AUDIOMOUTH: ARTICULATION-GROUNDED STRUCTURED GENERATION FOR SPEECH-DRIVEN MOUTH ANIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech-driven mouth animation aims to produce facial motion synchronized with speech, and existing audio-driven methods generally learn this mapping from acoustic features while leaving the articulatory structure of speech implicit in the waveform. Such implicit learning forces the generator to infer which mouth shape a sound requires from acoustics alone, and token-level objectives do not explicitly supervise the dynamics or geometry of complete motion sequences. In this paper, we aim to advance speech-driven mouth animation from implicit acoustic regression to explicit articulation-grounded generation. To this end, we propose AudioMouth, which formulates mouth animation as schema-constrained sequence generation with a multimodal language model. AudioMouth first segments long recordings into semantically coherent units sharing temporal boundaries across audio, transcript, phonemes, and target motion, so each phoneme sequence is paired with motion from the same segment. Conditioned on these units, AudioMouth generates fixed-schema ARKit coefficient sequences under explicit phoneme-to-articulation grounding, and reweights coefficient-value tokens to reduce the relative contribution of frequent near-zero targets to supervision. To supervise the trajectory as a whole, AudioMouth further refines the generator with Group Relative Policy Optimization, whose reward measures coefficient accuracy, temporal dynamics, and a simplified lip-geometry discrepancy that reads 3D mouth deformation out of a fixed blendshape basis without rendering or landmark detection. On our selected subject-disjoint BEAT split, using the provided linguistic annotations, AudioMouth improves every reported metric over the four evaluated baseline configurations, reducing Fréchet distance by 57% and coefficient MSE by 70% relative to the best baseline in each metric. Our ablations support the benefits of articulation guidance for distributional fidelity, balanced supervision for coefficient reconstruction, and geometric feedback for reduced across-script FD dispersion.

1 INTRODUCTION

Speech-driven facial animation aims to generate mouth motion temporally synchronized with speech, supporting digital avatars, game production, and embodied interaction. Recent work has steadily improved the visual quality of talking faces across portrait animation (Ji et al., 2021; Zhang et al., 2023; Xu et al., 2024b; Jiang et al., 2024; Xu et al., 2024a), 3D facial animation (Fan et al., 2022a; Xing et al., 2023), and audio-conditioned neural rendering (Guo et al., 2021; Chen et al., 2025; Ma et al., 2026). Compact facial representations such as 3D morphable models, FLAME, and blendshape parameterizations provide practical control spaces (Blanz & Vetter, 2023; Li et al., 2017; Menzel et al., 2022), among which ARKit defines semantically meaningful facial action coefficients (Apple Inc., 2023) and has become a standard interface between speech and mouth motion in production. However, how a model decides which mouth shape a given speech sound requires is still left to be discovered from acoustics.

The difficulty lies in the size of the mapping that must be learned from the waveform alone. For instance, a 10-second utterance at 30 Hz spans 300 frames over 32 mouth-related ARKit channels, so a model must emit 9,600 coefficients whose correct values are governed by only about one hundred

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

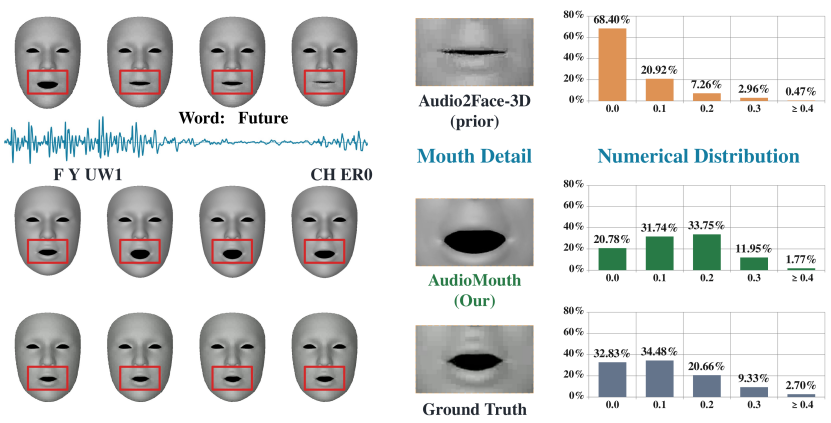

Figure 1: Comparison of Audio2Face-3D, AudioMouth, and ground truth. In the illustrated portion of "future", the snapshots (left) and mouth close-ups (middle) show Audio2Face-3D closing the mouth while AudioMouth and the reference maintain an open configuration. The histograms (right) show a near-zero coefficient bias in Audio2Face-3D, whereas the reference has more nonzero values, motivating explicit articulation grounding and distribution-balanced supervision.

phonemes. Existing audio-driven methods generally predict motion from learned acoustic representations (Cudeiro et al., 2019; Richard et al., 2021; Fan et al., 2022a; Park & Cho, 2023), which asks the model to rediscover phoneme identity, phoneme boundaries, and the articulatory action of each phoneme as a by-product of fitting motion. Maintaining local audio–motion correspondence in long utterances remains challenging, motivating explicit priors that distinguish articulatory events such as bilabial closure and jaw opening. Figure 1 illustrates a failure case in which a baseline does not sustain the reference mouth configuration within a word, alongside a concentration of predicted coefficients near zero.

Existing attempts to inject linguistic structure into this mapping fall into two lines. Early efforts build viseme inventories and phoneme-to-shape rules (Cohen & Massaro, 1993; Taylor et al., 2017; Zhou et al., 2018), which make articulation explicit but, without additional coarticulation modeling (Edwards et al., 2016), compress speech into a small discrete inventory, losing the coarticulation that determines how strongly each action is realized. More recently, a line of work injects textual or linguistic cues as auxiliary features of an audio-driven generator (Fan et al., 2022b; Xu et al., 2024c), and language-aware facial models (Kang et al., 2026; Wu et al., 2025) show that language-aligned supervision helps estimate facial actions. Although these methods enrich the input, the cue is consumed as one more feature vector and the model is not explicitly told what action a phoneme category names. In essence, both lines leave the connection between a phonetic unit and a named facial control outside explicit natural-language instructions. Fortunately, recent multimodal language models (Xu et al., 2025) consume audio and text jointly and follow natural-language instructions, which makes it possible to state that connection explicitly rather than relying solely on the model to learn it gradually from data.

We aim to advance speech-driven mouth animation from implicit acoustic regression to explicit articulation-grounded generation: linguistic content, phonemes, and their articulatory meaning condition the model, and named facial controls form its output. Two obstacles arise: *boundary mismatch*, since phoneme sequences and target motion must cover the same segment while long unsegmented recordings hinder local correspondence, and *value imbalance*, since most channels are inactive at any instant and frequent near-zero values receive more supervision than the larger values that shape visible motion. We propose AudioMouth, which formulates this task as schema-constrained generation with a multimodal language model. It constructs semantically aligned audio-linguistic units with shared boundaries across audio, transcript, phonemes, and motion (Figure 2). Explicit *articulation grounding* names the controls driven by each phoneme group, while *distribution-balanced supervision* downweights frequent near-zero targets to give active channels proportionally stronger supervision. Since next-token likelihood does not explicitly evaluate decoded temporal or geometric discrepancies, AudioMouth finally refines the generator with Group Relative Policy Optimization using coefficient accuracy, temporal dynamics, and a *simplified lip-geometry discrepancy* (SLMD)

that computes 3D mouth deformation from a fixed blendshape basis without rendering or landmark detection.

Experiments on our selected subject-disjoint BEAT split (Liu et al., 2022), using the provided linguistic annotations, show that AudioMouth improves every reported metric over the four evaluated baseline configurations, reducing Fréchet distance by 57% and coefficient MSE by 70% relative to the best baseline in each metric. In summary, our key contributions are three-fold.

