MOTION-OMNI: END-TO-END JOINT SPEECH AND FULL-BODY MOTION FOR SPOKEN DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

An avatar that holds a conversation should decide what to say and to move while saying it, yet these abilities live in separate model families: spoken dialogue models produce speech without motion, and co-speech motion models produce motion only from audio handed to them. The standard remedy is a cascade that first generates the spoken response and then runs a motion model over the finished audio, which requires a second full inference pass and precludes any joint optimisation between the two. We present **Motion-Omni**, an end-to-end framework in which a spoken dialogue model natively outputs explicit facial expression together with hand, upper-body and lower-body motion, generated directly from the hidden states that produce the speech. Joint training is not optional here: with the speech pathway frozen, motion remains misaligned with the audio, and co-adapting the LLM, Speech Generator and Motion Generator under both objectives is what recovers alignment while retaining spoken-dialogue ability. Supervision comes from a scalable, model-agnostic pipeline that pseudo-labels consistent-voice speech responses with a replaceable motion teacher, yielding $422{,}856$ quality-ranked pairs ($1{,}402$ hours). We further release SwDA-500 and, to our knowledge, the first public evaluation protocol for stochastic open-ended full-body spoken dialogue, matching audio across motion systems while unifying rendering, automatic metrics, human evaluation, and latency measurement. Instantiated with a Qwen2.5-7B-Instruct backbone, **Motion-Omni-Q7** matches the same-audio teacher cascade to within 2% on beat-correlation and lip-sync metrics while responding $5.4\times$ faster ($\mathrm{RTF} = 0.78$, faster than real time), surpasses all non-teacher cascades on beat correlation and diversity, and reaches a 2.62% word error rate, the lowest among the omni-modal systems compared. Code and video demonstrations are provided in the supplementary material.

1 INTRODUCTION

Speech and full-body co-speech motion, including facial expression and hand, upper-body, and lower-body movement, are tightly coupled in human communication: these movements carry meaning that complements and reinforces what is being said. We call a model that produces both a spoken response and this accompanying motion, conditioned on the same dialogue context, a *spoken motion model*. Figure 1 illustrates this interaction paradigm. Throughout this paper, *motion* denotes co-speech communicative motion; locomotion, dance, sports, and generic action generation are outside our scope.

Spoken dialogue models (SDMs) (OpenAI, 2024; Yang et al., 2024; Fang et al., 2025) and co-speech motion generation models (Liu et al., 2024; Yi et al., 2023; Chen et al., 2025a) have each progressed rapidly, and the direct way to obtain both outputs is to run them in sequence: an SDM produces the spoken response, and its audio is then passed to a motion model. This cascade does produce speech and motion, and it is the baseline we compare against throughout, but it has two structural costs: the motion model runs as a second full inference pass after the audio is complete, and no motion objective can ever update the speech or dialogue parameters. This paper asks whether both costs can be removed at once, that is, whether full-body motion can be a native output of an SDM, generated from the same states that produce the speech, without giving up motion quality or spoken-dialogue ability. Recent spoken motion models (Jiang et al., 2025; Deng et al., 2026; Zhang et al., 2026;

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.

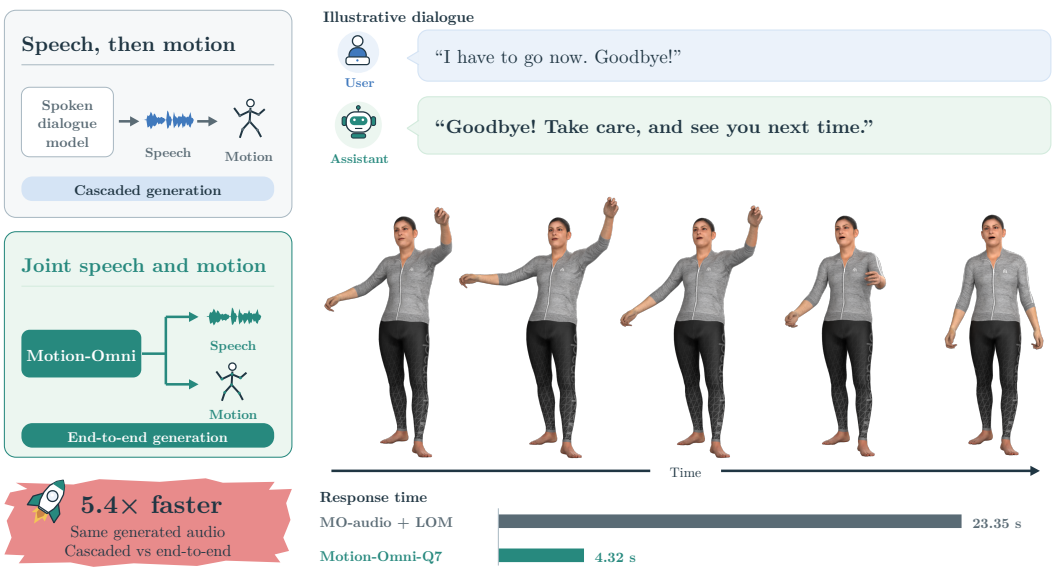


Figure 1: Motion-Omni jointly generates speech and full-body motion. The timing comparison uses the matched-audio, offline SwDA-500 benchmark (Table 4).

2025a; Mughal et al., 2026) relax parts of this recipe, but none natively combines facial expression, hand, upper-body, and lower-body output with co-adaptation of the response-generation pathway.

Building and evaluating such a model presents three challenges. First, making motion a native output of the SDM requires jointly optimising the speech and motion objectives without degrading either output. Joint optimisation enables reciprocal transfer through shared states: motion supervision can preserve motion-relevant timing and contextual cues in the states that generate speech, while speech supervision constrains those states to retain spoken-dialogue ability (Baltrusaitis et al., 2019). Realising this transfer is nontrivial, because the two outputs live at heterogeneous rates (12.5 Hz speech units and 30 Hz motion) and share parameters through which the two losses can interfere, so the architecture must bridge the rate mismatch and the training procedure must prevent motion gradients from degrading the dialogue pathway. Second, such joint training needs supervision at a scale no existing corpus provides. Many SDMs are trained towards a consistent target voice, so its motion supervision should be paired with responses in that same voice; captured audiovisual corpora record many speakers and voices (Agrawal et al., 2025) and cannot supply this at scale. We therefore adopt teacher pseudo-labeling, accepting the teacher as a quality ceiling in exchange for scale and voice consistency, which makes quality ranking of the generated supervision essential. Third, to our knowledge, no public benchmark is designed for stochastic open-ended full-body spoken dialogue. Existing full-body motion benchmarks assume fixed supplied speech (Kucherenko et al., 2024; Mughal et al., 2026), while spoken-dialogue evaluations with controlled generated speech cover facial animation only (Zhang et al., 2026); neither evaluates whether a model's own valid but unpredictable spoken response is accompanied by appropriate whole-body motion. Fair evaluation therefore requires open-ended prompts, matched audio across motion systems to prevent differences in response content and prosody from confounding motion comparisons, and complementary measurements of speech, motion, human perception, and latency.

In this work we propose **Motion-Omni**, an end-to-end spoken motion framework. To our knowledge, it is the first open-ended spoken dialogue model that natively generates explicit facial expression, hand, upper-body, and lower-body motion while allowing the motion objective to update both the LLM and a distinct Speech Generator. For the architecture challenge, our four-component framework (Speech Projector, LLM Backbone, Speech Generator, Part-Aware Motion Generator) bridges the rate mismatch through a dual-input conditioning interface: the Motion Generator attends to the Speech Generator's hidden states (key/value) and consumes the embeddings of emitted speech tokens (query), interpolated from the speech-unit rate to the motion rate, so motion is generated directly from the representations that produce the speech, without using the rendered waveform as an intermediate input. A four-stage training progression (Automatic Speech Recognition (ASR),

Text-to-Speech (TTS), TTS-with-Motion (TTSM) curriculum, joint mixture) co-adapts the Speech Projector, LLM, Speech Generator and Motion Generator end-to-end while keeping each as a distinct module and preserving spoken-dialogue ability. An ablation confirms that this co-adaptation is necessary: with the speech pathway frozen, generated motion remains visibly misaligned with the speech (Section 3.3). For the data challenge, we introduce a scalable, model-agnostic route from open-ended speech-instruction responses in a consistent target voice to explicit facial and full-body motion supervision: a replaceable motion teacher (LOM (Chen et al., 2025a) in this instantiation) labels every response waveform, and a dual-metric quality score induces the training curriculum. Applied to InstructS2S-200K (Fang et al., 2025), this route yields $422{,}856$ teacher-generated pseudo-labeled speech-motion samples ($1{,}402$ hours). For the evaluation challenge, we introduce and release SwDA-500 and a reproducible protocol for stochastic open-ended spoken motion: matched generated audio controls response content and prosody, while a shared renderer, reference-aware automatic metrics, paired human evaluation, and a common latency protocol cover the complete output. Instantiated as **Motion-Omni-Q7** (denoted as **MO**) with a Qwen2.5-7B-Instruct (Yang et al., 2024) backbone, the resulting model matches the same-audio teacher cascade to within 2% on beat-correlation and lip-sync metrics while completing speech-and-motion responses faster than real time ($\mathrm{RTF} = 0.78$), $5.4\times$ faster than that cascade; among the remaining systems, which do not run the motion teacher at inference time, it obtains the best beat correlation and diversity, and its speech reaches a 2.62% word error rate, the lowest among the omni-modal systems in our comparison.

In summary, our contributions are: (1) an end-to-end spoken motion framework in which motion is generated from the hidden states that produce the speech, with motion supervision jointly updating the LLM, Speech Generator and Motion Generator; ablations show this co-adaptation is necessary for speech-motion alignment, and the resulting model matches the same-audio teacher cascade on motion quality while removing its separate audio-to-motion stage and responding $5.4\times$ faster; (2) a scalable, model-agnostic route from consistent-voice speech instructions to quality-ranked facial and full-body motion supervision, yielding $422{,}856$ paired samples ($1{,}402$ hours); (3) SwDA-500 together with, to our knowledge, the first publicly released evaluation protocol for stochastic open-ended full-body spoken dialogue, matching audio across motion systems and unifying rendering, automatic metrics, human evaluation, and latency measurement.

2 Related Work

Motion-Omni sits at the intersection of four lines of work: (i) audio-conditioned co-speech motion generation; (ii) text-driven integrated speech-and-gesture synthesis; (iii) spoken dialogue models; and (iv) dialogue systems that emit both speech and articulated motion.

2.1 Motion Generation Models

Co-speech motion generation synthesises body motion from supplied speech audio (Nyatsanga et al., 2023). Audio-driven facial animation (FaceFormer (Fan et al., 2022)) produces face-only motion with periodic positional encodings that our Motion Generator inherits via Ex-Omni (Zhang et al., 2026). Full-body co-speech models (TalkShow (Yi et al., 2023), Listen, Denoise, Action (Alexanderson et al., 2023), EMAGE (Liu et al., 2024), MambaTalk (Xu et al., 2024b), GestureLSM (Liu et al., 2025), SemTalk (Zhang et al., 2025b)) extend the recipe to 3D motion using diffusion, discrete, state-space, or flow-matching representations; the Language of Motion (LOM) (Chen et al., 2025a) uses four part-specific VQ-VAE codebooks (face, hand, upper, lower) that we reuse as a frozen detokeniser and as the motion teacher in our data pipeline. These systems can be connected to synthetic speech in a cascade, but they do not themselves plan a dialogue response.

2.2 Integrated Speech and Gesture Synthesis

Joint optimisation lets objectives from related modalities shape shared representations, a central form of multimodal co-learning (Baltrusaitis et al., 2019). Integrated speech-and-gesture synthesis applies this principle to prescribed text scripts: Wang et al. (2021) adapted neural TTS architectures to jointly predict speech acoustics and 3D gesture from text. Diff-TTSG (Mehta et al., 2023) introduced probabilistic diffusion with parallel speech and gesture heads, while Match-TTSG (Mehta et al., 2024b) used one conditional-flow-matching decoder to model their joint distri-

bution. MAGI (Mehta et al., 2024a) added synthetic pre-training, multi-speaker support and prosody control. FastTalker (Guo & Zhang, 2024) reuses intermediate TTS timing and prosodic features for efficient full-body gesture decoding, while Gelina (Guichoux et al., 2026) autoregressively interleaves discrete speech and gesture tokens. Unlike these systems, which synthesise speech and gesture for a prescribed script, Motion-Omni takes a user turn as input and generates an open-ended spoken response together with its motion.

