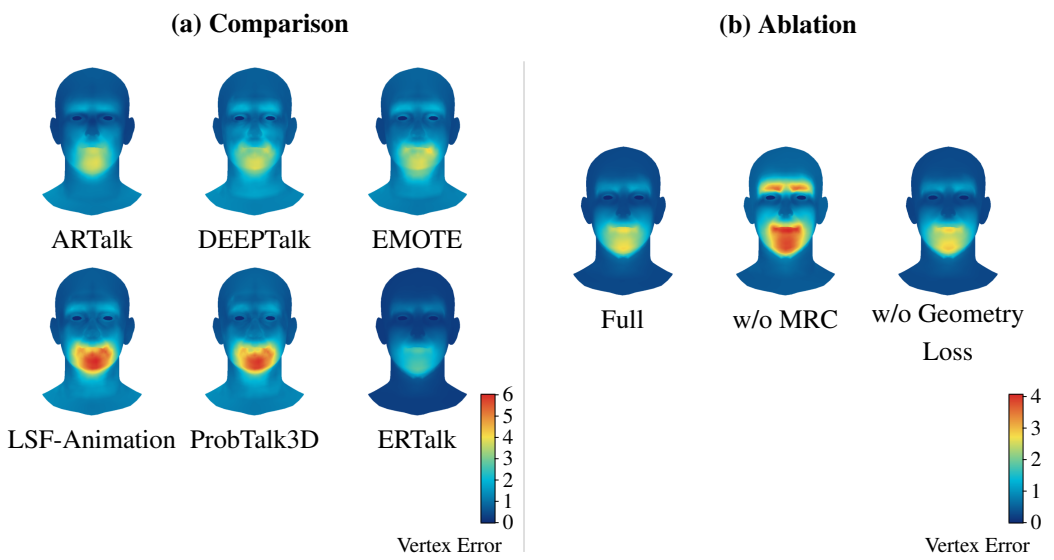

Figure 3: Vertex error maps for (a) comparisons and (b) ablations, from the same evaluation data as Tables 1 and 3, respectively. Methods share a color scale within each panel; ranges differ between panels. Lower is better.

shows a compact mouth shape. Several baselines deviate in mouth opening or attenuate these local cues.

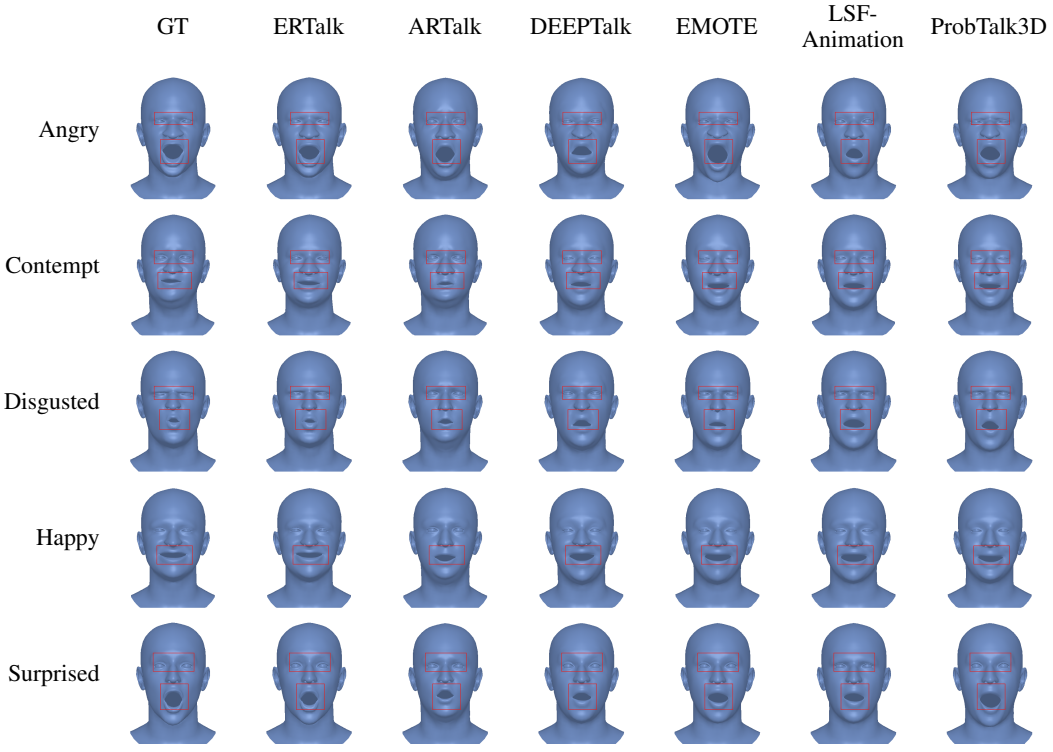

Figure 4: Reconstructed ground truth and six methods' outputs at matched time points across five emotions in held-out test examples. Red boxes highlight key facial expressions.