- **Articulation-Grounded Generator:** To the best of our knowledge, AudioMouth is the first framework that generates ARKit mouth trajectories inside a multimodal language model, where phonemes and their articulatory meaning are explicit inputs and the named facial controls are the output schema.
- **Motion-Aware Refinement:** We propose a sequence-level refinement stage whose reward evaluates coefficient accuracy, temporal dynamics, and mouth geometry, and we introduce SLMD to obtain 3D geometric feedback directly from coefficient differences without rendering or landmark detection.
- **Strong Performance:** AudioMouth outperforms the four evaluated baseline configurations on all five metrics under our selected subject-disjoint BEAT split with provided linguistic annotations. Our ablations support the benefits of articulation guidance for distributional fidelity, balanced supervision for coefficient reconstruction, and geometric feedback for reduced across-script FD dispersion.

2 Related Work

Speech-Driven 3D Facial Animation. Speech-driven 3D facial animation generates temporally synchronized facial motion from audio, and existing methods learn a direct mapping from acoustic features to facial geometry. Early efforts predict vertex displacements, where VOCA (Cudeiro et al., 2019) produces speaker-specific vertices and MeshTalk (Richard et al., 2021) learns a categorical latent space covering non-verbal motion, and Transformer-based models later improved long-range temporal modeling (Fan et al., 2022a; 2024). Because the audio-to-motion mapping is one-to-many, these regressors can average over plausible motions and produce over-smoothed results, which CodeTalker (Xing et al., 2023) mitigates with discrete motion priors and which diffusion-based methods (Park & Cho, 2023; Stan et al., 2023; Sun et al., 2025) address by modeling motion stochastically. In production settings, Audio2Face-3D (Chung et al., 2025) and related avatar pipelines (Wei et al., 2024; Lin et al., 2025; Sun et al., 2024; Zhen et al., 2025) provide real-time animation, SentiAvatar (Jin et al., 2026) plans at the sentence level, and Ex-Omni (Zhang et al., 2026) uses an omni-modal model for semantic understanding while delegating coefficient prediction to a lightweight generator. Across this line, the linguistic structure of speech does not enter the module that decides mouth shape as explicit phoneme-to-control instructions.

Linguistic and Articulatory Priors. Speech articulation is tied to phonetic structure, where bilabial consonants require lip closure, open vowels require larger jaw opening, and phoneme transitions shape the temporal profile of motion (Cohen & Massaro, 1993; Taylor et al., 2012; Edwards et al., 2016). Viseme-based methods encode these relations as explicit rules or discrete inventories (Zhou et al., 2018; Taylor et al., 2017; Bao et al., 2023), which makes articulation interpretable at the potential cost of collapsing continuous coarticulation into a few categories. Later methods relax this constraint by supplying higher-level cues as auxiliary conditions, including emotion signals (Peng et al., 2023; Daněček et al., 2023; Gan et al., 2023; Jin et al., 2026) and linguistic features (Fan et al., 2022b; Xu et al., 2024c; Wu et al., 2025). In these designs, linguistic or affective cues enter as feature representations rather than explicit natural-language instructions linking phoneme groups to named facial controls.

Positioning of AudioMouth. The output representation affects how directly phoneme-to-control guidance can be expressed. Vertex-based representations (Cudeiro et al., 2019; Fan et al., 2022a; Aloni et al., 2025) preserve fine detail without directly naming articulatory actions, and parametric models such as FLAME (Li et al., 2017) include compact statistical coordinates that are less directly associated with named articulatory actions. ARKit blendshapes (Apple Inc., 2023) instead define named facial actions, which lets one vocabulary describe the articulatory cause and the generated control. In this work, we focus on the articulatory structure of speech, with the goal of making it an explicit part of both the input and the output of mouth-motion generation. Unlike previous ap-

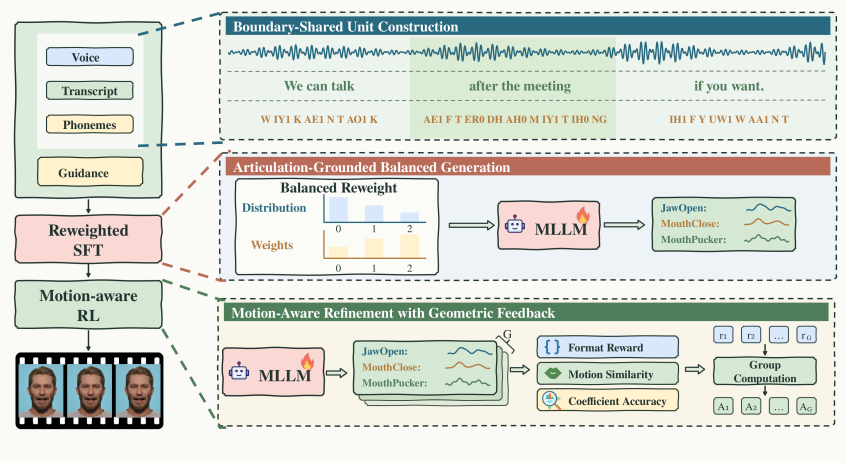

Figure 2: AudioMouth makes articulation explicit at every stage. Aligned units tie each phoneme sequence to motion from the same segment (upper right), grounding and balanced supervision guide the mapping from those phonemes to named facial controls (middle right), and a motion-level reward including 3D lip geometry explicitly evaluates decoded motion beyond token likelihood (lower right).

proaches that treat linguistic information as auxiliary features of an acoustic regressor, AudioMouth adopts a more deliberate design by generating named ARKit controls inside a multimodal language model that is told qualitative associations between phoneme groups and named facial controls, and by evaluating the decoded trajectory as a whole through a reward whose geometric term reads out how the predicted coefficients deform the mouth surface.

3 AUDIOMOUTH

We introduce AudioMouth, which generates mouth-related ARKit trajectories by making the articulatory structure of speech explicit on both sides of the mapping. We first organize long recordings into units where a phoneme sequence and the target motion share one segment, then generate a fixed-schema coefficient sequence under phoneme-to-articulation grounding with supervision balanced across coefficient values, and finally refine the generator with a reward computed on the decoded motion. The formulation, the three stages, and inference are introduced as follows.

3.1 STRUCTURED MOTION GENERATION

Given a speech waveform $\mathcal{A} = \{a_n\}_{n=1}^{N}$ of N samples, our goal is to predict a synchronized coefficient sequence $\mathcal{V} = \{v_t\}_{t=1}^{T}$ over T motion frames, where each $v_t \in \mathbb{R}^K$ holds the $K = 32$ mouth-related ARKit controls (Apple Inc., 2023) retained by SAiD (Park & Cho, 2023). We divide each recording into J aligned units, so unit j carries the inputs and target

$$\mathcal{X}^j = \left(\mathcal{A}^j, \mathcal{W}^j, \mathcal{P}^j, \mathcal{G}\right), \quad \mathcal{Y}^j = f_{\text{json}}\left(\mathcal{V}^j\right), \tag{1}$$

where $\mathcal{W}^j$ and $\mathcal{P}^j$ are the transcript and phoneme sequence of the audio segment $\mathcal{A}^j$, $\mathcal{G}$ is the phoneme-to-articulation grounding of Sec.3.3, $\mathcal{V}^j$ is the target motion, and $f_{\text{json}}(\cdot)$ serializes each ARKit channel into a predefined key holding that channel in temporal order. A fixed schema turns animation into generation without a regression head. The generator $f_{\text{gen}}(\cdot)$ with parameters θ models $\mathcal{Y}^j = (y_1, \cdots, y_{M_j})$ autoregressively as

$$p_\theta\left(\mathcal{Y}^j \mid \mathcal{X}^j\right) = \prod_{m=1}^{M_j} p_\theta\left(y_m \mid y_{<m}, \mathcal{X}^j\right), \tag{2}$$

where M_j is the output length and $y_{<m}$ the preceding tokens. We therefore align the conditioning and target segments, reweight frequent coefficient values in the token-level objective, and add explicit evaluation of the decoded motion during refinement.