2.3 Spoken Dialogue Models

Spoken dialogue models (SDMs) equip LLMs with speech input and output. SpeechGPT (Zhang et al., 2023) helped popularise discrete audio tokens as a language-model vocabulary; LLaMA-Omni (Fang et al., 2025) adds a speech-unit decoder and the InstructS2S-200K dataset that we build on; GLM-4-Voice (Zeng et al., 2024) contributes the discrete 12.5 Hz speech tokenizer and CosyVoice (Du et al., 2024b)-style flow-matching decoder that MO adopts; Qwen2.5-Omni (Xu et al., 2025) introduces a Thinker-Talker architecture; and Moshi (Défossez et al., 2024) achieves full-duplex dialogue. More recently, WavAlign (Chen et al., 2026b) proposes modality-aware adaptive post-training for semantic quality and speech expressiveness, while AV-Dialog (Chen et al., 2026a) adds visual cues for target-speaker tracking and turn-taking in noisy multi-speaker settings. These standalone SDMs do not produce body motion.

2.4 Spoken Motion Models

The most directly related line conditions joint speech and motion output on an open-ended dialogue context. **SOLAMI** (Jiang et al., 2025) jointly predicts speech and body/hand motion tokens with an AnyGPT/LLaMA2-based autoregressive backbone, but generates facial animation post hoc with an audio-to-face model. **U-Mind** (Deng et al., 2026) uses one shared autoregressive backbone to generate response text, acoustic tokens, and SMPL-X pose tokens, but does not specify native facial-expression output or isolate how motion supervision affects speech-generation states. **Ex-Omni** (Zhang et al., 2026) jointly optimises an LLM, Speech Generator and 52-dimensional ARKit facial decoder, but addresses only the face. **ViBES** (Zhang et al., 2025a) combines a frozen GLM-4-Voice speech expert with trainable face and body experts; freezing preserves its speech model but prevents motion gradients from updating that component. **MIBURI** (Mughal et al., 2026) causally generates full-body gestures and facial expressions from Moshi's streaming states, but does not report motion-loss adaptation of Moshi. It addresses streaming interaction, whereas Motion-Omni synthesises a complete response offline, so their response times are not directly comparable. Motion-Omni uniquely combines explicit face, hand, upper- and lower-body output with end-to-end co-adaptation of distinct LLM, speech and motion modules, supported by a model-agnostic pseudo-labeling curriculum.

3 Method

3.1 System Overview

Motion-Omni comprises four components that process speech (or text) and jointly generate a spoken response and full-body motion (Figure 2). The framework fixes their interfaces and conditioning topology while allowing different model instantiations. Our Motion-Omni-Q7 instance uses a frozen Whisper-large-v3 (Radford et al., 2023) **Speech Encoder** and trainable projector, a Qwen2.5-7B-Instruct (Yang et al., 2024) **LLM Backbone**, and a Qwen2.5-0.5B-based **Speech Generator** emitting GLM-4-Voice (Zeng et al., 2024) units at 12.5 Hz. Token-as-Query Gated Fusion (TQGF) (Zhang et al., 2026) lets token embeddings query the LLM states through learned gates. The **Motion Generator** predicts four parallel LOM (Chen et al., 2025a) VQ-code streams at 30 Hz for face, hands, upper body and lower body. The decoders share speech context without cross-conditioning on one another's sampled outputs. Speech units are decoded to audio, and the frozen motion VQ-VAE produces SMPL-X/FLAME parameters; output decoding is detailed in Appendix A.1.

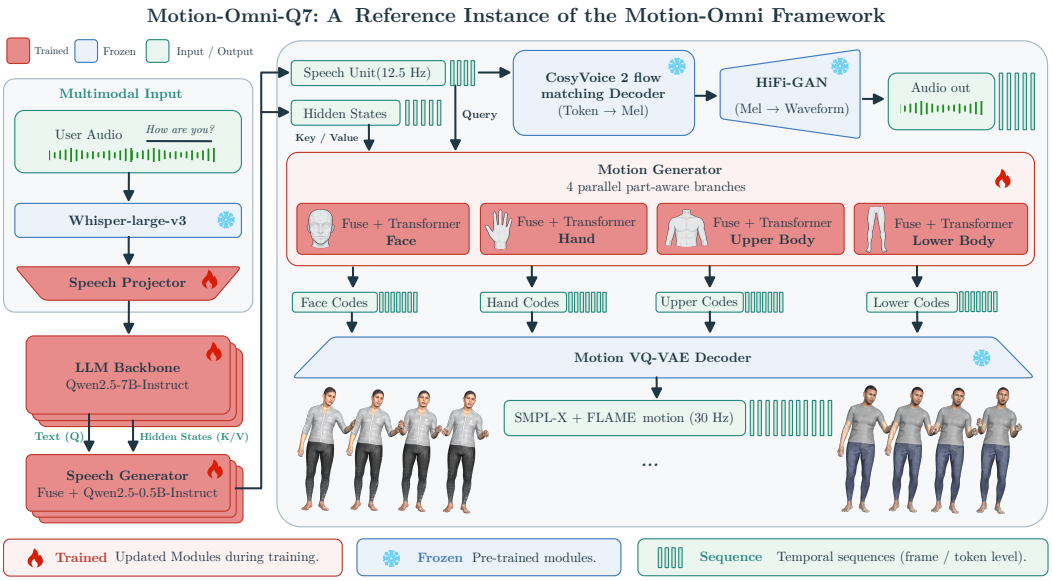


Figure 2: Motion-Omni framework with four components and their conditioning topology.

3.2 Motion Generator

Each body-part decoder stacks two TQGF layers at working dimension $d_m = 512$, with eight attention heads of dimension 64, followed by six causal self-attention layers with periodic rotary positional encoding of period 30 frames (Zhang et al., 2026). Its two conditioning streams are

$$\mathbf{Z} = \mathbf{H}_s \mathbf{W}_h, \tag{1}$$

$$\mathbf{Q} = \mathrm{Interp}\big(\mathbf{E}[\mathbf{u}]\mathbf{W}_e\big), \tag{2}$$

where $\mathbf{H}_s$ contains the Speech Generator's last-layer hidden states, $\mathbf{u}$ is the speech-unit sequence, $\mathbf{E}$ is initialised from the CosyVoice flow embedding (Du et al., 2024b), and $\mathbf{W}_h, \mathbf{W}_e$ are learned projections. Linear interpolation aligns the 12.5 Hz token queries with the 30 Hz motion rate. All four decoders read the same $\mathbf{Z}$ and $\mathbf{Q}$ with independent parameters. Training uses target speech units; inference uses the units just emitted by the Speech Generator.

The projected hidden states supply keys and values carrying acoustic timing and response semantics, while the queries identify the speech units and the gate selects context for each body part. This interface removes the separate waveform-to-motion inference pass and lets motion supervision update the speech-generation pathway. Each decoder predicts a 256-entry code distribution per frame, trained with

$$\mathcal{L}_{\text{motion}} = \sum_b w_b \, \mathcal{L}_{\text{CE}}^{(b)}, \qquad w_{\text{face}} : w_{\text{hand}} : w_{\text{upper}} : w_{\text{lower}} = 106 : 180 : 78 : 61, \tag{3}$$

where $\mathcal{L}_{\text{CE}}^{(b)}$ is the per-frame cross-entropy for part b with label smoothing 0.1. The weights are normalised to sum to one and are proportional to the underlying SMPL-X+FLAME feature dimensions, giving each VQ-code prediction equal per-feature-dimension importance. Appendix A.4 gives further details.

3.3 Progressive Training Curriculum

Training has four stages; the Whisper encoder and motion VQ-VAE remain frozen throughout. **Stage 1** trains the speech projector on ASR, and **Stage 2** trains the Speech Generator on TTS, with the LLM frozen in both. **Stage 3** jointly trains the Speech and Motion Generators on TTS-with-Motion (TTSM), keeping the LLM and projector frozen. Four warm-started substages progressively expand the training pool; Section 3.4 and Appendix B specify selection. In an initial pilot with the Speech Generator frozen, motion loss plateaued above the jointly trained model and rendered

motions remained misaligned with speech. Joint training lowered the loss and improved alignment, motivating co-adaptation throughout Stages 3a–3d. **Stage 4** jointly fine-tunes the LLM, projector and both generators on an ASR/TTS/Speech-to-Speech-with-Motion (S2SM)/T2T mixture to preserve dialogue ability. Input construction and optimisation details are in Appendix A.6.

3.4 Data Construction Pipeline

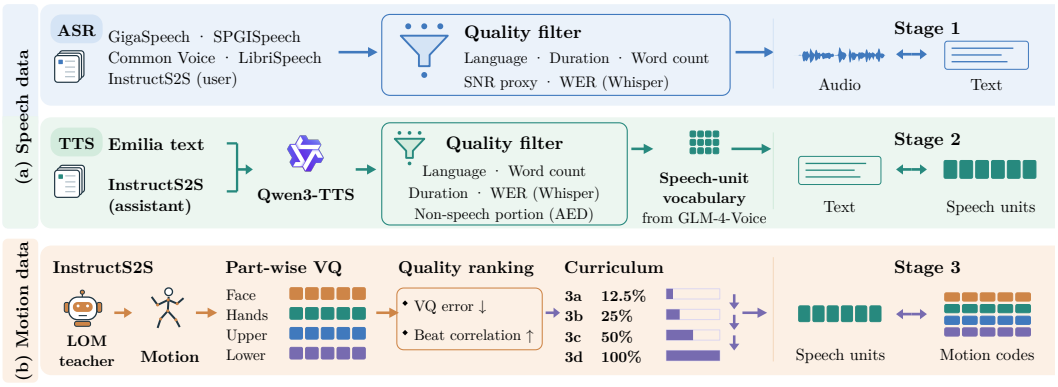


Figure 3: Speech-data curation and motion supervision. Insets schematically show the paired modalities: audio–text in Stage 1, text–speech units in Stage 2, and speech units–motion codes in Stage 3.

Figure 3 summarises speech-data preparation and the construction of motion supervision. Training a spoken motion model requires large-scale supervision that pairs a dialogue response with the motion accompanying it. Captured audiovisual resources such as the dyadic Seamless Interaction corpus (Agrawal et al., 2025) record conversations between speakers, but their many voices are not matched to the single target voice our Speech Generator is trained to produce. We therefore construct supervision by pseudo-labeling: a *generic* pipeline takes a speech-instruction corpus already paired with response audio together with a co-speech motion generation teacher, and produces paired (text, speech, full-body motion) supervision in two steps.

(1) Motion supervision via a teacher: a pre-trained co-speech motion generation teacher (LOM (Chen et al., 2025a) in our run) is run on every response waveform; its four per-part VQ code streams form the motion target. The teacher is a replaceable component, so a future implementation can regenerate supervision with a stronger teacher. **(2) Dual-metric quality scoring and curriculum**: every sample is scored along (i) a *weighted VQ-VAE reconstruction error* L_1^{vq} that flags motion outside the codebook's expressive range, and (ii) a *beat correlation* (BC) score (Li et al., 2021b) that flags teacher motion weakly coupled to the speech; after robust 5th–95th percentile normalisation each sample receives a combined score $s(x) = \alpha(1 - \widetilde{L_1^{\text{vq}}}) + (1-\alpha)\widetilde{\text{BC}}$ with $\alpha = 0.5$, which drives the four-substage Stage 3 curriculum and selects the Stage 4 S2SM pool. The exact reconstruction-error computation is given in Appendix B.4.