Changing the reference with fixed audio. Figure 5 uses a randomly selected target utterance absent from the training, validation, and test sets, with happy and angry reference motions drawn from the test set. With target audio fixed, changing the reference alters lip-corner and brow configurations while retaining speech-related mouth-opening changes at five corresponding time points.

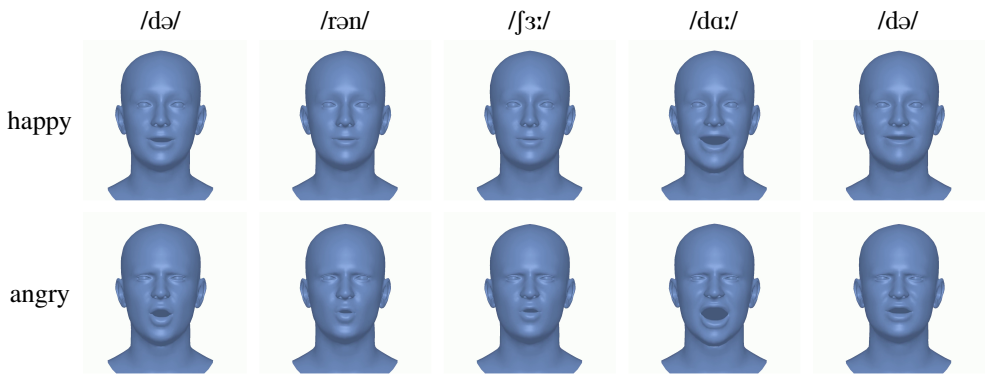

Figure 5: Generated facial motion for fixed target audio, conditioned on happy (top) and angry (bottom) reference motion. Columns show corresponding time points in the target utterance.

4.3 User Study

The same 15 participants completed ten method-blinded, randomized A/B trials, comparing ERTalk with each baseline twice across anger, contempt, happiness, and surprise. Animations shared audio and rendering settings and were synchronized with the target RGB video. Participants chose one method without ties for four criteria: lip synchronization and facial-motion accuracy, emotion accuracy, temporal stability, and overall naturalness. This yields 40 choices per participant (600 total) and 30 votes per baseline per criterion. Table 2 reports preference proportions with nominal 95% vote-level Wilson intervals.

Table 2: Pairwise preference for ERTalk (%) with 95% vote-level Wilson intervals.

Baseline	**Lip/Face Acc.**	**Emotion Acc.**	**Stability**	**Naturalness**
ARTalk	90.0 [74.4, 96.5]	96.7 [83.3, 99.4]	93.3 [78.7, 98.2]	86.7 [70.3, 94.7]
DEEPTalk	86.7 [70.3, 94.7]	90.0 [74.4, 96.5]	96.7 [83.3, 99.4]	93.3 [78.7, 98.2]
EMOTE	66.7 [48.8, 80.8]	83.3 [66.4, 92.7]	96.7 [83.3, 99.4]	93.3 [78.7, 98.2]
LSF-Animation	56.7 [39.2, 72.6]	56.7 [39.2, 72.6]	73.3 [55.6, 85.8]	73.3 [55.6, 85.8]
ProbTalk3D	70.0 [52.1, 83.3]	66.7 [48.8, 80.8]	70.0 [52.1, 83.3]	80.0 [62.7, 90.5]

ERTalk receives majority preference across all baselines and criteria, with larger margins against ARTalk and DEEPTalk than against LSF-Animation.

4.4 Ablation Study

The *w/o MRC* variant removes only layer-wise reference cross-attention, retaining the global prefix, speech, motion history, and geometric supervision. *w/o Geometry Loss* retains MRC and full-model sampling but removes vertex and velocity losses over all valid frames and vertices during retraining.