3.2 SEMANTIC-ALIGNED AUDIO-LINGUISTIC UNITS

Boundary-Shared Unit Construction. To make a phoneme sequence informative about motion, the sequence and the motion must refer to the same segment, while long recordings make local correspondence harder to learn. We therefore build units whose boundaries are shared by all four signals, using automatic speech recognition $f_{\mathrm{asr}}(\cdot)$ (Gao et al., 2023; Zhang et al., 2022) for the transcript and word-level timestamps τ, and semantic segmentation $f_{\mathrm{seg}}(\cdot)$ (Frohmann et al., 2024) for coherent segments,

$$\left(\tilde{\mathcal{W}}, \tau\right) = f_{\mathrm{asr}}(\mathcal{A}), \quad \{\mathcal{W}^j\}_{j=1}^{J} = f_{\mathrm{seg}}\left(\tilde{\mathcal{W}}\right). \tag{3}$$

The first and last words of $\mathcal{W}^j$ determine one interval through τ, which selects $\mathcal{A}^j$, $\mathcal{P}^j$, and, during training, $\mathcal{V}^j$. We use semantic boundaries to preserve local linguistic context, since a fixed-length window can split a word and its surrounding phonetic context. At inference the same construction runs without the motion branch.

3.3 ARTICULATION-GROUNDED BALANCED GENERATION

Articulation Grounding. Aligned phoneme sequences tell the model which speech sounds occur within a segment but not what the mouth does to produce it, and an audio-driven regressor generally has to recover that relation from motion data. AudioMouth instead states it as a prior. Specifically, the grounding $\mathcal{G}$ in Eq.(1) is a fixed qualitative statement mapping phonetic categories to the named ARKit controls they drive, connecting bilabial consonants to lip closure and open vowels to larger jaw opening. The output schema reuses those control names, providing a qualitative prior for learning detailed trajectories from training targets; the audio supplies timing and acoustic cues for realizing each action. We call this division *articulation grounding*, and it is why phonemes and their meaning are supplied together.

Distribution-Balanced Supervision. Grounded conditioning still leaves the objective unbalanced, because most mouth channels are inactive at any instant and their near-zero values dominate the token stream. Under the plain likelihood of Eq.(2), frequent low-valued coefficients appear at more supervised positions than the large values that produce visible motion. We therefore reweight the loss with a fixed function $\omega(\cdot)$ that downweights the frequent low-valued tokens, giving

$$\mathcal{L}_{\mathrm{db}} = -\sum_{m=1}^{M_j} \omega(y_m) \log p_\theta\left(y_m \mid y_{<m}, \mathcal{X}^j\right), \tag{4}$$

where $\omega(y_m)$ depends only on the value a coefficient token carries and equals one on every schema token, so keys and punctuation retain unit loss weights. The weighting is a monotone table fixed before training, with values in Appendix C.1. Supervised fine-tuning with Eq.(4) yields θ_{sft}, which initializes refinement.

3.4 MOTION-AWARE REFINEMENT WITH GEOMETRIC FEEDBACK

Even a well-balanced token objective does not explicitly measure temporal or geometric discrepancies in the decoded trajectory. We close this gap by refining θ_{sft} with Group Relative Policy Optimization (GRPO) (Shao et al., 2024) against a reward computed on the decoded motion, where we sample G candidates per input $\mathcal{X}^j$, parse each into a trajectory through $f_{\mathrm{parse}}(\cdot)$, and compare it with the reference $\mathcal{V}^j$.

Simplified Lip-Geometry Discrepancy. Coefficient-wise errors treat the K controls as independent while the visible mouth is their joint result, so two candidates with equal coefficient error can differ greatly in shape, motivating a geometric discrepancy that can be computed without rendering each candidate inside the sampling loop. We therefore introduce the *simplified lip-geometry discrepancy* (SLMD), which reads the shape difference out of the coefficients through a fixed blendshape basis. Let $\mathbf{C}^p, \mathbf{C}^g \in \mathbb{R}^{T\times K}$ denote the candidate and reference sequences, and let $\mathbf{b}_{k,m} \in \mathbb{R}^3$ be the displacement the k-th blendshape applies to mouth-region vertex $m \in \mathcal{M}$ relative to the neutral mesh. Since both sequences are driven from the same neutral geometry, the neutral term cancels and

the per-vertex displacement difference $\delta \mathbf{x}_{t,m} = f_{\text{geo}}(\mathbf{C}^p - \mathbf{C}^g)$ and the resulting discrepancy are

$$\delta \mathbf{x}_{t,m} = \sum_{k=1}^{K} \left(c^p_{t,k} - c^g_{t,k} \right) \mathbf{b}_{k,m}, \tag{5}$$

$$\mathcal{L}_{\text{slmd}} = \frac{\sum_{t=1}^{T} \sum_{m \in \mathcal{M}} w_m \left\| \delta \mathbf{x}_{t,m} \right\|_2}{T \sum_{m \in \mathcal{M}} w_m}, \tag{6}$$

where the fixed vertex weights w_m emphasize the lip region while retaining lower-weighted surrounding vertices, and T is the compared length. SLMD therefore costs one matrix product per candidate, couples all K controls through their shared effect on the surface, and stays in the coefficient domain, which keeps it distinct from the rendered-video landmark metric used for evaluation.

Motion-Level Reward. Mouth quality involves both the shape at each instant and how that shape evolves, so SLMD alone is not sufficient. We combine it with the coefficient error $\mathcal{L}_{\text{mse}}$ and the first- and second-order temporal difference discrepancies $\mathcal{L}_{\text{vel}}$ and $\mathcal{L}_{\text{acc}}$, which compare how the candidate changes against how the reference changes instead of penalizing movement itself. The four terms live on different numerical scales, so each is mapped to a bounded score before aggregation, giving

$$S_q = (1 + \mathcal{L}_q / s_q)^{-1}, \quad r_{\text{motion}} = \sum_{q \in \mathcal{Q}} \lambda_q S_q, \quad \mathcal{Q} = \{\text{mse}, \text{vel}, \text{acc}, \text{slmd}\}, \tag{7}$$

where each scale s_q is the median of $\mathcal{L}_q$ measured once on reference data so the four scores occupy a comparable range, and the nonnegative weights λ_q sum to one. Invalid schema instances receive a fixed negative reward and valid candidates of the wrong length are penalized in proportion to their length error, which keeps format compliance and duration control inside one signal. The scales, weights, and penalties are listed in Appendix C.4.

Group-Relative Policy Update. To turn these absolute rewards into a learning signal without a value network, we normalize them within each sampled group as $\hat{A}_i = (r_i - \bar{r})/(\sigma_r + \epsilon)$, where $\bar{r}$ and σ_r are the group mean and standard deviation and ϵ ensures numerical stability. Since all G candidates answer the same input, the update favors candidates with higher combined rewards for that input. Starting from θ_{sft}, refinement adds explicit sequence-level feedback on decoded motion, format validity, and duration.

Inference. Given test speech, we build input units as in Sec.3.2 and generate one coefficient JSON per unit with θ_{grpo}. The decoded segments are concatenated in temporal order and passed through $f_{\text{post}}(\cdot)$, which suppresses residual jitter and the small discontinuities appearing where two independently generated units meet, giving

$$\hat{\mathcal{V}} = f_{\text{post}}\left(\text{Concat}\left(\left\{ f_{\text{parse}}\left(\hat{\mathcal{Y}}^j \right) \right\}_{j=1}^{J} \right) \right). \tag{8}$$

Algorithm 1 in Appendix D summarizes the procedure, and Appendix E verifies that $f_{\text{post}}(\cdot)$ complements refinement instead of substituting for it.