Speech-data filtering. ASR clips must last 1–30 s and satisfy Whisper-large-v3 WER $\leq 5\%$. For all-uppercase labels, we adopt Whisper's capitalisation to avoid teaching all-caps output. We require matching normalised words (WER $= 0$) to avoid introducing recognition errors; otherwise the sample is discarded. TTS clips require 1–25 s, 2–200 words and WER $< 5\%$. Both tasks reject non-English characters and texts shorter than two words; Appendix B details SNR and AED checks.

Reference instantiation. Stage 1 combines GigaSpeech (Chen et al., 2021), SPGISpeech (O'Neill et al., 2021), Common Voice (Ardila et al., 2020), InstructS2S-200K (Fang et al., 2025), and LibriSpeech (Panayotov et al., 2015). For Stage 2, Qwen3-TTS (Team, 2026) synthesises speech separately for each text sample from English Emilia (He et al., 2024) or InstructS2S assistant responses. Filtering retains 4,156,067 ASR records and 1,043,309 TTS records, each containing text and its corresponding speech (Appendix B.1). Stage 3 supplies 422,856 motion-supervised samples (1,402 hours). Unranked full-corpus training diverged; teacher-reference FGD falls from 0.3974 at Stage 3a to 0.3040 after Stage 4 (Appendix G). Stage 4 includes text-only data; omitting it increased repetition of the user's input.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

For speech and motion evaluation we use *SwDA-500*, an external 500-prompt dialogue-text evaluation set derived from the Switchboard Dialog Act Corpus (SwDA) (Stolcke et al., 2000). The set covers all 66 SwDA topic descriptions with 7 or 8 semantically complete speaker turns per topic, keeps prompts of moderate length, and removes transcription artifacts not intended to be spoken. SwDA-500 provides real conversational wording, but it is not a paired real-motion benchmark; all motion references below are teacher-generated or baseline-generated under matched prompts.

4.2 SPOKEN DIALOGUE QUALITY

The remaining evaluation assumes that the backbone can still hold a spoken conversation after joint motion training. We verify this on VoiceBench (Chen et al., 2026c), a nine-subset benchmark covering open-ended dialogue, factual QA, reasoning, instruction-following and safety, where Motion-Omni-Q7 reaches an Overall of 47.63, above LLaMA-Omni (41.12), Ex-Omni (43.57), Mini-Omni2 (33.49) and Moshi (29.51). Appendix E reports all per-subset scores.

Table 1: WER (%) on Seed-TTS-Eval.

Model	test-en↓
Human Reference	
Human	2.14
Dedicated TTS	
CosyVoice (Du et al., 2024a)	4.29
CosyVoice 2 (Du et al., 2024b)	2.57
FireRedTTS (Guo et al., 2024)	3.82
Qwen3-TTS-12Hz-1.7B (Team, 2026)	**1.24**
Qwen3-TTS-12Hz-0.6B (Team, 2026)	<u>1.32</u>
CosyVoice 3-1.5B RL (Du et al., 2025)	1.45
F5-TTS (Chen et al., 2025b)	2.04
IndexTTS2 (Zhou et al., 2026)	2.18
Omni-modal LLMs	
Qwen2.5-Omni (Xu et al., 2025)	2.72
Ex-Omni (Zhang et al., 2026)	<u>2.67</u>
Motion-Omni-Q7 (ours)	**2.62**

4.3 TEXT-TO-SPEECH (SEED-TTS-EVAL)

Setting. We evaluate speech intelligibility on Seed-TTS-Eval (Anastassiou et al., 2024) (English split, 1,088 samples), reporting corpus-level Word Error Rate (WER) computed by Whisper-large-v3 (Radford et al., 2023) on the model's synthesised speech against the ground-truth transcript.

Results. Table 1 reports WER. In the table, lower is better. **Bold** marks the best value and <u>underline</u> the second best *within each system category*, since dedicated TTS systems synthesise a supplied sentence whereas omni-modal LLMs also plan the response; the Human row is a reference and is not ranked. Motion-Omni-Q7 reaches 2.62% on `test-en`, the lowest among the omni-modal LLMs, while jointly producing speech and SMPL-X/FLAME co-speech motion; several recent dedicated TTS systems report lower WER on this speech-only metric.

Speech naturalness proxy. On SwDA-500 generated audio, Motion-Omni-Q7 scores 3.77 on UTMOSv2 (Baba et al., 2024), an automatic speech-naturalness proxy that does not replace a listening test. Appendix F gives full comparisons.

4.4 SPEECH-TO-MOTION

We compare MO against controlled cascades that combine either MO audio or Qwen2.5-Omni (Xu et al., 2025) audio with recent or previously used audio-to-motion generators: MambaTalk (Xu et al., 2024b), GestureLSM (Liu et al., 2025), EMAGE (Liu et al., 2024), and LOM (Chen et al., 2025a). Throughout, a cascade is named *speech source* + *audio-to-motion model*, in the order the two run. Rows using MO audio isolate the motion-generation pathway under the same speech as our model; rows using Qwen2.5-Omni audio test the same motion generators under a stronger external speech model. LOM rows are teacher-reference cascades because LOM also supplies the reference distribution for teacher-based FGD.

4.4.1 AUTOMATIC EVALUATION

Setting. Following LOM (Chen et al., 2025a), we report FGD (Yoon et al., 2020), which compares the distribution of generated motion features against a reference distribution, together with Diversity (Li et al., 2021a) and BC (Li et al., 2021b). Because MO generates speech and motion

Table 2: Motion, facial geometry and lip-sync metrics on SwDA-500. Teacher-referenced metrics compare against LOM-generated motion and thus favour LOM-based cascades; for FGD, BC and BC_{bi}, displayed values = raw values $\times 10$. **Bold**/underline mark the best/second-best; dashes denote values not computed.

	Reference-free					Teacher-referenced		
System	**Div.↑**	**BC↑**	**BC_{bi} ↑**	**LSE-C↑**	**LSE-D↓**	**FGD↓**	**MSE↓**	**LVD↓**
Qwen2.5-Omni + MambaTalk	11.92	7.17	8.56	6.853	7.920	4.08	7.092	6.851
MO-audio + MambaTalk	12.07	7.03	8.60	6.797	7.955	4.15	7.128	7.019
Qwen2.5-Omni + GestureLSM	13.61	7.44	9.04	–	–	3.29	–	–
MO-audio + GestureLSM	13.55	7.49	9.11	–	–	3.17	–	–
Qwen2.5-Omni + EMAGE	11.67	7.31	8.66	6.578	8.151	3.75	7.553	7.205
MO-audio + EMAGE	10.95	7.32	8.73	6.643	8.094	3.29	7.627	7.291
Qwen2.5-Omni + LOM	**13.82**	7.56	**9.13**	**7.136**	7.128	–	**6.581**	**6.243**
MO-audio + LOM	13.07	**7.67**	9.05	6.985	**7.011**	–	–	–
Motion-Omni-Q7 (ours)	13.67	7.59	9.07	7.011	7.023	**3.04**	6.723	6.427

Table 3: Human preference on motion-related dimensions. **Bold**/underline mark the largest/second-largest pooled wins-minus-losses margin.

Comparison	**R1 Rhythm**	**R2 Semantics**	**R3 Naturalness**	**Pooled W/T/L (W–L)**
Qwen2.5-Omni + LOM	12 3 10	13 12	10 3 12	35/6/34 (+1)
Qwen2.5-Omni + EMAGE	7 2 16	10 3 12	16 4 5	33/9/33 (0)
Same MO speech audio:				
MO-audio + LOM	5 12 8	9 8 8	11 7 7	25/27/23 (+2)
MO-audio + EMAGE	13 7 5	15 3 7	17 8	**45/10/20 (+25)**

MO wins ties MO losses

jointly, we cannot fix its audio output to match a standard benchmark corpus (e.g. BEAT2); instead we use LOM's predictions on the *same* generated audio as the FGD reference. This is a teacher-reference FGD: it measures fidelity to the LOM-generated teacher distribution. We omit FGD for LOM rows because their motion supplies the reference distribution. The exploratory BC_{bi} is defined and analysed in Appendix B.5.

Facial geometry and lip synchronisation. MSE and landmark velocity difference (LVD) measure fidelity to LOM-generated facial geometry, not captured ground truth. Rendered-video SyncNet scores (LSE-C/LSE-D) (Prajwal et al., 2020) evaluate audio–lip synchronisation without a motion reference. All systems share the renderer and preprocessing; metric definitions and settings are in Appendix C.1.

Results. Table 2 compares all systems on SwDA-500 under matched prompts and a single shared metric pipeline. Motion-Omni-Q7 places first or second on seven of the eight metrics. Among systems that do not run LOM at motion-inference time it obtains the highest beat correlation (7.59), the highest diversity (13.67) and the best score on every facial and lip-sync metric; the rows that surpass it are the LOM teacher-reference cascades, which invoke at motion-inference time the same model that supplied its training targets. It also obtains the lowest teacher-reference FGD, which by construction measures distance to the LOM-generated distribution.

The integrated model matches strong cascades in motion quality while avoiding a separate audio-to-motion inference stage, and Section 4.5 quantifies what that stage costs.

4.4.2 Human Evaluation

Setting. Four trained annotators independently rated 25 paired clips per comparison arm, using hidden system names and randomised A/B order. Annotator characteristics, instructions and the full protocol are in Appendix D. Each pair was rated on three motion-related dimensions: **R1** rhythm, **R2** semantic alignment, and **R3** body naturalness, judged from rendered video+audio. Per-pair verdicts were determined by majority vote.

Table 4: Response latency on SwDA-500. **Bold**/underline mark the best/second-best value; Slowdown is relative to our model.

Model	T_{resp}(s)↓	RTF↓	Slowdown
Qwen2.5-Omni + LOM	99.08	18.34	$22.9\times$
Qwen2.5-Omni + EMAGE	48.65	8.79	$11.3\times$
MO-audio + LOM	23.35	4.39	$5.4\times$
MO-audio + EMAGE	<u>4.63</u>	<u>0.84</u>	$1.1\times$
Motion-Omni-Q7 (ours)	**4.32**	**0.78**	$1.0\times$

Results. Table 3 reports win/tie/loss for Motion-Omni-Q7. The two MO-audio arms hold speech fixed, so differences there reflect the motion pathway alone. Against MO-audio + EMAGE the pooled margin is $+25$ (45 wins, 10 ties, 20 losses). Against MO-audio + LOM, that is, against the teacher itself, the outcome is $25/27/23$, with ties the single most frequent verdict. With four annotators and 25 pairs per arm, we do not read the smaller margins as evidence of a reliable preference.

The same-audio comparisons suggest that Motion-Omni-Q7 can produce motion preferred over EMAGE and broadly comparable to LOM in this limited setting; Appendix H shows representative outputs and matched-audio comparisons.

4.5 Latency

Setting. We measure total response time T_{resp} (wall-clock seconds from receiving the user audio to completing all audio and motion generation) and the real-time factor $\text{RTF} = T_{\text{resp}}/T_{\text{user}}$, where T_{user} is the user-utterance duration; $\text{RTF} < 1$ means the system responds faster than real time. All measurements use the same SwDA-500 evaluation set on a single GPU. These measurements evaluate offline response generation.

Results. Table 4 shows that Motion-Omni-Q7 completes a full response (speech + motion) in $4.32\,\text{s}$ ($\text{RTF} = 0.78$), faster than real time. The matched comparison is MO-audio + LOM, which uses the same speech model and the teacher motion model but invokes the latter as a separate audio-to-motion stage: at $23.35\,\text{s}$ it is $5.4\times$ slower.

Appendix G.3 compares motion quality and latency jointly.

4.6 Controlled Fusion Ablation

All variants start from the same Stage 2 checkpoint and use the same Stage 3a data and optimisation schedule, changing only how the motion query incorporates context. TQGF with Speech Generator states achieves BC 0.912, versus 0.827 with LLM states and 0.905 for plain cross-attention with Speech Generator states. These comparisons support speech-generation context and a modest TQGF advantage under this fixed setup; they do not isolate every internal TQGF component. Appendix G reports all six variants.