Removing MRC increases geometric errors, particularly around the brows and mouth (Table 3, Figure 3(b)), despite retaining the global reference prefix. Removing vertex and velocity supervision also worsens all four metrics, supporting its contribution to geometric accuracy and motion continuity.

Table 3: Ablations on 1,620 MEAD test clips. All variants retain the global reference prefix.

Configuration	MVE ↓	LVE ↓	EVE ↓	AFDD ↓
Full ERTalk	0.52846	0.53777	0.47619	0.38181
w/o MRC	0.93336	0.76504	1.03364	0.39674
w/o Geometry Loss	0.53843	0.55860	0.48787	0.39884

4.5 Gaussian Avatar Rendering

We render ERTalk-generated facial motion using FlashAvatar (Xiang et al., 2024), a Gaussian avatar renderer. Figure 6 shows examples driven by happy and angry reference motions. See the supplementary material for more detailed visual demonstrations.

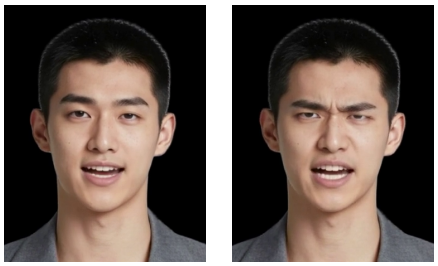

Figure 6: FlashAvatar renderings driven by happy (left) and angry (right) reference motion.

5 Discussion and Future Work

Spatiotemporal expression control. MRC provides context-dependent access to a fixed reference clip, but does not explicitly specify expression onset, duration, or transitions. Multiple references could enable control of individual facial regions or successive expression states within an utterance. Time-dependent conditioning, text instructions, or sparse keyframes could provide editable expression timing and regional control while preserving articulation and motion continuity.

Interaction with digital humans. A further direction is to extend reference-guided facial animation to conversations between users and digital humans. Facial responses in such interactions should account for the user's utterances, emotional cues, and the evolving dialogue. Updating reference conditions from conversational context could support responsive expression changes. Coordinating facial expression with gaze and head motion would help digital humans convey attention, turn-taking, and emotional responses throughout a conversation.

6 Conclusion

We present ERTalk, an autoregressive framework for speech-driven 3D facial animation guided by expressive reference motion. MRC provides fine-grained reference states that complement speech-driven articulation, while differentiable decoding connects discrete motion prediction to vertex and velocity supervision. Representing expression separately from target geometry allows the same deformation sequence to be decoded for different identities. Together, these components support facial animation that combines accurate articulation, fine-grained emotion control, and motion transfer across identities.

7 Use of Generative AI

Generative AI tools assisted with drafting, translating, and editing the paper, writing code for experimental-data processing and logging, and preparing figures, tables, and mathematical expressions.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS, 2022.

Shivangi Aneja, Justus Thies, Angela Dai, et al. FaceTalk: Audio-Driven Motion Diffusion for Neural Parametric Head Models. CVPR, 2024.

Xuangeng Chu, Nabarun Goswami, Ziteng Cui, et al. ARTalk: Speech-Driven 3D Head Animation via Autoregressive Model. SIGGRAPH Asia Conference Papers, 2025.

Xuangeng Chu, Yu Han, Wei Mao, et al. Personalizing Causal Audio-Driven Facial Motion via Dynamic Multi-modal Retrieval. arXiv:2604.23692v1, 2026.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, et al. Capture, Learning, and Synthesis of 3D Speaking Styles. CVPR, 2019.

Radek Daněček, Michael J. Black, and Timo Bolkart. EMOCA: Emotion Driven Monocular Face Capture and Animation. CVPR, 2022.

Radek Daněček, Kiran Chhatre, Shashank Tripathi, et al. Emotional Speech-Driven Animation with Content-Emotion Disentanglement. SIGGRAPH Asia Conference Papers, 2023.

Pif Edwards, Chris Landreth, Eugene Fiume, et al. JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization. ACM Transactions on Graphics, 2016.

Xiangyu Fan, Jiaqi Li, Zhiqian Lin, et al. UniTalker: Scaling up Audio-Driven 3D Facial Animation through A Unified Model. ECCV, 2024.

Yingruo Fan, Zhaojiang Lin, Jun Saito, et al. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. CVPR, 2022.

Yao Feng, Haiwen Feng, Michael J. Black, et al. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. ACM Transactions on Graphics, 2021.

Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, et al. Learning Neural Parametric Head Models. CVPR, 2023.