Unless explicitly stated otherwise in a targeted ablation, the AudioMouth settings described in the method, experimental setup, and appendix are the default settings shared across experiments. Each targeted ablation changes only the component or setting specified in its description.

4 Experiments

4.1 Experimental Setting

Benchmark. We conduct experiments on BEAT (Liu et al., 2022), which pairs natural speech with ARKit coefficients captured at 60 Hz and is therefore directly usable for an Audio2ARKit comparison. We use its English subset and screen it by rendering the captured coefficients against the audio, discarding sequences with tracking failures or audible misalignment, since a reference that does not match its own audio cannot support training or evaluation. The retained data are split by speaker into 176 training, 40 validation, and 60 test scripts, and semantic segmentation

Table 1: AudioMouth achieves the lowest values on all five metrics among the evaluated configurations on our selected subject-disjoint BEAT test split. All metrics are lower-is-better. FD, WInD, and LMD are mean ± sample standard deviation across evaluation scripts, **bold** marks the best result per column.

Methods	MSE ↓	MAE ↓	FD ↓	WInD ↓	LMD ↓
LAM-A2E	0.023	0.109	38.403 ± 10.149	41.751 ± 10.885	7.648 ± 0.497
Audio2Face-3D	0.020	0.106	25.228 ± 3.560	29.879 ± 4.454	7.879 ± 0.438
SAiD	0.021	0.108	58.357 ± 21.580	62.256 ± 21.911	7.651 ± 0.504
SentiAvatar	0.022	0.106	40.609 ± 10.277	42.153 ± 10.542	8.131 ± 1.005
AudioMouth	**0.006**	**0.053**	$\mathbf{10.820 \pm 2.635}$	$\mathbf{16.513 \pm 3.584}$	$\mathbf{6.249 \pm 0.723}$

yields 15,655 training, 2,858 validation, and 2,946 test units. For this benchmark, transcripts and phonemes are obtained from the provided annotations in all splits, including testing. Appendix B details the procedure.

Metrics. We report five metrics covering accuracy, distribution, and appearance, where MSE and MAE measure frame-level coefficient error. Fréchet distance (FD) and Wasserstein inception distance (WInD) measure discrepancies between predicted and reference motion distributions, following the feature-extraction and evaluation protocol of SAiD (Park & Cho, 2023) with its released VAE weights. Lip landmark distance (LMD) measures the mean Euclidean distance between 2D mouth landmarks in paired rendered videos, and is the only metric outside the coefficient domain, which makes it distinct from the geometric term inside our reward. LMD follows the protocol of prior work (Chen et al., 2018).

Baselines. We compare AudioMouth with four Audio2ARKit methods, namely LAM-A2E (He et al., 2025), NVIDIA Audio2Face-3D (Chung et al., 2025), SAiD (Park & Cho, 2023), and SentiAvatar (Jin et al., 2026), covering real-time production pipelines, diffusion-based generation, and sentence-level planning. SAiD is fine-tuned from its public checkpoint on our training split, while the other three are evaluated with released weights because their training code or data is unavailable. Two asymmetries follow. The three unadapted baselines are not fine-tuned on our training split, and all four consume audio alone while AudioMouth also consumes the transcript and phonemes its formulation requires. Table 1 therefore compares the reported system configurations, and the Audio Only ablation of Sec.5.1 is the matched-input reference under our own backbone and data.

Configuration. We build AudioMouth on Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025) and adapt it with LoRA (Hu et al., 2021) on the language decoder, keeping the audio encoder and multimodal alignment modules frozen so adaptation concerns only how motion is generated. General preprocessing combines automatic speech recognition with SAT (Frohmann et al., 2024); dataset-specific handling of the linguistic annotations is described in Appendix B. Supervised fine-tuning and GRPO each run for four epochs, refinement samples six candidates per prompt, all training and inference use NVIDIA A100 GPUs, and predictions are rendered with MetaHuman (Epic Games, 2025). Appendix C.3 reports the full configuration.

4.2 Main Results

Capability on Speech-Driven Coefficient Prediction. We evaluate all methods of Sec.4.1 on the same speech intervals and 32 channels. Table 1 shows that AudioMouth achieves the lowest values on all five metrics among the evaluated configurations, reducing MSE by 70% and Fréchet distance by 57% relative to the best baseline in each column. The four baselines differ by at most 15% in frame-level MSE and MAE, but by more than a factor of two in FD and WInD, showing that distributional metrics provide complementary information. AudioMouth improves both metric groups and has the lowest across-script FD standard deviation, supporting improved reconstruction and motion-distribution fidelity under this setting. The conditioning ablation in Sec.5.1 further evaluates articulation guidance.

Capability on Rendered Mouth Appearance. We render every prediction and reference with MetaHuman and evaluate LMD outside the coefficient domain. AudioMouth reduces LMD by 18% relative to the best baseline (Table 1), compared with 57% on FD. These reductions describe different aspects of quality: LMD measures paired landmark distances, whereas FD measures distributional discrepancies in motion features. In Figure 3, some baselines produce a smaller mouth opening on /L/ in "family" and a less open configuration on /K/ in "cat". AudioMouth more closely matches

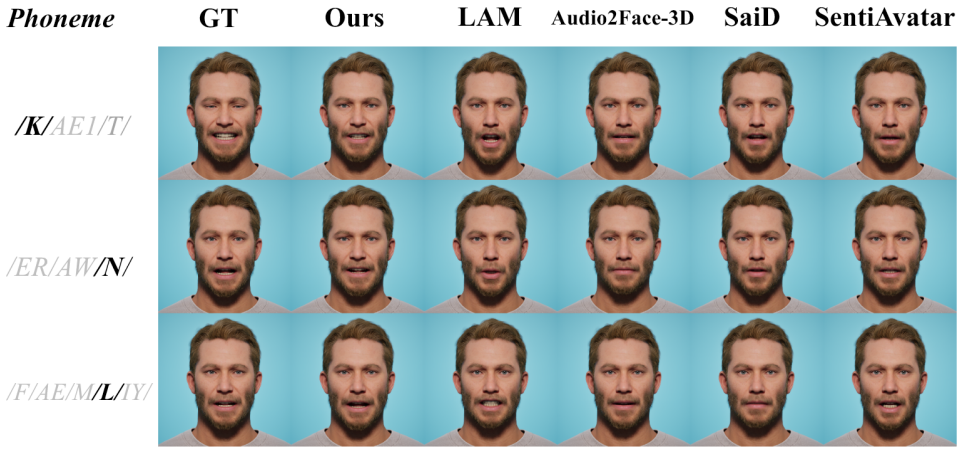

Figure 3: AudioMouth reproduces the reference mouth configuration at the highlighted phonemes, maintaining a moderate opening with relaxed lips on /L/ in "family" and an open, laterally spread mouth on /K/ in "cat", where some baselines keep a smaller, weakly differentiated mouth opening. The leftmost column shows the phoneme context with the active phoneme in bold.