5 Conclusion

We have presented Motion-Omni, an end-to-end framework that enables a conversational LLM to jointly generate intelligible speech and synchronised full-body co-speech motion, together with a model-agnostic pseudo-labeling pipeline that supplies $1{,}402$ hours of teacher-generated supervision and a reproducible evaluation setup for this task class. Across the reported benchmarks, the reference instance Motion-Omni-Q7 records the lowest word error rate among the omni-modal systems compared, and on SwDA-500 the highest beat correlation and diversity among systems that do not run the motion teacher at inference time, along with better rendered-video lip synchronisation than the evaluated EMAGE and MambaTalk cascades. Conditioning motion on the Speech Generator's hidden states rather than on a decoded waveform keeps beat-correlation and lip-sync metrics within 2% of the same-audio teacher cascade while removing a separate audio-to-motion stage, which accounts for the $5.4\times$ difference in response time. All components are swappable; additional instances follow from re-running the recipe with stronger backbones or teachers.

REPRODUCIBILITY STATEMENT

The architecture and its conditioning topology are specified in Section 3, and the four training stages are described in Section 3.3, with per-stage hyperparameters, optimiser settings and numerical-stability measures given in Appendix A.6. The data construction pipeline, including the teacher model, the two quality metrics and the curriculum quantiles, is described in Section 3.4, with the exact definitions of the weighted VQ-VAE reconstruction error and the beat correlation score in Appendix B.4 and Appendix B.5; all source corpora are public. The evaluation setup is described in Section 4.1: SwDA-500 is derived from the public Switchboard Dialog Act Corpus by the filtering procedure stated there, cascade baselines are driven by audio generated by our own model so that speech is held fixed across systems, and every system is rendered through the shared pipeline detailed in Appendix C before lip-sync scoring. The human evaluation protocol, annotator instructions and rubric are given in Appendix D and Appendix D.1.

ETHICS STATEMENT

This work involves a small human preference study. Annotators rated pairs of rendered animation clips with anonymised system labels and randomised presentation order; the protocol is described in Appendix D. No personally identifying information was collected from annotators; the paper reports only aggregate demographic characteristics of the evaluator group. The evaluation prompts are drawn from the public Switchboard Dialog Act Corpus rather than from newly recorded human subjects. The model generates a synthetic speaking avatar with a single fixed voice and a generic body mesh; it is not designed to imitate a specific person's voice, face or motion style, and it does not condition on speaker identity. Training data are entirely in English, so the learned motion style reflects the conversational register of that data and should not be assumed to transfer across languages or cultures.

AI USE STATEMENT

We used generative AI tools as writing assistants: to polish phrasing, tighten wording and check the internal consistency of tables and cross-references in the manuscript. Research ideation, the system design, the implementation, the experiments and the analysis of results were carried out by the authors. AI tools were also used for routine coding assistance during development; all such code was reviewed and tested by the authors. We have reviewed all AI-assisted content and take full responsibility for the final content of this work, including its text, claims and artifacts.

REFERENCES

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeffrey M. Girard, Martin Gleize, Sahir Gomez, Hongyu Gong, Srivathsan Govindarajan, Brandon Han, Sen He, Denise Hernandez, Yordan Hristov, Rongjie Huang, Hirofumi Inaguma, Somya Jain, Raj Janardhan, Qingyao Jia, Christopher Klaiber, Dejan Kovachev, Moneish Kumar, Hang Li, Yilei Li, Pavel Litvin, Wei Liu, Guangyao Ma, Jing Ma, Martin Ma, Xutai Ma, Lucas Mantovani, Sagar Miglani, Sreyas Mohan, Louis-Philippe Morency, Evonne Ng, Kam-Woh Ng, Tu Anh Nguyen, Amia Oberai, Benjamin N. Peloquin, Juan Pino, Jovan Popovic, Omid Poursaeed, Fabian Prada, Alice Rakotoarison, Rakesh Ranjan, Alexander Richard, Christophe Ropers, Safiyyah Saleem, Vasu Sharma, Alex Shcherbyna, Jia Shen, Jie Shen, Anastasis Stathopoulos, Anna Y. Sun, Paden Tomasello, Tuan Tran, Arina Turkatenko, Bo Wan, Chao Wang, Jeff Wang, Mary Williamson, Carleigh Wood, Tao Xiang, Yilin Yang, Julien Yao, Chen Zhang, Jiemin Zhang, Xinyue Zhang, Jason Zheng, Pavlo Zhyzheria, Jan Zikes, and Michael Zollhöfer. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *CoRR*, abs/2506.22554, 2025. doi: 10.48550/ARXIV.2506.22554. URL https://doi.org/10.48550/arXiv.2506.22554.

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.*, 42(4):44:1–44:20, 2023. doi: 10.1145/3592458. URL https://doi.org/10.1145/3592458.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models. *CoRR*, abs/2406.02430, 2024. doi: 10.48550/ARXIV.2406.02430. URL https://doi.org/10.48550/arXiv.2406.02430.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4218–4222, 2020.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824. IEEE, 2024.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607. URL https://doi.org/10.1109/TPAMI.2018.2798607.

Changan Chen, Juze Zhang, Shrinidhi K. Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 6200–6211. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.00581. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_The_Language_of_Motion_Unifying_Verbal_and_Non-verbal_Language_of_CVPR_2025_paper.html.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek (eds.), *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pp. 3670–3674. ISCA, 2021. doi: 10.21437/INTERSPEECH.2021-1965. URL https://doi.org/10.21437/Interspeech.2021-1965.

Tuochao Chen, Bandhav Veluri, Hongyu Gong, and Shyamnath Gollakota. Av-dialog: Spoken dialogue models with audio-visual input. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 42208–42225. Association for Computational Linguistics, 2026a. doi: 10.18653/V1/2026.ACL-LONG.1954. URL https://doi.org/10.18653/v1/2026.acl-long.1954.

Yifu Chen, Shengpeng Ji, Qian Chen, Tianle Liang, Yangzhuo Li, Ziqing Wang, Wen Wang, Jingyu Lu, Haoxiao Wang, Xueyi Pu, Fan Zhuo, and Zhou Zhao. Wavalign: Enhancing intelligence and expressiveness in spoken dialogue models via adaptive hybrid post-training. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 2421–2444. Association for Computational Linguistics, 2026b. doi: 10.18653/V1/2026.FINDINGS-ACL.114. URL https://doi.org/10.18653/v1/2026.findings-acl.114.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *Trans. Assoc. Comput. Linguistics*, 14:378–398, 2026c. doi: 10.1162/TACL.A.628. URL https://doi.org/10.1162/tacl.a.628.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 6255–6271. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.ACL-LONG.313. URL https://doi.org/10.18653/v1/2025.acl-long.313.

Jiayan Cui, Zhihan Yang, Naihan Li, Jiankun Tian, Xingyu Ma, Yi Zhang, Guangyu Chen, Runxuan Yang, Yuqing Cheng, Yizhi Zhou, Guochen Yu, Xiaotao Gu, and Jie Tang. GLM-TTS technical report. *CoRR*, abs/2512.14291, 2025. doi: 10.48550/ARXIV.2512.14291. URL https://doi.org/10.48550/arXiv.2512.14291.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *CoRR*, abs/2410.00037, 2024. doi: 10.48550/ARXIV.2410.00037. URL https://doi.org/10.48550/arXiv.2410.00037.

Xiang Deng, Feng Gao, Yong Zhang, Youxin Pang, Xu Xiaoming, Zhuoliang Kang, Xiaoming Wei, and Yebin Liu. U-mind: A unified framework for real-time multimodal interaction with audiovisual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10874–10886, June 2026.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *CoRR*, abs/2407.05407, 2024a. doi: 10.48550/ARXIV.2407.05407. URL https://doi.org/10.48550/arXiv.2407.05407.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117, 2024b. doi: 10.48550/ARXIV.2412.10117. URL https://doi.org/10.48550/arXiv.2412.10117.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *CoRR*, abs/2505.17589, 2025. doi: 10.48550/ARXIV.2505.17589. URL https://doi.org/10.48550/arXiv.2505.17589.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 18749–18758. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01821. URL https://doi.org/10.1109/CVPR52688.2022.01821.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=PYmrUQmMEw.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yifan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. VITA-1.5: towards gpt-4o level real-time vision and speech interaction. *CoRR*, abs/2501.01957, 2025. doi: 10.48550/ARXIV.2501.01957. URL https://doi.org/10.48550/arXiv.2501.01957.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL https://arxiv.org/abs/2506.04178.

Téo Guichoux, Théodor Lemerle, Shivam Mehta, Jonas Beskow, Gustav Eje Henter, Laure Soulier, Catherine Pelachaud, and Nicolas Obin. Gelina: Unified speech and gesture synthesis via interleaved token prediction. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16122–16126, 2026. doi: 10.1109/ICASSP55912.2026.11464562.

Haohan Guo, Kun Liu, Feiyu Shen, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kaituo Xu. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *CoRR*, abs/2409.03283, 2024. doi: 10.48550/ARXIV.2409.03283. URL https://doi.org/10.48550/arXiv.2409.03283.

Zixin Guo and Jian Zhang. Fasttalker: Jointly generating speech and conversational gestures from text. In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi (eds.), *Computer Vision - ECCV 2024 Workshops - Milan, Italy, September 29-October 4, 2024, Proceedings, Part XIX*, volume 15641 of *Lecture Notes in Computer Science*, pp. 177–194. Springer, 2024. doi: 10.1007/978-3-031-93806-1_14. URL https://doi.org/10.1007/978-3-031-93806-1_14.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pp. 885–890. IEEE, 2024. doi: 10.1109/SLT61566.2024.10832365. URL https://doi.org/10.1109/SLT61566.2024.10832365.

Jianping Jiang, Weiye Xiao, Zhengyu Lin, Huaizhong Zhang, Tianxiang Ren, Yang Gao, Zhiqian Lin, Zhongang Cai, Lei Yang, and Ziwei Liu. SOLAMI: social vision-language-action modeling for immersive interaction with 3d autonomous characters. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 26887–26898. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02504. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jiang_SOLAMI_Social_Vision-Language-Action_Modeling_for_Immersive_Interaction_with_3D_Autonomous_CVPR_2025_paper.html.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. SODA: million-scale dialogue distillation with social commonsense contextualization. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 12930–12949. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.799. URL https://doi.org/10.18653/v1/2023.emnlp-main.799.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report. *CoRR*, abs/2504.18425, 2025. doi: 10.48550/ARXIV.2504.18425. URL https://doi.org/10.48550/arXiv.2504.18425.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html.

Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. Evaluating gesture generation in a large-scale open challenge: The GENEA challenge 2022. *ACM Trans. Graph.*, 43(3):32:1–32:28, 2024. doi: 10.1145/3656374. URL https://doi.org/10.1145/3656374.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124, 2024. doi: 10.48550/ARXIV.2411.15124. URL https://doi.org/10.48550/arXiv.2411.15124.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf, 2024.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 11273–11282. IEEE, 2021a. doi: 10.1109/ICCV48922.2021.01110. URL https://doi.org/10.1109/ICCV48922.2021.01110.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3d dance generation with AIST++. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 13381–13392. IEEE, 2021b. doi: 10.1109/ICCV48922.2021.01315. URL https://doi.org/10.1109/ICCV48922.2021.01315.

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):194:1–194:17, 2017. doi: 10.1145/3130800.3130813. URL https://doi.org/10.1145/3130800.3130813.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. EMAGE: towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 1144–1154. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00115. URL https://doi.org/10.1109/CVPR52733.2024.00115.

Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 10929–10939. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01017. URL https://doi.org/10.1109/ICCV51701.2025.01017.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arik. Vista: A test-time self-improving video generation agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6021–6032, June 2026.

Chengqian Ma, Wei Tao, and Steven Y. Guo. C3: A bilingual benchmark for spoken dialogue models exploring challenges in complex conversations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou,*

China, November 4-9, 2025, pp. 22778–22796. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1160. URL https://doi.org/10.18653/v1/2025.emnlp-main.1160.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In Kathryn D. Huff and James Bergstra (eds.), *Proceedings of the 14th Python in Science Conference, SciPy 2015, Austin, Texas, USA, July 6-12, 2015*, pp. 18–24. scipy.org, 2015. doi: 10.25080/MAJORA-7B98E3ED-003. URL https://doi.org/10.25080/Majora-7b98e3ed-003.