Kazi Injamamul Haque and Zerrin Yumak. FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning. ACM International Conference on Multimodal Interaction (ICMI), 2023.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. arXiv:2207.12598, 2022.

Ari Holtzman, Jan Buys, Li Du, et al. The Curious Case of Neural Text Degeneration. ICLR, 2020.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, et al. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. ICLR, 2017.

Tero Karras, Timo Aila, Samuli Laine, et al. Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion. ACM Transactions on Graphics, 2017.

Hyung Kyu Kim, Sangmin Lee, and Hak Gu Kim. MemoryTalker: Personalized Speech-Driven 3D Facial Animation via Audio-Guided Stylization. ICCV, 2025a.

Jisoo Kim, Jungbin Cho, Joonho Park, et al. DEEPTalk: Dynamic Emotion Embedding for Probabilistic Speech-Driven 3D Face Animation. AAAI, 2025b.

Tianye Li, Timo Bolkart, Michael J. Black, et al. Learning a Model of Facial Shape and Expression from 4D Scans. ACM Transactions on Graphics, 2017.

Chang Liu, Ye Pan, Chenyang Ding, et al. MEDTalk: Multimodal Controlled 3D Facial Animation with Dynamic Emotions by Disentangled Embedding. ACM Multimedia, 2025.

Xin Lu, Chuanqing Zhuang, Chenxi Jin, et al. LSF-Animation: Label-Free Speech-Driven Facial Animation via Implicit Feature Representation. SIGGRAPH Asia Conference Papers, 2025.

Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, et al. DiffSpeaker: Speech-Driven 3D Facial Animation with Diffusion Transformer. arXiv:2402.05712, 2024.

Federico Nocentini, Thomas Besnier, Claudio Ferrari, et al. ScanTalk: 3D Talking Heads from Unregistered Scans. ECCV, 2024.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, et al. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. CVPR, 2019.

Ziqiao Peng, Yihao Luo, Yue Shi, et al. SelfTalk: A Self-Supervised Commutative Training Diagram to Comprehend 3D Talking Faces. ACM Multimedia, 2023a.

Ziqiao Peng, Haoyu Wu, Zhenbo Song, et al. EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation. ICCV, 2023b.

Zhan Qu, Shengyu Zhang, Mengze Li, et al. ExpTalk: Diverse Emotional Expression via Adaptive Disentanglement and Refined Alignment for Speech-Driven 3D Facial Animation. IJCAI, 2025.

Alexander Richard, Michael Zollhöfer, Yandong Wen, et al. MeshTalk: 3D Face Animation from Speech Using Cross-Modality Disentanglement. ICCV, 2021.

Kang Shen, Haifeng Xia, Guangxing Geng, et al. DEITalk: Speech-Driven 3D Facial Animation with Dynamic Emotional Intensity Modeling. ACM Multimedia, 2024.

Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion. ACM SIGGRAPH Motion, Interaction and Games (MIG), 2023.

Zhiyao Sun, Tian Lv, Sheng Ye, et al. DiffPoseTalk: Speech-Driven Stylistic 3D Facial Animation and Head Pose Generation via Diffusion Models. ACM Transactions on Graphics, 2024.

Kim Sung-Bin, Lee Chae-Yeon, Gihun Son, et al. MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset. Interspeech, 2024.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. NeurIPS, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention Is All You Need. NeurIPS, 2017.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, et al. MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation. ECCV, 2020.

Sichun Wu, Kazi Injamamul Haque, and Zerrin Yumak. ProbTalk3D: Non-Deterministic Emotion Controllable Speech-Driven 3D Facial Animation Synthesis Using VQ-VAE. ACM SIGGRAPH Motion, Interaction and Games (MIG), 2024.

Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. FlashAvatar: High-fidelity Head Avatar with Efficient Gaussian Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1802–1812, 2024.

Jinbo Xing, Menghan Xia, Yuechen Zhang, et al. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. CVPR, 2023.

An Yang, Chenyu Liu, Pengcheng Xia, et al. StyleSpeaker: Audio-Enhanced Fine-Grained Style Modeling for Speech-Driven 3D Facial Animation. arXiv:2503.09852v1, 2025.

Yixiang Zhuang, Baoping Cheng, Yao Cheng, et al. Learn2Talk: 3D Talking Face Learns from 2D Talking Face. IEEE Transactions on Visualization and Computer Graphics, 2025.