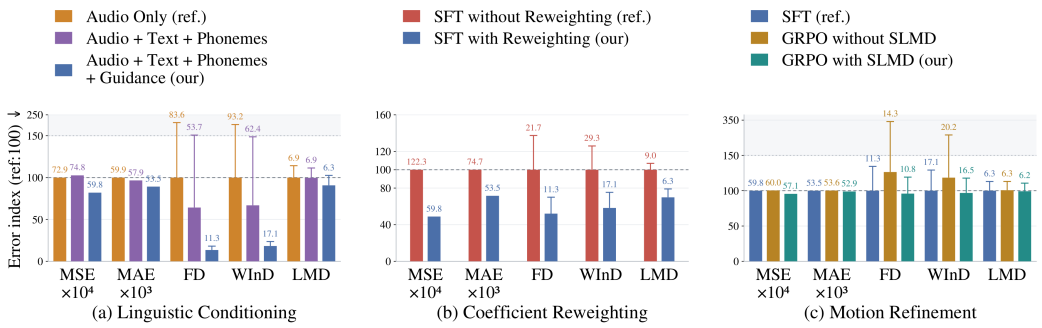

Figure 4: Ablations of (a) linguistic and articulatory conditioning, (b) coefficient-value reweighting, and (c) motion-aware refinement. Panels (a) and (b) compare SFT models. Within each panel, bar heights are normalized to the reference mean for each metric (reference = 100; lower is better). Upper error bars show one across-script sample standard deviation for FD, WInD, and LMD, scaled by the same reference mean. Numeric labels show measured values, with MSE and MAE scaled as indicated. Panel y-axis ranges differ; panels (a) and (c) use continuous piecewise-linear axes with vertical distances above 150 compressed by a factor of four.

the reference lip and jaw configuration in both displayed contexts, showing that its coefficient-level gains also yield improved rendered appearance in these examples.

5 Analysis

5.1 Ablation Study on Articulation Grounding

To verify that articulation grounding contributes beyond the capacity of the multimodal backbone, we ablate the linguistic inputs at the SFT stage while holding the backbone, the training data, the objective, and the post-processing identical. The fully conditioned SFT model receives audio, transcript, phonemes, and the grounding statement. The w/o Guidance variant keeps transcript and phonemes but deletes the statement naming which control each phoneme group drives, and the Audio Only variant deletes all three, which reduces the model to a structured audio-to-coefficient regressor.

As shown in Figure 4(a), the three settings differ by at most a quarter in MSE and about a tenth in MAE, yet by almost an order of magnitude in FD and WInD, where the fully conditioned SFT model substantially reduces FD by about 79% relative to w/o Guidance. Severe deviations on a small number of scripts drive the large across-script standard deviations of the ungrounded variants, which

reach values comparable to their means, whereas the fully conditioned SFT model has a standard deviation of only a few units. Adding the guidance block substantially lowers FD and WInD while producing smaller relative changes in coefficient errors. The w/o Guidance variant already cuts the FD of Audio Only by more than a third while the two stay comparable at the frame level, indicating a benefit from the joint transcript and phoneme inputs. The results show that the phoneme-specific guidance block improves distributional fidelity beyond conditioning on transcripts and phonemes alone.

5.2 Superiority of Distribution-Balanced Supervision

To assess the effect of coefficient-value imbalance, we compare the weighted SFT objective with SFT without reweighting, which assigns unit weight to coefficient-value tokens while retaining the same conditioning and post-processing (Figure 4(b)). Reweighting improves all five reported metrics, indicating benefits for both coefficient reconstruction and the resulting motion distribution and mouth shape. This comparison evaluates the contribution of the supervised objective independently of adding linguistic inputs or GRPO refinement. Downweighting frequent low-valued targets reduces their dominance in the loss and gives larger coefficient values greater relative emphasis during training. The observed gains are consistent with this motivation and support distribution-balanced supervision for structured coefficient generation.

5.3 Effect of Motion-Aware Refinement and Geometric Feedback

To evaluate the effect of the GRPO refinement stage, we compare the supervised and refined generators under identical post-processing. As shown in Figure 4(c), refinement improves all five metrics, and the gain in the means is modest at roughly 4% on MSE and FD. The relative decrease in across-script FD dispersion is larger than the decrease in mean FD. In particular, GRPO reduces the across-script FD standard deviation from 3.888 to 2.635, a decrease of about 32%, indicating lower FD dispersion across the evaluated speech inputs.

Supervised fine-tuning maximizes reference-token likelihood, while GRPO compares candidates for the same input using rewards for decoded coefficient accuracy, temporal dynamics, and mouth geometry. This trajectory-level feedback complements token-level supervision and is consistent with the observed accuracy gains and lower across-script FD dispersion.

We next isolate the geometric term itself. To verify that SLMD contributes something beyond the coefficient and temporal terms already in the reward, we remove it from Eq.(7) and renormalize the remaining weights, keeping the sampling procedure, candidate count, and schedule unchanged, so the w/o SLMD variant still receives coefficient and temporal feedback and tests the geometric term specifically.

As shown in Figure 4(c), removing SLMD does more than reduce the benefit of refinement, since the resulting FD is worse than that of the supervised model it started from and its across-script deviation exceeds its own mean. In this experiment, refinement without geometric feedback produces higher FD and greater across-script variation. The joint geometric effect of the controls provides a motivation for including SLMD. Coefficient error and temporal differences score the K channels independently with equal channel weights, so candidates with similar coefficient and temporal errors can still produce very different mouth surfaces through the joint geometric effect of their controls. SLMD helps distinguish these candidates, because Eq.(5) maps every channel difference onto shared mouth vertices and sums the resulting displacements before measuring the surface discrepancy, allowing geometric displacement compensation while favoring the resulting surface closest to the reference. The results support the contribution of SLMD to lower FD and reduced across-script variation compared with refinement using coefficient and temporal feedback alone.

6 Conclusion

We present AudioMouth, which generates ARKit mouth trajectories by making articulation explicit on both sides of the speech-to-motion mapping. On our selected BEAT split with provided linguistic annotations, it improves every reported metric over the four evaluated baseline configurations. Our ablations support the benefits of articulation guidance for distributional fidelity, balanced supervision for coefficient reconstruction, and geometric feedback for reduced across-script FD dispersion.

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, implement methods, and interpret results. We did not use generative AI tools to generate synthetic datasets, assist with translation, clean or reformat datasets, or support qualitative and thematic data analysis. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to summarize or analyze existing literature, discover research topics or identify gaps, brainstorm, and revise the wording and organization of the manuscript. We have reviewed all AI-assisted work. For manuscript preparation, we reviewed AI-assisted drafts through iterative revision, checking that the wording and organization reflected the intended method. We also corrected the presentation of pseudocode, citation formatting, and figure captions to maintain consistency with the manuscript and figures. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yaron Aloni, Rotem Shalev-Arkushin, Yonatan Shafir, Guy Tevet, Ohad Fried, and Amit Haim Bermano. Express4d: Expressive, friendly, and extensible 4d facial motion generation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 485–494, 2025.

Apple Inc. Arkit. `https://developer.apple.com/documentation/arkit`, 2023. URL `https://developer.apple.com/documentation/arkit`. Apple Developer Documentation.

Linchao Bao, Haoxian Zhang, Yue Qian, Tangli Xue, Changhai Chen, Xuefei Zhe, and Di Kang. Learning audio-driven viseme dynamics for 3d face animation. *arXiv preprint arXiv:2301.06059*, 2023.

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

Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 520–535, 2018.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2403–2410, 2025.

Chaeyeon Chung, Ilya Fedorov, Michael Huang, Aleksey Karmanov, Dmitry Korobchenko, Roger Ribera, Yeongho Seol, et al. Audio2face-3d: Audio-driven realistic facial animation for digital avatars. *arXiv preprint arXiv:2508.16401*, 2025.

Michael M Cohen and Dominic W Massaro. Modeling coarticulation in synthetic visual speech. In *Models and techniques in computer animation*, pp. 139–156. Springer, 1993.

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.

Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–13, 2023.

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11, 2016.

Epic Games. Metahuman: High-fidelity digital humans made easy. `https://www.metahuman.com/en-US`, 2025. Accessed: 2025-06-30.

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, 2022a.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Joint audio-text model for expressive speech-driven 3d facial animation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–15, 2022b.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11908–11941, 2024.