Shivam Mehta, Siyang Wang, Simon Alexanderson, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Diff-ttsg: Denoising probabilistic integrated speech and gesture synthesis. In Gérard Bailly, Thomas Hueber, Damien Lolive, Nicolas Obin, and Olivier Perrotin (eds.), *12th ISCA Speech Synthesis Workshop, SSW 2023, Grenoble, France, August 26-28, 2023*, pp. 150–156. ISCA, 2023. doi: 10.21437/SSW.2023-24. URL https://doi.org/10.21437/SSW.2023-24.

Shivam Mehta, Anna Deichler, Jim O'Regan, Birger Moëll, Jonas Beskow, Gustav Eje Henter, and Simon Alexanderson. Fake it to make it: Using synthetic data to remedy the data shortage in joint multimodal speech-and-gesture synthesis. *CoRR*, abs/2404.19622, 2024a. doi: 10.48550/ARXIV.2404.19622. URL https://doi.org/10.48550/arXiv.2404.19622.

Shivam Mehta, Ruibo Tu, Simon Alexanderson, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Unified speech and gesture synthesis using flow matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 8220–8224. IEEE, 2024b. doi: 10.1109/ICASSP48485.2024.10445998. URL https://doi.org/10.1109/ICASSP48485.2024.10445998.

Muhammad Hamza Mughal, Rishabh Dabral, Vera Demberg, and Christian Theobalt. MIBURI: towards expressive interactive gesture synthesis. *CoRR*, abs/2603.03282, 2026. doi: 10.48550/ARXIV.2603.03282. URL https://doi.org/10.48550/arXiv.2603.03282.

Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. *Comput. Graph. Forum*, 42(2):569–596, 2023. doi: 10.1111/CGF.14776. URL https://doi.org/10.1111/cgf.14776.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. Spgispeech: 5, 000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek (eds.), *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pp. 1434–1438. ISCA, 2021. doi: 10.21437/INTERSPEECH.2021-1860. URL https://doi.org/10.21437/Interspeech.2021-1860.

OpenAI. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024. doi: 10.48550/ARXIV.2410.21276. URL https://doi.org/10.48550/arXiv.2410.21276.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pp. 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 10975–10985. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01123. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Pavlakos_Expressive_Body_Capture_3D_Hands_Face_and_Body_From_a_CVPR_2019_paper.html.

K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (eds.), *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pp. 484–492. ACM, 2020. doi: 10.1145/3394171.3413532. URL https://doi.org/10.1145/3394171.3413532.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 3505–3506. ACM, 2020. doi: 10.1145/3394486.3406703. URL https://doi.org/10.1145/3394486.3406703.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *CoRR*, cs.CL/0006023, 2000. URL https://arxiv.org/abs/cs/0006023.

Qwen Team. Qwen3-tts technical report. *CoRR*, abs/2601.15621, 2026. doi: 10.48550/ARXIV.2601.15621. URL https://doi.org/10.48550/arXiv.2601.15621.

Siyang Wang, Simon Alexanderson, Joakim Gustafson, Jonas Beskow, Gustav Eje Henter, and Éva Székely. Integrated speech and gesture synthesis. In Zakia Hammal, Carlos Busso, Catherine Pelachaud, Sharon L. Oviatt, Albert Ali Salah, and Guoying Zhao (eds.), *ICMI '21: International Conference on Multimodal Interaction, Montréal, QC, Canada, October 18-22, 2021*, pp. 177–185. ACM, 2021. doi: 10.1145/3462244.3479914. URL https://doi.org/10.1145/3462244.3479914.

Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *CoRR*, abs/2410.11190, 2024. doi: 10.48550/ARXIV.2410.11190. URL https://doi.org/10.48550/arXiv.2410.11190.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=CfXh93NDgH.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025. doi: 10.48550/ARXIV.2503.20215. URL https://doi.org/10.48550/arXiv.2503.20215.

Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/23c9c94227f937cfb50592a15e7fbb63-Abstract-Conference.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang,

Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3d human motion from speech. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 469–480. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00053. URL https://doi.org/10.1109/CVPR52729.2023.00053.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph.*, 39(6):222:1–222:16, 2020. doi: 10.1145/3414685.3417838. URL https://doi.org/10.1145/3414685.3417838.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *CoRR*, abs/2412.02612, 2024. doi: 10.48550/ARXIV.2412.02612. URL https://doi.org/10.48550/arXiv.2412.02612.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pp. 15757–15773. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.1055. URL https://doi.org/10.18653/v1/2023.findings-emnlp.1055.

Haoyu Zhang, Zhipeng Li, Yiwen Guo, and Tianshu Yu. Ex-omni: Enabling 3d facial animation generation for omni-modal large language models. *CoRR*, abs/2602.07106, 2026. doi: 10.48550/ARXIV.2602.07106. URL https://doi.org/10.48550/arXiv.2602.07106.

Juze Zhang, Philip Chen, Xin Chen, Heng Yu, Tiange Xiang, Ali Sartaz Khan, Shrinidhi K. Lakshmikanth, and Ehsan Adeli. Vibes: A conversational agent with behaviorally-intelligent 3d virtual body. *CoRR*, abs/2512.14234, 2025a. doi: 10.48550/ARXIV.2512.14234. URL https://doi.org/10.48550/arXiv.2512.14234.

Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Ziqiang Dang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13761–13771, October 2025b.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 35139–35148. AAAI Press, 2026. doi: 10.1609/AAAI.V40I41.40820. URL https://doi.org/10.1609/aaai.v40i41.40820.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. Voxcpm: Tokenizer-free TTS for context-aware speech generation and true-to-life voice cloning. *CoRR*, abs/2509.24650, 2025. doi: 10.48550/ARXIV.2509.24650. URL https://doi.org/10.48550/arXiv.2509.24650.

A Model Architecture Details

A.1 Reference Components and Output Decoding

The Motion-Omni framework specifies four components that together process a user's speech (or text) input and autoregressively generate a spoken response together with synchronised full-body co-speech motion (Figure 2). For each component, the framework prescribes only the input/output interface and the conditioning topology; the concrete model class, parameter count and tokenizer are choices made by a particular instance. Motion-Omni-Q7, the reference instance reported here, makes the following choices. The **Speech Encoder** is a frozen Whisper-large-v3 (Radford et al., 2023) encoder (hidden dimension 1280) that maps a 16 kHz waveform to continuous representations; a *speech projector* concatenates every five consecutive frames and passes them through a two-layer MLP into the LLM embedding space, downsampling the sequence fivefold. The **LLM Backbone** is a Qwen2.5-7B-Instruct (Yang et al., 2024) model. For speech input, projected Whisper features replace a designated `<speech>` placeholder in the token sequence; the LLM then processes the resulting continuous speech segment and surrounding text tokens. It is frozen during Stages 1–3 and fine-tuned with a small learning rate in Stage 4. The **Speech Generator** is a Qwen2-style transformer initialised from Qwen2.5-0.5B-Instruct (Yang et al., 2024) that autoregressively emits GLM-4-Voice (Zeng et al., 2024) discrete speech units at 12.5 Hz over a vocabulary of $16{,}384$ units (plus three control tokens). A Token-as-Query Gated Fusion (TQGF) block (Zhang et al., 2026) lets token embeddings query the LLM's contextualised hidden states through learned head-wise sigmoid gates; its full formulation is provided in Appendix A.2. The **Motion Generator** comprises four parallel per-part decoders that emit LOM (Chen et al., 2025a) VQ codes at 30 Hz for the face, hands, upper body and lower body, conditioned on the Speech Generator's last-layer hidden states (key/value) and on a learned speech-token query embedding (initialised from the pre-trained flow embedding of CosyVoice (Du et al., 2024b)). At inference time, the four part decoders share the Speech Generator context but do not explicitly cross-condition on one another's sampled motion outputs.

At inference time, speech units are converted to mel spectrograms by a CosyVoice (Du et al., 2024b) chunk-aware flow-matching decoder and then to 22.05 kHz waveforms by a HiFi-GAN (Kong et al., 2020)-style vocoder. Motion codes are decoded by the frozen LOM VQ-VAE into SMPL-X (Pavlakos et al., 2019) body/hand parameters and FLAME (Li et al., 2017) facial-expression coefficients. For the human evaluation we render each response as a video using an SMPL-X mesh (Appendix C).

A.2 Speech Generator

The Speech Generator is a Qwen2-style transformer initialised from Qwen2.5-0.5B-Instruct (Yang et al., 2024) and trained under a next-token-prediction objective over the GLM-4-Voice (Zeng et al., 2024) discrete speech-unit vocabulary. It runs at width $d = 896$ while the LLM backbone emits 3584-dimensional states, so the two are joined by a two-layer GELU-MLP and a reshape that together form the conditioning inputs of a Token-as-Query Gated Fusion (TQGF) block (Zhang et al., 2026):

$$\mathbf{Z} = \mathrm{Reshape}\big(\mathrm{MLP}(\mathbf{H}_{\text{llm}})\big) \in \mathbb{R}^{4T_t \times d}, \qquad \mathbf{Q} = \mathbf{E}_{\text{sg}}[\mathbf{w}] \in \mathbb{R}^{T_t \times d}, \tag{4}$$

where $\mathbf{w}$ are the T_t text tokens the LLM has already generated for its response, and $\mathbf{H}_{\text{llm}} \in \mathbb{R}^{T_t \times 3584}$ are the LLM's last-layer hidden states at those same positions. The conditioning stream $\mathbf{Z}$ supplies both the keys and the values of the block. The MLP widens each hidden state to $4d$ dimensions, and the reshape reads that vector as four consecutive d-dimensional positions, so $\mathbf{Z}$ is four times as long as the text sequence. The query stream, by contrast, looks the same tokens up in the Speech Generator's own embedding table $\mathbf{E}_{\text{sg}}$, giving one un-contextualised vector per token.

Each of the two TQGF layers then injects $\mathbf{Z}$ into the query stream. Writing $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{Z}}$ for the layer-normalised inputs, a layer computes

$$\mathbf{C} = \Big(\sigma(\tilde{\mathbf{Q}}\mathbf{W}_g) \odot \mathrm{Attn}\big(\tilde{\mathbf{Q}}\mathbf{W}_q,\, \tilde{\mathbf{Z}}\mathbf{W}_k,\, \tilde{\mathbf{Z}}\mathbf{W}_v\big)\Big)\mathbf{W}_o, \tag{5}$$

$$\mathbf{Q}' = \mathbf{Q} + \mathbf{C}, \qquad \mathbf{Q}'' = \mathbf{Q}' + \mathrm{FFN}(\mathbf{Q}'), \tag{6}$$

where Attn is multi-head scaled dot-product attention with h heads and no output projection, so $\mathrm{Attn}(\cdot) \in \mathbb{R}^{T_t \times h d_h}$; FFN is a position-wise feed-forward network with pre-norm and inner

width $4d$, σ is the logistic sigmoid and $\odot$ is element-wise multiplication. The projections satisfy $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_g \in \mathbb{R}^{d \times hd_h}$ and $\mathbf{W}_o \in \mathbb{R}^{hd_h \times d}$. We use $h = 16$ heads of dimension $d_h = 256$, so the gate acts in a 4096-dimensional space rather than in the 896-dimensional working space.

Plain cross-attention would send $\mathrm{Attn}(\cdot)\mathbf{W}_o$ straight into the residual stream; the factor $\sigma(\tilde{\mathbf{Q}}\mathbf{W}_g)$ in Equation (5) filters it first. The query holds one un-contextualised embedding per token, whereas the keys and values carry the LLM's semantic plan for the utterance, its prosodic intent and the dialogue history, along with information the speech head does not need. Gating before $\mathbf{W}_o$ lets each query token decide, coordinate by coordinate, how much of that signal to admit. Because Equation (6) adds the gated branch to $\mathbf{Q}$ instead of replacing it, a closed gate falls back on the un-contextualised query rather than producing a zero-information state.

A.3 VQ-VAE Body Representation

Following LOM (Chen et al., 2025a), we operate on 6D-rotation SMPL-X (Pavlakos et al., 2019) parameters for body and hand articulation, together with FLAME (Li et al., 2017) expression coefficients for the face (the same face parameterisation that is embedded inside the SMPL-X face model). The per-frame body state is decomposed into four parts as summarized in Table 5.

Table 5: SMPL-X + FLAME feature decomposition used by the four LOM VQ-VAEs. The face branch combines the SMPL-X jaw rotation with FLAME (Li et al., 2017) expression coefficients; the remaining three branches are SMPL-X joint rotations (6D rotation representation).

Part	Dim	Composition
Face	106	jaw (1×6) + FLAME expression (100)
Hand	180	30 hand joints $\times$ 6
Upper	78	13 upper-body joints $\times$ 6
Lower	61	9 lower-body joints $\times$ 6 + translation (3) + foot contact (4)

The four per-part VQ-VAEs share codebook size $|\mathcal{C}| = 256$ and are taken verbatim from the LOM `emage_vq` release (architecture `VQVAEConvZero`, 30 Hz, no temporal downsampling). We do not modify or fine-tune them.

A.4 Motion Decoder and Training Objective

Motion Generator. The Motion Generator instantiates one independent decoder per body part $b \in \{\mathrm{face}, \mathrm{hand}, \mathrm{upper}, \mathrm{lower}\}$. Each decoder stacks $L_{\mathrm{TQGF}} = 2$ TQGF layers, which reuse Equations (5)–(6) at $d_m = 512$, $h = 8$ and $d_h = 64$, followed by an $L_{\mathrm{self}} = 6$-layer self-attention Transformer with periodic rotary positional encoding (Zhang et al., 2026) of period $\mathcal{T} = 30$, that is, one second at 30 Hz. Its two input streams are defined in Equations (1)–(2).

Here $\mathbf{H}_s$ are the Speech Generator's last-layer hidden states and $\mathbf{W}_h \in \mathbb{R}^{896 \times d_m}$ projects them into the motion working dimension. The sequence $\mathbf{u} \in \{0, \ldots, 16{,}383\}^{T_s}$ holds speech units, $\mathbf{E}$ is a 512-dimensional look-up table initialised from the pre-trained CosyVoice flow embedding (Du et al., 2024b), $\mathbf{W}_e$ is a learned projection, and Interp linearly interpolates along the time axis from the 12.5 Hz speech-unit rate to the 30 Hz motion rate. All four decoders read the same $\mathbf{Z}$ and $\mathbf{Q}$ and differ only in their parameters. The self-attention block that follows operates at motion rate under a causal mask, so each frame attends only to earlier frames. During training $\mathbf{u}$ are the ground-truth units of the target response; at inference they are the units the Speech Generator has just emitted.

Conditioning on $\mathbf{H}_s$ in Equation (1) rather than on the decoded waveform is what removes a separate audio-to-motion stage at inference time, and it gives each decoder a representation that already carries acoustic timing together with response semantics. The same query/key asymmetry holds here: Equation (2) supplies one embedding per speech unit, identifying what is being said at that instant, while the keys and values additionally carry speaker timbre, which does not drive body movement, and the gate selects what each body part needs. A part-specific MLP head produces per-frame logits over the $|\mathcal{C}| = 256$ LOM codebook entries, trained with the weighted cross-entropy in Equation (3).

A.5 MOTION GENERATOR HYPERPARAMETERS

Table 6: Motion Generator architectural hyperparameters.

Parameter	**Value**
Working dimension d_m	512
Number of attention heads h	8
Per-head dimension d_h	64
TQGF depth L_{TQGF}	2
Self-attention depth L_{self}	6
TQGF FFN multiplier	4
Self-attention FFN multiplier	2
Dropout	0.0
Periodic RoPE period $\mathcal{T}$	30 frames
RoPE cache length	10,000
Max motion length $T_m^{\max}$	600 frames (20 s)
Per-part codebook size $\lvert\mathcal{C}\rvert$	256
Motion frame rate	30 Hz
Speech-token rate	12.5 Hz
Ratio T_m/T_s	2.4

A.6 STAGE-WISE TRAINING HYPERPARAMETERS

We train Motion-Omni-Q7 in four stages; throughout, the Whisper speech encoder and the LOM VQ-VAE are frozen. **Stage 1** trains only the speech projector with ASR supervision. Each sample contains a `<speech>` placeholder followed by its transcript target; projected Whisper features replace the placeholder, labels at the inserted speech positions are masked, and next-token cross-entropy is applied only to the transcript tokens. Because the LLM is frozen, this objective trains the projector to produce representations from which the LLM can decode the transcript. **Stage 2** trains the Speech Generator on TTS-style pairs while the LLM remains frozen. For example, the text-side instruction asks the model to say or restate a supplied sentence, the LLM produces a semantically consistent response representation, and the target is the corresponding sequence of discrete speech units. **Stage 3** attaches the Motion Generator and jointly trains it with the Speech Generator, while keeping the LLM and projector frozen, on the TTSM corpus produced by our data construction pipeline (Section 3.4); the data is consumed under a four-substage curriculum that exposes the network to progressively larger quality quantiles in turn, warm-starting each substage from the previous one. **Stage 4** unfreezes the LLM backbone, the speech projector, the Speech Generator and the Motion Generator simultaneously and optimises a four-task ASR/TTS/Speech-to-Speech-with-Motion (S2SM)/T2T mixture with a constant-with-warmup schedule, warm-started from the final Stage 3d checkpoint. The per-stage settings are summarised below, followed by implementation and numerical-stability details.

Table 7: Training configuration of the four stages (rows are settings; columns are stages). †: Stage 4 starts from the best Stage 3d checkpoint. "MG": Motion Generator; "SG": Speech Generator; "SP": Speech Projector; "LR" is the peak learning rate. Warmup is given as a fraction of total steps unless noted.

Setting	**Stage 1**	**Stage 2**	**Stage 3 (a–d)**	**Stage 4**†
Trainable modules	SP	SG	SG + MG	LLM + SP + SG + MG
LR (SP)	1×10^{-3}	–	–	1×10^{-6}
LR (SG)	–	1×10^{-4}	1×10^{-5}	5×10^{-7}
LR (MG)	–	–	2×10^{-4}	2×10^{-6}
LR (LLM)	–	–	–	2×10^{-7}
LR schedule	cosine	cosine	const + warmup	const + warmup
Warmup	0.30	0.10	0.03	3000 steps
Grad clip	1.0	1.0	1.0	0.5 + NaN guard
Effective batch size	128	128	128	128
Epochs	1	2	$\sim$5 per substage	until val. plateau

A.7 Implementation Details

Hardware and software. All training is conducted on 4 GPUs using DeepSpeed ZeRO-2 (Rasley et al., 2020), BF16 mixed precision and gradient checkpointing. The effective batch size is 128 across all stages.

Model parameters. The total parameter count of Motion-Omni-Q7 is approximately 8.3B, broken down as follows:

- LLM Backbone (Qwen2.5-7B-Instruct): $\sim$7.6B
- Speech Generator (Qwen2.5-0.5B-Instruct): $\sim$0.5B
- Motion Generator (4 part-specific decoders, d=512): $\sim$150M
- Speech Projector (2-layer MLP with stride-5): $\sim$5M

Training cost. The full four-stage training of Motion-Omni-Q7 takes approximately 960 GPU-hours in total.

A.8 Loss Composition and Multi-Task Balancing

Let $\mathcal{L}_{\text{LM}}$ denote the LLM next-token cross-entropy, $\mathcal{L}_{\text{sp}}$ the speech-unit label-smoothed cross-entropy, and $\mathcal{L}_{\text{mo},b}$ the per-part motion cross-entropy on body part $b \in \{\text{face}, \text{hand}, \text{upper}, \text{lower}\}$. The per-batch loss is

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{sp}} \cdot \mathbb{1}[\text{speech task}] + \sum_{b} w_b \, \mathcal{L}_{\text{mo},b} \cdot \mathbb{1}[\text{motion task}], \tag{7}$$

with $(w_{\text{face}}, w_{\text{hand}}, w_{\text{upper}}, w_{\text{lower}})$ proportional to the feature dimensions of Table 5 (i.e. $(106, 180, 78, 61)/425$). The Speech Projector has no dedicated loss term; it is trained through $\mathcal{L}_{\text{LM}}$ via backpropagation when the input contains audio (the projected Whisper features feed into the LLM, and the LLM CE loss gradient flows back through the projector). The masks $\mathbb{1}[\cdot]$ select the samples whose task requires the corresponding supervision: ASR contributes only $\mathcal{L}_{\text{LM}}$; TTS, TTSM and S2SM contribute $\mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{sp}}$; TTSM and S2SM additionally contribute $\sum_b w_b \mathcal{L}_{\text{mo},b}$; T2T contributes only $\mathcal{L}_{\text{LM}}$ on text-only samples. When logging training curves, we average each per-task loss only over batches that contain at least one sample of the corresponding task, so that zero-padding from other tasks does not artificially depress the reported loss.

B Data Pipeline Details

B.1 Training Corpora and Stage Statistics

Reference instantiation. Table 8 summarises the per-stage data statistics. **Stage 1** uses the filtered English ASR mixture depicted in Figure 3. After filtering and audio-path deduplication, the retained records comprise GigaSpeech (Chen et al., 2021) (1,601,834), SPGISpeech (O'Neill et al., 2021) (1,436,627), Common Voice (Ardila et al., 2020) (615,844), InstructS2S-200K (Fang et al., 2025) (340,163), and LibriSpeech (Panayotov et al., 2015) (161,599), summing to 4,156,067 ASR records. The retained LibriSpeech records comprise 16,969 from `train-clean-100`, 61,316 from `train-clean-360`, and 83,314 from `train-other-500`. **Stage 2** synthesises speech with Qwen3-TTS (Team, 2026) for each text sample from either English Emilia (He et al., 2024) or InstructS2S assistant responses. Filtering retains 715,628 Emilia-derived records and 327,681 InstructS2S-derived records, summing to 1,043,309 TTS records. Each record contains one text sample and its corresponding synthesised speech, represented by GLM-4-Voice speech-unit targets (Zeng et al., 2024); texts from the two datasets are not paired with each other. These counts refer to dataset records, not repeated presentations over training epochs. **Stage 3** applies the pipeline to InstructS2S-200K with the LOM teacher. Training directly on the complete unranked corpus diverged, so we instead expose the network to progressively larger quantiles of the dual-metric score ($12.5\%/25\%/50\%/100\%$), warm-starting each substage from the previous checkpoint. The stage-wise teacher-reference Fréchet Gesture Distance (FGD, defined in Section 4.4) reported in Table 12 decreases from 0.3974 at Stage 3a to 0.3040 after Stage 4. **Stage 4** assembles a four-task ASR/TTS/S2SM/T2T mixture. The T2T component ($\sim$213K samples) is drawn from six text-only datasets (SODA (Kim et al., 2023), WildChat (Zhao et al., 2024), Tulu-3 (Lambert et al., 2024), WizardLM (Xu et al., 2024a), OpenThoughts (Guha et al., 2025), NuminaMath-CoT (LI et al., 2024)) to preserve general text-response planning and multi-turn dialogue behaviour during joint fine-tuning. Without this component, responses more often repeated the user's input. All sources are English-filtered with `langdetect` and capped at $10{,}000$ characters per dialogue. The following subsections detail filtering and quality scoring.

Table 8: Training records by stage.

Stage	Samples
Stage 1: ASR	4,156,067
Stage 2: TTS	1,043,309
Stage 3: TTSM	422,856
Stage 4: Mixture	724,508

B.2 ASR Audio Filtering

For Stage 1 ASR supervision, we discard transcripts containing non-English characters, transcripts shorter than two words, and utterances outside the 1–30 s duration range. We require an energy-distribution SNR proxy of at least 15 dB, except for InstructS2S and SPGISpeech, which bypass this check. Whisper-large-v3 (Radford et al., 2023) then retranscribes every candidate, and only pairs with $\mathrm{WER} \leq 5\%$ are retained. For all-uppercase transcripts, we use the Whisper-cased transcript only when the normalised word sequence matches exactly ($\mathrm{WER} = 0$); otherwise the sample is removed. Dataset-specific punctuation placeholders are normalised after filtering, and duplicate audio paths are removed.