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22634–22645, 2023.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5784–5794, 2021.

Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Papers*, pp. 1–13, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14080–14089, 2021.

Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024.

Chuhao Jin, Rui Zhang, Qingzhe Gao, Haoyu Shi, Dayu Wu, Yichen Jiang, Yihan Wu, and Ruihua Song. Sentiavatar: Towards expressive and interactive digital humans. *arXiv preprint arXiv:2604.02908*, 2026.

Zejian Kang, Kai Zheng, Yuanchen Fei, Wentao Yang, Hongyuan Zou, and Xiangru Huang. Semanticface: Semantic facial action estimation via semantic distillation in interpretable space. *arXiv preprint arXiv:2603.14827*, 2026.

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

Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, Zerong Zheng, and Yanbo Zheng. Cyberhost: A one-stage diffusion framework for audio-driven talking body generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pp. 612–630. Springer, 2022.

Chuhang Ma, Shuai Tan, Ye Pan, Jiaolong Yang, and Xin Tong. Esgaussianface: Emotional and stylized audio-driven facial animation via 3d gaussian splatting. *IEEE Transactions on Visualization and Computer Graphics*, 2026.

Timo Menzel, Mario Botsch, and Marc Erich Latoschik. Automated blendshape personalization for faithful face animations using commodity smartphones. In *Proceedings of the 28th ACM Symposium on virtual reality software and technology*, pp. 1–9, 2022.

Inkyu Park and Jaewoong Cho. Said: Speech-driven blendshape facial animation with diffusion. *arXiv preprint arXiv:2401.08655*, 2023.

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, 2023.

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

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

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.

Wenzhang Sun, Xiang Li, Donglin Di, Zhuding Liang, Qiyuan Zhang, Hao Li, Wei Chen, and Jianxun Cui. Uniavatar: Taming lifelike audio-driven talking head generation with comprehensive motion and lighting control. *arXiv preprint arXiv:2412.19860*, 2024.

Xusen Sun, Longhao Zhang, Hao Zhu, Peng Zhang, Bang Zhang, Xinya Ji, Kangneng Zhou, Daiheng Gao, Liefeng Bo, and Xun Cao. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. In *2025 International Conference on 3D Vision (3DV)*, pp. 713–722. IEEE, 2025.

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017.

Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pp. 275–284, 2012.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Jingchao Wu, Zejian Kang, Haibo Liu, Yuanchen Fei, and Xiangru Huang. Keyframeface: From text to expressive facial keyframes. *arXiv preprint arXiv:2512.11321*, 2025.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024a.

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024b.

Zhihao Xu, Shengjie Gong, Jiapeng Tang, Lingyu Liang, Yining Huang, Haojie Li, and Shuangping Huang. Kmtalk: Speech-driven 3d facial animation with key motion embedding. In *European Conference on Computer Vision*, pp. 236–253. Springer, 2024c.

Haoyu Zhang, Zhipeng Li, Yiwen Guo, and Tianshu Yu. Ex-omni: Enabling 3d facial animation generation for omni-modal large language models. *arXiv preprint arXiv:2602.07106*, 2026.

Hui Zhang, Tian Yuan, Junkun Chen, Xintong Li, Renjie Zheng, Yuxin Huang, Xiaojie Chen, Enlei Gong, Zeyu Chen, Xiaoguang Hu, et al. Paddlespeech: An easy-to-use all-in-one speech toolkit. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pp. 114–123, 2022.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8652–8661, 2023.

Dingcheng Zhen, Shunshun Yin, Shiyang Qin, Hou Yi, Ziwei Zhang, Siyuan Liu, Gan Qi, and Ming Tao. Teller: Real-time streaming audio-driven portrait animation with autoregressive motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21075–21085, 2025.

Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (ToG)*, 37(4):1–10, 2018.

AudioMouth: Articulation-Grounded Structured Generation for Speech-Driven Mouth Animation
Supplementary Material

A Qualitative Comparison



Figure 5: Qualitative comparison on a held-out BEAT recording rendered with the same MetaHuman character. From top to bottom: ground truth, AudioMouth, LAM-A2E, SAiD, SentiAvatar, and AudioFace-3D. Columns show ten synchronized frames sampled at 5 Hz from 2.0 to 3.8 seconds.

B Dataset Preparation

Source Data. BEAT Liu et al. (2022) contains 76 hours of multimodal recordings from 30 speakers across multiple languages and emotional conditions. Its facial data comprise 52-dimensional ARKit coefficients captured at 60 Hz with an iPhone 12 Pro, synchronized with 48-kHz stereo audio. We use its English subset. The test split contains 60 scripts, yielding 2,946 aligned units.

Screening. We render coefficient sequences with MetaHuman Epic Games (2025) and manually inspect them together with the synchronized audio. We discard samples with corrupted or implausible mouth motion, including occlusion-induced tracking failures, or clear audio-motion misalignment.

General and In-the-Wild Recordings. The general pipeline uses ASR to obtain transcripts and word-level timestamps, followed by SAT Frohmann et al. (2024) for semantic sentence segmentation. The resulting word spans define shared temporal boundaries for the audio and linguistic inputs. Where paired motion is available for training, the same boundaries select the target coefficient subsequences.

Benchmark Inputs. The source dataset provides word and phoneme annotations, which we use for the selected data in place of automatically estimated linguistic annotations. Transcript segments are matched to the word intervals in its TextGrid files; the first and last matched words determine each segment's temporal boundaries. The same interval selects the audio, phonemes, and training coefficient subsequences. Phonemes come from the provided annotations in all three splits, including testing, and retain CMU ARPAbet lexical stress markers (0/1/2) on vowels. The segment transcript is constructed from the aligned words. Target coefficient values provide training supervision and are withheld from the generator at test time.

C Additional Implementation Details

C.1 Coefficient Representation and JSON Format

Temporal Resolution and Quantization. Following SAiD Park & Cho (2023), we retain 32 mouth-related ARKit channels. The preprocessing pipeline converts the original 60-Hz trajectories to 30 Hz by retaining every second frame within each aligned segment. For a coefficient $v_{t,k} \in [0, 1]$, the serialized integer is

$$q_{t,k} = \mathrm{round}(10 v_{t,k}) \in \{0, 1, \ldots, 10\}. \tag{9}$$

The implementation uses Python's nearest-integer rounding, including ties to even. Decoded integers are divided by 10 to recover coefficients on the original scale before animation post-processing.

Output Schema. Each assistant response is a JSON object with the 32 channel names below, serialized in this order. Each value is an integer array of length T_j, ordered chronologically at 30 Hz; every channel in a response has the same length. The target text uses compact JSON without extra spaces. These strings are processed by the pretrained model's tokenizer.

Index	Key	Index	Key
1	`JawForward`	17	`MouthStretchRight`
2	`JawRight`	18	`MouthRollLower`
3	`JawLeft`	19	`MouthRollUpper`
4	`JawOpen`	20	`MouthShrugLower`
5	`MouthClose`	21	`MouthShrugUpper`
6	`MouthFunnel`	22	`MouthPressLeft`
7	`MouthPucker`	23	`MouthPressRight`
8	`MouthRight`	24	`MouthLowerDownLeft`
9	`MouthLeft`	25	`MouthLowerDownRight`
10	`MouthSmileLeft`	26	`MouthUpperUpLeft`
11	`MouthSmileRight`	27	`MouthUpperUpRight`
12	`MouthFrownLeft`	28	`CheekPuff`
13	`MouthFrownRight`	29	`CheekSquintLeft`
14	`MouthDimpleLeft`	30	`CheekSquintRight`
15	`MouthDimpleRight`	31	`NoseSneerLeft`
16	`MouthStretchLeft`	32	`NoseSneerRight`

For illustration, a three-frame channel entry can be written as `"JawOpen":[2,4,3]`, corresponding to coefficients $(0.2, 0.4, 0.3)$. A complete response contains all 32 entries with the same array length.

Model Input and Training Record. Each JSONL training record contains a system message specifying the schema and articulation guidance, a user message with the audio placeholder, transcript, phoneme sequence, and requested frame count, and an assistant message containing the coefficient JSON as a string. The audio file is supplied through the record's audio field. The requested frame count defines the output duration, while the assistant coefficient sequence is the supervised target. The articulation guidance connects phoneme groups to named controls, for example bilabial consonants to lip closure and open vowels to jaw opening.

Coefficient-Value Weights. The SFT loss uses a fixed lookup table for integers in trajectory arrays:

$$\omega(q) = \begin{cases} 0.60, & q = 0, \\ 0.70, & q = 1, \\ 0.85, & q = 2, \\ 1.00, & q \in \{3, \ldots, 10\}. \end{cases} \tag{10}$$

These are manually specified weights that reduce the contribution of frequent low-valued coefficients. JSON keys and structural punctuation retain unit weight; the reweighting applies only to array-value text.

C.2 ARTICULATION GUIDANCE

The generator receives a fixed articulation-guidance block in its system message, together with the output-schema constraints described above. The selected-data configuration uses ARPAbet phonemes and retains vowel stress suffixes. The guidance is shared across inputs; the audio, transcript, phoneme sequence, and requested frame count are supplied separately for each unit.

The wording below is reproduced from the current system prompt. These are qualitative prompt heuristics: terms such as "strong," "moderate," and "subtle" do not specify fixed coefficient values or impose hard constraints on decoded motion. The detailed mapping is learned from the training targets.

Timing and Motion Instructions. The system message includes the following general instructions before the phoneme-specific block:

> Use the audio as the main timing source. Use the subtitle and phonemes as alignment hints.
>
> Higher values mean stronger activation. Neutral or silent frames should stay near zero. Left and right channels should usually be similar unless the audio suggests clear asymmetry.
>
> Use smooth frame to frame motion. Avoid sudden jumps. Let mouth shapes anticipate upcoming phonemes and relax naturally after them.

Verbatim Phoneme-to-Articulation Guidance.

> Learn the detailed mapping from training samples. Use the following English ARPAbet articulation rules as guidance.
>
> **English consonants:**
>
> Bilabials B P M use strong MouthClose, quick release, and moderate MouthPressLeft and MouthPressRight during closure.
>
> Labiodentals F V use strong MouthPressLeft and MouthPressRight, moderate MouthClose, and subtle MouthLowerDownLeft and MouthLowerDownRight.
>
> Dental fricatives TH DH use moderate MouthClose with subtle MouthUpperUpLeft and MouthUpperUpRight.
>
> Alveolar stops D T N use a brief MouthClose peak and quick release. N may add subtle MouthShrugUpper.
>
> Alveolar fricatives S Z use moderate MouthStretchLeft and MouthStretchRight, moderate MouthPressLeft and MouthPressRight, and moderate MouthClose.
>
> Alveolar liquid L uses slight MouthRollLower and moderate MouthClose.
>
> Alveolar approximant R uses moderate MouthFunnel, MouthPucker, and MouthRollUpper.
>
> Postalveolars SH ZH CH JH use stronger MouthFunnel and MouthPucker with moderate MouthRollUpper and JawOpen.
>
> Velars G K NG use moderate JawOpen and a brief MouthClose release while keeping lips mostly neutral.
>
> Glottal HH uses moderate JawOpen with relaxed lips and minimal MouthClose.
>
> Semivowel W uses strong MouthPucker and MouthFunnel. Semivowel Y uses MouthStretchLeft, MouthStretchRight, MouthSmileLeft, and MouthSmileRight.
>
> **English vowels:**
>
> Open vowels AA AE AH use stronger JawOpen, low MouthClose, and moderate MouthStretchLeft and MouthStretchRight. AE also uses moderate MouthSmileLeft and MouthSmileRight.
>
> Mid vowel EH uses moderate JawOpen and slight MouthStretchLeft and MouthStretchRight.
>
> Close front vowels IY IH use strong MouthStretchLeft and MouthStretchRight with moderate MouthSmileLeft and MouthSmileRight. JawOpen stays low.

Close back vowels UW UH use strong MouthPucker and MouthFunnel, supported by MouthRollUpper, MouthRollLower, and moderate JawOpen.

Mid back vowels AO OW use moderate MouthPucker, MouthFunnel, and JawOpen.

Diphthongs AY EY AW OW OY should transition smoothly from the first vowel shape to the second.

R colored vowel ER uses moderate MouthFunnel, MouthRollLower, MouthRollUpper, and JawOpen.

English stress:

Primary stress 1 should have stronger peaks and wider JawOpen.

Secondary stress 2 should be moderate and less intense than primary stress.

Unstressed 0 should have lower values, shorter motion, and more neutral mouth shapes.

Conditioning Ablation. The w/o Guidance condition removes the verbatim phoneme-to-articulation block above, from "Learn the detailed mapping from training samples" through the unstressed-vowel instruction. It retains the general timing and motion instructions, JSON constraints, final validity checks, and the transcript and phoneme inputs. Audio Only additionally removes the transcript and phoneme sequence from the user message while retaining the requested output length and audio input.

C.3 Model Adaptation and Training Configuration

We initialize the generator from Qwen3-Omni-30B-A3B-Instruct Xu et al. (2025) and train with LoRA Hu et al. (2021) using four NVIDIA A100 GPUs and BF16 precision. The audio encoder and multimodal alignment modules remain frozen in both stages. SFT adapts the decoder's linear layers. GRPO starts from the merged SFT generator and trains new LoRA adapters on the attention query/key/value and output projections. The two stages use the following settings; batch sizes are global across the training run, rather than per GPU.

Setting	SFT	GRPO
Epochs	4	4
LoRA rank / α	32 / 32	8 / 16
Micro-batch size	1	1
Global batch size	8	24
Learning rate	10^{-4}	10^{-6}
Minimum learning rate	10^{-5}	10^{-7}
Warmup fraction	0.03	0.05
Tensor parallel size	2	2
Context parallel size	2	1
Pipeline parallel size	1	1

SFT enables sequence packing with a maximum sequence length of 15,360 tokens. GRPO permits up to 4,096 prompt tokens and 12,400 completion tokens, uses expert parallel size 4, and clips the gradient norm at 0.5.

C.4 Sequence-Level Refinement

Sampling and Policy Updates. GRPO samples $G = 6$ candidate sequences per training prompt, with temperature 0.4, top-p $= 0.95$, and repetition penalty 1.0. The training launcher uses one policy-update iteration per sampled group, a policy-ratio clipping parameter of 0.2, and a reference-policy KL coefficient of 0.04. The KL coefficient controls regularization separately from the four motion-reward weights. Validation-time sampling uses four candidates per prompt.

Reward Normalization. The coefficients and scales in Eq. 7 are fixed during refinement:

Discrepancy	Weight λ_q	Scale s_q
Coefficient MSE	0.30	0.660
Velocity	0.20	0.0380
Acceleration	0.15	0.0746
SLMD	0.35	0.0285

The scales are stored reference-median values in the reward implementation. FD and WInD are not reward terms. Auxiliary diagnostic scores are logged with zero optimization weight.