B.3 TTS Audio Filtering

All Stage 2 TTS waveforms, including those derived from Emilia text (He et al., 2024) and InstructS2S assistant-response text (Fang et al., 2025), are synthesised with Qwen3-TTS (Team, 2026); the speech-unit targets use the GLM-4-Voice tokenizer and vocabulary (Zeng et al., 2024). For Stage 2 TTS supervision, we first remove text containing non-English characters or outside the range of 2–200 words, then require that the waveform is present and decodable, lasts 1–25 s. Text–audio agreement is checked by Whisper-large-v3, retaining only samples with $\mathrm{WER} < 5\%$. We further apply acoustic event detection (AED) to each waveform and reject those whose non-speech portion exceeds 30% of the total duration.

B.4 WEIGHTED VQ-VAE RECONSTRUCTION ERROR

The weighted VQ-VAE reconstruction error L_1^{vq} measures how faithfully the frozen LOM VQ-VAE codebook can represent a given motion sample. For each sample in the Stage 3 corpus, we compute this metric as follows.

Step 1: Feature extraction. The original motion parameters (stored as SMPL-X axis-angle rotations, FLAME expression coefficients and global translation) are converted into per-part continuous feature vectors following the body decomposition in Table 5: face ($D_{\text{face}} = 106$), hand ($D_{\text{hand}} = 180$), upper body ($D_{\text{upper}} = 78$) and lower body ($D_{\text{lower}} = 61$). Axis-angle joint rotations are converted to 6D rotation representations before concatenation.

Step 2: Encode, quantise, decode. Each part's feature sequence $\hat{m}_{1:T}^{(b)} \in \mathbb{R}^{T \times D_b}$ is passed through the frozen LOM VQ-VAE for that part:

$$\hat{m}^{(b)} \xrightarrow{\text{encoder}} z^{(b)} \xrightarrow{\text{quantise}} c^{(b)} \xrightarrow{\text{decoder}} \tilde{m}^{(b)},$$

where the quantisation step maps each latent vector to its nearest codebook entry ($|\mathcal{C}| = 256$). The reconstructed sequence $\tilde{m}_{1:T}^{(b)}$ has the same dimensionality and frame rate as the original.

Step 3: Per-part L1 error. For each body part b, we compute the mean absolute error averaged over all frames and all feature dimensions:

$$\ell_b \;=\; \frac{1}{T \cdot D_b} \sum_{t=1}^{T} \big\| \hat{m}_t^{(b)} - \tilde{m}_t^{(b)} \big\|_1. \tag{8}$$

Step 4: Dimension-weighted aggregation. The final score is the dimension-weighted average across parts:

$$\begin{aligned} L_1^{\text{vq}}(\hat{m}) &= \frac{\sum_b D_b \cdot \ell_b}{\sum_b D_b} \\ &= \frac{106\,\ell_{\text{face}} + 180\,\ell_{\text{hand}} + 78\,\ell_{\text{upper}} + 61\,\ell_{\text{lower}}}{425}, \end{aligned} \tag{9}$$

so that each feature dimension contributes equally to the aggregate score regardless of which body part it belongs to.

Interpretation. L_1^{vq} measures *supervision fidelity*: it quantifies how faithfully the discrete VQ codes (which serve as training targets for the Motion Generator) represent the underlying continuous motion. A high L_1^{vq} indicates that the quantisation step maps the latent to a distant codebook entry, so the code assigned by the teacher VQ-VAE is a poor proxy for the intended motion and training on such samples injects noisy label supervision. However, a low L_1^{vq} alone does not guarantee the motion is useful as supervision: the codebook may faithfully represent a motion sequence that is nevertheless poorly aligned with the driving speech. The complementary beat correlation score (Section B.5) catches exactly this failure mode by measuring speech-motion rhythmic coupling. Together the two metrics form a dual filter: L_1^{vq} ensures label fidelity (the codes reliably represent the motion), while BC ensures speech-motion coupling (the motion is rhythmically aligned with the audio). The curriculum (Section 3.4) leverages the combined score to expose the network to high-fidelity, well-aligned samples first.

B.5 BEAT CORRELATION SCORE (BC)

Given the assistant audio a and LOM-predicted motion $\hat{m}$, we first extract a body-level kinematic envelope $v_t = \|\dot{X}_{\text{upper},t}\|_2$ on the upper body at 30 Hz, and detect kinematic beats $\mathcal{M} = \{t : v_t < v_{t-1},\, v_t < v_{t+1},\, v_t > 0.3\bar{v}\}$ as the local minima of v above a relative threshold. We extract audio onsets $\mathcal{O}$ from a via librosa (McFee et al., 2015) and compute the forward (audio→motion)

Gaussian-kernel beat-distance score

$$\mathrm{BC}(a, \hat{m}) = \frac{1}{|\mathcal{O}|} \sum_{t_o \in \mathcal{O}} \exp\left(-\frac{\min_{t_m \in \mathcal{M}} (t_o - t_m)^2}{2\sigma^2} \right), \quad \sigma = 9 \text{ frames}. \tag{10}$$

This is the audio→motion direction of the EMAGE BC formulation (Liu et al., 2024), restricted to a single body-level envelope. We use this established BC direction for curriculum scoring and the main descriptive comparison. The bidirectional $\mathrm{BC}_{\mathrm{bi}}$ defined in Equations 11–13 is evaluated only as an exploratory diagnostic.

Bidirectional beat diagnostic. We additionally examine a bidirectional beat-correlation diagnostic, $\mathrm{BC}_{\mathrm{bi}}$. Standard BC matches audio onsets $\mathcal{O}$ to nearby motion beats $\mathcal{M}$ in one direction. Since both audio and motion are generated in our setting, we average both matching directions:

$$\mathrm{BC}_{a\to m} = \frac{1}{|\mathcal{O}|} \sum_{t_o \in \mathcal{O}} \exp\left(-\frac{\min_{t_m \in \mathcal{M}} (t_o - t_m)^2}{2\sigma^2} \right), \tag{11}$$

$$\mathrm{BC}_{m\to a} = \frac{1}{|\mathcal{M}|} \sum_{t_m \in \mathcal{M}} \exp\left(-\frac{\min_{t_o \in \mathcal{O}} (t_o - t_m)^2}{2\sigma^2} \right), \tag{12}$$

$$\mathrm{BC}_{\mathrm{bi}} = \tfrac{1}{2} \left(\mathrm{BC}_{a\to m} + \mathrm{BC}_{m\to a} \right), \tag{13}$$

where audio onsets are detected with `librosa.onset.onset_detect` (McFee et al., 2015), motion beats are local minima of the upper-body kinetic-energy envelope above a relative threshold, and $\sigma = 0.3\,\mathrm{s}$ follows the LOM setup. The two directions respectively penalise unmatched audio onsets and unmatched motion beats.

We compare the per-pair ordering induced by $\mathrm{BC}_{\mathrm{bi}}$ with the majority R1 rhythm verdict on the 100 human-evaluated pairs. Treating score differences below 0.07 as ties gives 46% agreement, against 40% for standard BC under the same tie-aware protocol. On a set of this size neither figure separates from the chance reference, so we report $\mathrm{BC}_{\mathrm{bi}}$ as an exploratory diagnostic rather than a validated perceptual metric.

B.6 Robust Percentile Normalisation

To form the combined quality score $s(x)$ in Section 3.4, we robustly normalise each metric: the 5th percentile is mapped to 0, the 95th percentile to 1, and values outside the $[0, 1]$ range are clipped. The L_1^{vq} axis is additionally inverted (large → 0, small → 1) so that 1 denotes the highest quality in both axes. Samples with either metric unavailable are excluded from the ranked early subsets. The final Stage 3d pool includes all samples with valid motion codes, including those without both quality scores.

C Rendering Pipeline

To produce the video stimuli used in the human preference study, we render each motion sequence as a 30 fps video with synchronised response audio. The rendering pipeline uses the following assets and tools:

- **Mesh template**: Meshcapade SMPL-X female mesh (`SMPLX-female.obj`, CC-BY-4.0 licence), which provides a canonical UV layout with 42,064 UV-seam-duplicated vertices mapped to 10,475 unique SMPL-X vertices.
- **Material**: PBR metallic-roughness material with an albedo map (`f_01_alb.002.palmfix.png`, palm-tone-corrected) and a normal map (`f_01_nrm.002.png`) from the Meshcapade texture samples; metallic factor 0.0, roughness factor 0.6.
- **Lighting**: three-point directional lights (key, fill, rim) plus an ambient term estimated from the median of an HDRI environment map (`studio_small_03_4k.hdr`).

- **Renderer**: `pyrender` with headless EGL backend (`PYOPENGL_PLATFORM=egl`), producing 720×720 frames.
- **Body model**: SMPL-X (Pavlakos et al., 2019) with FLAME (Li et al., 2017) expression blendshapes, driven by the decoded per-frame 6D-rotation parameters and expression coefficients from the LOM VQ-VAE output.

Frames are composited into an MP4 (H.264) with the response audio muxed in, using FFmpeg. The same rendering configuration is applied identically to all systems (Motion-Omni-Q7 and cascaded baselines) so that visual differences arise solely from the motion stream.

C.1 Facial Geometry and Lip-Synchronisation Metrics

For facial geometry, we compute the mean squared error (MSE) and the landmark velocity difference (LVD) on facial vertices generated from jaw pose and FLAME expression coefficients, using the LOM-generated motion as the pseudo-reference; these two metrics therefore measure fidelity to the LOM teacher facial-motion distribution rather than direct agreement with newly captured real facial motion. For lip synchronisation, we render the generated SMPL-X+FLAME motion into standardised frontal face videos and evaluate them with the lip-sync error scripts released in the official Wav2Lip evaluation code (Prajwal et al., 2020), which score a rendered video against its audio with a pretrained SyncNet and report a confidence (LSE-C, higher is better) and a feature distance (LSE-D, lower is better). Neither needs a motion reference. All systems use the same renderer, avatar, face crop, 25 FPS video, and 16 kHz mono audio.

D Human Evaluation Protocol

Because automatic gesture metrics (FGD, BC) are imperfect proxies for perceived motion quality, we additionally conduct a paired A/B/tie human preference study.

Evaluation set. We select 25 clips per comparison arm from the external SwDA-500 evaluation set, covering both short (4–10 s) and long ($\geq$ 10 s) responses across the sampled dialogue topics. Clips with degenerate responses ($<$ 40 characters) or missing motion files are excluded.

Study design. Each of the 25 clips is compared across four arms, yielding 100 total pairs: B1 is Qwen2.5-Omni + LOM, B2 is Qwen2.5-Omni + EMAGE, B3 is MO-audio + LOM, and B4 is MO-audio + EMAGE. Arms B1/B2 use a different speech model, so both speech and motion differ from Motion-Omni-Q7; arms B3/B4 share Motion-Omni-Q7's own audio, isolating the motion contribution.

Interface and rating. Each pair is presented as two side-by-side rendered videos (left = System A, right = System B) with the system assignment randomised per pair (fixed seed = 42). Annotators judge each pair on **three motion-related dimensions** independently, choosing one of *A is better*, *B is better*, or *Tie* per dimension.

Annotators. Prior video generation work has used 3 expert annotators for evaluation (Long et al., 2026). Our four annotators were young-adult males (mean age approximately 26), all fluent in English and experienced in evaluating motion-generation or spoken-dialogue outputs. Before annotation, all received the rubric in Appendix D.1. Four annotators independently rate all 100 pairs on dimensions R1–R3 (video+audio). Per-pair verdicts are determined by majority vote across the four raters. Inter-annotator agreement is measured by mean pairwise Cohen's kappa.