Validity and Length Handling. A candidate must be a parseable JSON object with the expected channels, equally long nonempty arrays, and finite integer values in $[0, 10]$. The reward assigns -1.5 to unparseable outputs, -1.0 to schema violations, and -0.8 to invalid numeric values. For otherwise valid outputs, let $e_T = |T_p - T_g| / \max(T_g, 1)$, where T_p and T_g are the predicted and reference lengths. Candidates with $e_T > 0.20$ receive -0.6. For $e_T \leq 0.20$, motion discrepancies are computed on the shared prefix of length $\min(T_p, T_g)$, without interpolation, and the final reward is

$$r = r_{\text{motion}} - 0.5 \min(e_T, 1). \tag{11}$$

Thus, shorter or longer valid sequences incur a continuous penalty in addition to their motion error.

Fixed Geometry Asset. The SLMD asset is constructed from a neutral mesh and 32 ARKit blendshape meshes with identical topology. Each basis vector is the displacement from the neutral mesh. The supplied asset retains 606 vertices from a 1,220-vertex mesh, selected using the median of nonzero aggregate deformation magnitudes over the 32 controls. Fixed spatial weights emphasize the lip region: four nested ellipses in the neutral mesh's frontal plane assign weights of 1.0, 0.8, 0.6, and 0.45 from the innermost region outward; remaining selected vertices receive 0.1. Geometry is measured in the asset's native coordinate units. The basis and weights remain fixed throughout refinement.

C.5 Inference Post-Processing

Generated integer trajectories are decoded by dividing by 10 and assembled in temporal order on the recording timeline. For each coefficient channel, the post-processing pipeline applies deadzone suppression with threshold 0.02, Gaussian smoothing with standard deviation 1.0 frame and reflected boundaries, and Savitzky–Golay filtering with window length 7 and polynomial order 3. The Savitzky–Golay filter uses interpolated boundary handling and is applied only when at least seven frames are available. The final coefficients are clipped to $[0, 1]$.

C.6 Baseline Configuration and Visualization

Audio2Face-3D Chung et al. (2025), LAM-A2E He et al. (2025), and SentiAvatar Jin et al. (2026) use publicly released pretrained weights without additional fine-tuning on our dataset. SAiD Park & Cho (2023) is initialized from its public checkpoint and fine-tuned on our training split. The comparison uses the mouth-related ARKit channels and speaking intervals described in Section 4. Predicted coefficients are rendered with MetaHuman Epic Games (2025) for qualitative visualization.

D Training and Inference Procedure

Algorithm 1 summarizes the training and inference procedure described in Section 3.

E Additional Ablation Study

Ablation of Inference-Time Post-Processing. To distinguish the contribution of post-processing from that of learned motion refinement, we compare SFT and Refined (GRPO) under the reported settings with and without smoothing (Figure 6). Post-processing produces small improvements across the reported metrics for both generators, while refinement remains beneficial in both settings. The crossed comparison assesses whether the gains attributed to GRPO persist independently of smoothing. The filtering operations can suppress local fluctuations and segment-boundary discontinuities, which is consistent with their modest complementary benefit; GRPO instead changes the generator through motion-level feedback during training. These results support treating post-processing as a complement to motion-aware learning and show that the refinement gains persist without smoothing.

Algorithm 1 Training and Inference of AudioMouth

Require: Paired data $\mathcal{D}$, pretrained MLLM θ_0, articulation grounding $\mathcal{G}$, test speech $\mathcal{A}_{\text{test}}$
Ensure: Refined generator θ_{grpo} and predicted coefficients $\hat{\mathcal{V}}$

$\mathcal{U} \leftarrow \emptyset$ // Stage 1: semantic-aligned unit construction
for each $(\mathcal{A}, \mathcal{V}) \in \mathcal{D}$ **do**
  Obtain $\left(\tilde{\mathcal{W}}, \tau\right)$ and $\{\mathcal{W}^j\}_{j=1}^{J}$ by Eq.(3), and add units $(\mathcal{X}^j, \mathcal{V}^j, \mathcal{Y}^j)$ to $\mathcal{U}$ by Eq.(1)
end for
$\theta_{\text{sft}} \leftarrow \arg\min_\theta \mathcal{L}_{\text{db}}$ over $\mathcal{U}$ by Eq.(4) // Stage 2: grounded balanced generation
$\theta \leftarrow \theta_{\text{sft}}$ // Stage 3: motion-aware refinement
for each unit $(\mathcal{X}^j, \mathcal{V}^j) \in \mathcal{U}$ **do**
  Sample $\{\hat{\mathcal{Y}}_i^j\}_{i=1}^{G} \sim p_\theta\left(\cdot \mid \mathcal{X}^j\right)$ and parse each into a trajectory by $f_{\text{parse}}(\cdot)$
  Compute $\{r_i\}_{i=1}^{G}$ against $\mathcal{V}^j$ by Eq.(6) and Eq.(7), with validity and length handling
  $\hat{A}_i \leftarrow (r_i - \bar{r}) \,/\, (\sigma_r + \epsilon)$, and update θ with GRPO
end for
$\theta_{\text{grpo}} \leftarrow \theta$
Build $\{\mathcal{X}^j\}_{j=1}^{J}$ from $\mathcal{A}_{\text{test}}$, generate $\hat{\mathcal{Y}}^j$ with θ_{grpo}, and obtain $\hat{\mathcal{V}}$ by Eq.(8)
return θ_{grpo}, $\hat{\mathcal{V}}$

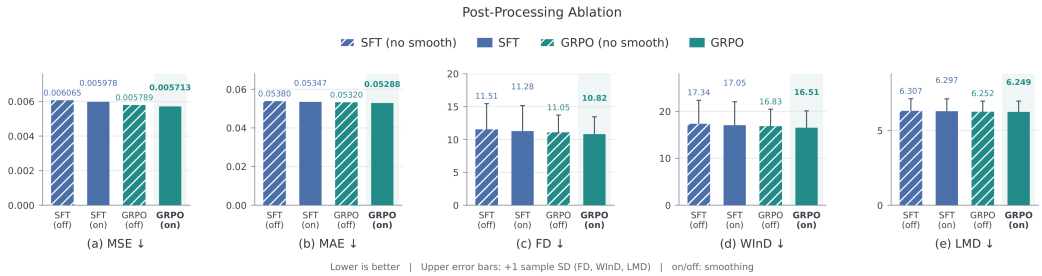

Figure 6: Inference-time post-processing for SFT and Refined (GRPO). The labels on/off denote the reported settings with and without smoothing. Results use 60 scripts from the selected test data. SFT is blue and GRPO is teal; hatching identifies the no-smooth variants. Lower values are better. Upper error bars show one sample standard deviation for FD, WInD, and LMD; MSE and MAE are aggregate values.

F Limitations and Future Work

Limitations. The experiments on the selected data use provided word and phoneme annotations at test time. Performance with automatically recognized phonemes remains to be established. The linguistic inputs and baseline adaptation settings differ across methods, so the benchmark compares their stated configurations rather than isolating architecture alone. The current scope is limited to 32 mouth-related ARKit coefficients and a fixed geometry basis for reward computation. Autoregressive JSON generation requires a long output sequence, and its latency has not been characterized for real-time deployment. SLMD remains a surrogate for mouth deformation: although the completed ablation supports its contribution to refinement, it does not establish equivalence to rendered-video LMD. The reported mean differences do not establish statistical significance, and some ablated conditions exhibit substantial variation across scripts. The conditioning ablation tests transcripts and phonemes jointly and does not isolate their individual effects or the shared segmentation stage. The post-processing comparison evaluates the reported pipeline settings rather than isolating each filtering operation. Moreover, improved facial animation techniques may also be misused for synthetic media generation, highlighting the importance of responsible deployment and appropriate safeguards.

Future Work. Leveraging the instruction-following ability of MLLMs, we plan to support explicit emotional control through natural language prompts, enabling richer and more expressive facial animations. We will also extend the framework to full-face coefficient generation, develop speaker-adaptive and few-shot personalization techniques, and improve robustness under in-the-wild, multilingual, and noisy conditions.