Interpretation of conversational motion preferences. We attribute this to a domain mismatch in the baselines' training data. The training corpus of LOM and EMAGE (BEAT2 (Liu et al., 2024)) contains not only conversational clips but also monologue and public-speaking recordings; as a result, their generated motion sometimes exhibits presentation-style behaviours (pacing, turning sideways, exaggerated arm swings) that annotators perceive as unnatural in a dialogue setting. MO, by contrast, is trained on dialogue-paired pseudo-motion produced by our data construction pipeline (Section 3.4), so its motion style is calibrated to a conversational register. This may contribute to the observed R3 preference over EMAGE under identical speech.

D.1 Annotation Rubric

The following rubric is provided to all annotators.

R1 – Rhythm. Does the body motion match the rhythm of the speech? We instruct annotators to check whether hand gestures or head movements accompany stressed words, whether the body pauses when the speech pauses (rather than continuing to gesture), and whether gesture density tracks speech rate. A candidate is preferred when it feels like "this person is actually saying these words."

R2 – Semantic Alignment. Do the gestures convey the *content* being spoken? Positive examples include counting fingers for enumerated items, pointing for spatial references, tracing shapes for described objects, and head nods/shakes for affirmation/negation. Generic rhythmic hand-waving (e.g. open-palm emphasis) does *not* count as semantic alignment. If neither side shows clear semantic gestures, annotators select tie.

R3 – Body Naturalness. Does the motion look like a real person in conversation? Annotators check for (a) mechanical artefacts (jitter, interpenetration, sudden joint pops, foot sliding), (b) posture issues (frozen pose, exaggerated public-speaking stance), and (c) overall relaxation (subtle weight shifts, natural idle motion). We emphasise that large motion amplitude does not imply naturalness.

Scoring principles. Annotators are instructed to (1) score each dimension independently without first forming an overall impression, (2) treat tie as a legitimate option.

E Spoken Dialogue Quality on VoiceBench

Public SDM benchmarks (Chen et al., 2026c; Ma et al., 2025) evaluate dialogue capabilities. Table 9 gives the per-subset VoiceBench (Chen et al., 2026c) scores summarised in Section 4.2. The listed systems differ by orders of magnitude in the amount of speech seen during pretraining and in whether speech generation is their primary objective, so we report the scores as context for the capability-retention check rather than as a like-for-like ranking, and do not mark best values.

Table 9: Spoken dialogue quality on VoiceBench (Chen et al., 2026c). Higher is better throughout. GPT-4o-Audio is served through an API and its weights are not released; every other system releases its weights. Dashes denote scores not reported.

Model	Alpaca	Common	Wild	SD-QA	MMSU	OBQA	BBH	IFEval	AdvBench	Overall
Closed-weight (API only)										
GPT-4o-Audio (OpenAI, 2024)	4.78	4.49	4.58	75.50	80.25	89.23	84.10	76.02	98.65	86.75
Open-weight										
Kimi-Audio (KimiTeam et al., 2025)	4.46	3.97	4.20	63.12	62.17	83.52	69.70	61.10	100.00	76.91
Qwen2.5-Omni (Xu et al., 2025)	4.49	3.93	–	55.71	61.32	81.10	–	52.87	99.42	–
VITA-1.5 (Fu et al., 2025)	4.21	3.66	3.48	38.88	52.15	71.65	55.30	38.14	97.69	64.53
LLaMA-Omni (Fang et al., 2025)	3.70	3.46	2.92	39.69	25.93	27.47	49.20	14.87	11.35	41.12
Ex-Omni (Zhang et al., 2026)	2.57	2.87	2.19	40.14	24.24	25.49	50.40	16.22	83.08	43.57
Moshi (Défossez et al., 2024)	2.01	1.60	1.30	15.64	24.04	25.93	47.40	10.12	44.23	29.51
Mini-Omni2 (Xie & Wu, 2024)	2.32	2.18	1.79	9.31	24.27	26.59	46.40	11.56	57.50	33.49
Motion-Omni-Q7 (ours)	3.17	3.35	2.97	39.78	27.36	26.59	51.00	20.87	73.27	47.63

F Speech Naturalness Proxy

Because WER measures only intelligibility, we additionally report UTMOSv2 (Baba et al., 2024) on SwDA-500 generated audio as an automatic proxy for speech naturalness. UTMOSv2 is a learned predictor of mean opinion score and does not replace a listening test; we report it as a lightweight speech-quality check under conversational wording rather than as a competitive comparison, and the reference systems below are dedicated TTS models that synthesise a supplied sentence rather than plan a dialogue response.

Table 10: Automatic speech naturalness proxy on SwDA-500 generated audio. In each column, **bold** marks the best value and underline the second best.

Model / System	UTMOSv2 en mean↑
Qwen3-TTS (Team, 2026)	**4.05**
CosyVoice 3 (Du et al., 2025)	<u>3.92</u>
GLM-TTS (Cui et al., 2025)	3.31
VoxCPM1.5 (Zhou et al., 2025)	3.17
F5-TTS (Chen et al., 2025b)	3.15
CosyVoice (Du et al., 2024a)	3.06
Motion-Omni-Q7 (ours)	3.77

Motion-Omni-Q7 scores 3.77, above GLM-TTS, VoxCPM1.5, F5-TTS and CosyVoice, and below Qwen3-TTS (4.05) and CosyVoice 3 (3.92).

G Fusion and Curriculum Ablations

G.1 Motion-Conditioning Fusion

To isolate the fusion design, all variants start from the same Stage 2 checkpoint and use the same Stage 3a data and optimisation schedule. The only changed factor is how the motion query incorporates contextual information. TQGF/LLM K/V reads keys and values directly from the LLM; the remaining variants read Speech Generator states. We report the established unidirectional BC metric only.

Table 11: Motion-conditioning ablation under the Stage 3a setting.

No.	Fusion / context source	BC↑
1	TQGF / LLM K/V	0.827
2	Concatenation / Speech Generator K/V	0.809
3	GLU / Speech Generator K/V	0.799
4	Cross-attention / Speech Generator K/V	0.905
5	FiLM / Speech Generator K/V	0.852
6	TQGF / Speech Generator K/V	**0.912**

The comparison supports two scoped observations: Speech Generator context is more useful than direct LLM context for the same TQGF operator (No. 6 vs. No. 1), and gated fusion is modestly better than plain cross-attention under this fixed Stage 3a setup (No. 6 vs. No. 4). It does not isolate every internal component of TQGF.

G.2 Stage-wise Teacher-Reference FGD

Using the same LOM latent encoder and teacher-reference definition as the main motion evaluation, FGD decreases through all curriculum stages and remains lowest after Stage 4 joint fine-tuning.

Table 12: Teacher-reference FGD across training stages.

Checkpoint	FGD↓
Stage 3a	0.3974
Stage 3b	0.3258
Stage 3c	0.3079
Stage 3d	0.3075
Stage 4	**0.3040**

These results are consistent with progressive improvement under the curriculum.

G.3 Quality–Latency Trade-off

Figure 4 jointly plots the BC values from Table 2 and the RTF values from Table 4. Motion-Omni-Q7 is the only evaluated system that combines faster-than-real-time generation ($\mathrm{RTF} < 1$) with $\mathrm{BC} > 7.5$. MO-audio + EMAGE has similar latency (0.84 versus 0.78) but lower BC (7.32 versus 7.59), whereas MO-audio + LOM attains higher BC (7.67) at an RTF of 4.39 and $5.4\times$ the response time.

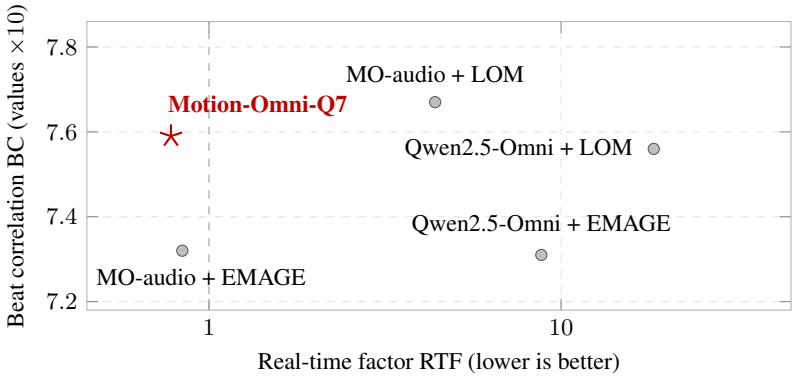
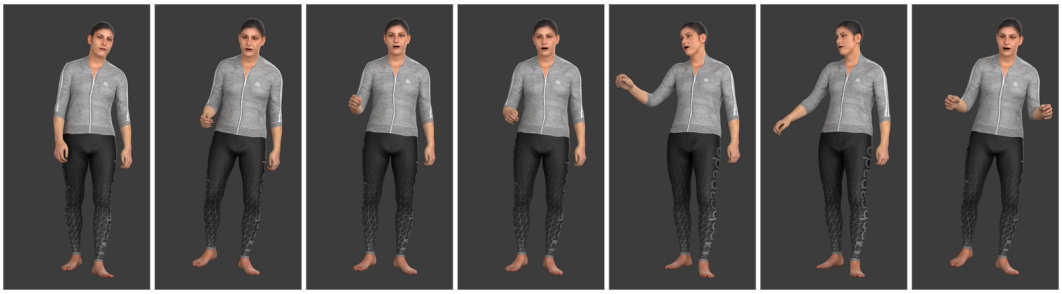


Figure 4: Motion quality against response latency on SwDA-500. Circles are two-stage cascades; the star is our end-to-end model. Points closer to the top-left are better. The dashed line marks $\mathrm{RTF} = 1$, left of which a system responds faster than real time.

H Qualitative Visualisations

This section shows rendered frames from the released demonstration clips. All frames come from the rendering pipeline of Section C; the full videos are part of uploaded files.

Frame selection. Uniform time sampling tends to return near-duplicate poses, so for Figure 5 we select the 7 frames that are mutually farthest apart in upper-body joint-angle space (spine through wrists) subject to a minimum temporal separation of 10% of the clip, then display them in temporal order. For the cross-system comparison in Figure 6 we deliberately do *not* use this criterion, because choosing timestamps from our own motion would bias the comparison toward moments at which our model happens to move; instead we sample seven evenly spaced timestamps over the middle of the clip and apply the same timestamps to every system. In both figures every tile is cropped with a single shared crop box, computed as the union of the avatar bounding boxes over all frames in the figure, so that the avatar keeps the same scale and image position in every tile and differences between tiles are differences in motion rather than in framing.



(a) “Here are five positive character traits: honesty, empathy, kindness, confidence, and curiosity.” Frames at $t = 1.1, 1.9, 2.6, 3.9, 4.8, 6.0, 6.9$ s.



(b) “Supervisors can provide constructive feedback by being specific, timely, and respectful, focusing on behavior rather than personality.” Frames at $t = 1.4, 2.3, 3.4, 4.3, 5.2, 6.6, 8.0$ s.



(c) “I think the show is trying to convey that social relationships are complex and can be both positive and negative, and that they can have a significant impact on our mental health and well-being.” Frames at $t = 1.2, 2.3, 4.2, 5.2, 7.0, 7.7, 8.5$ s.

Figure 5: Motion-Omni-Q7 outputs on three released clips. Each row shows seven frames of a single response in temporal order, together with the text the model spoke in that response; the speech and the motion are produced in the same autoregressive pass. Frames are the seven mutually most distinct upper-body poses subject to a minimum temporal gap, not uniform samples.



(a) Motion-Omni-Q7 (ours).



(b) MO-audio + LOM.



(c) MO-audio + GestureLSM.



(d) MO-audio + EMAGE.



(e) MO-audio + MambaTalk.

Figure 6: The same response audio driving five different motion sources. All rows use the speech Motion-Omni-Q7 generated for "Here are five positive character traits: honesty, empathy, kindness, confidence, and curiosity.", sampled at the same seven timestamps ($t = 1.1$ to 6.9 s in steps of about 0.98 s) and rendered with an identical pipeline, so the only variable across rows is the motion stream.