EX-OMNI: ENABLING 3D FACIAL ANIMATION GENERATION FOR OMNI-MODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (OLLMs) aim to unify multimodal understanding and generation, yet extending them to jointly produce speech and 3D facial animation remains largely underexplored. A key challenge is the mismatch between the discrete semantic reasoning of LLMs and the dense temporal dynamics required for 3D facial motion. We propose Expressive Omni (Ex-Omni), a framework that augments OLLMs with speech-accompanied 3D facial animation. Ex-Omni decouples semantic reasoning from temporal generation through a speech-unit generator with blendshape co-supervision and a non-autoregressive blendshape decoder, where speech units provide temporal scaffolding and hidden speech representations carry facially relevant cues. We further introduce a token-as-query gated fusion (TQGF) interface for controlled semantic injection, as well as InstructS2SF-1200K, a 1.2M-sample weakly supervised dataset for speech-accompanied facial animation. Extensive experiments show that Ex-Omni retains competitive speech QA capability while natively generating coordinated text, speech, and 3D facial animation, and approaches the Audio2Face-3D teacher cascade in synchronization and human preference. The demo, code and data can be found in **uploaded supplementary materials** and **`https://anonymous.4open.science/r/anonymity_adkshf`.**

1 INTRODUCTION

Large language models (LLMs) have achieved strong progress across a wide range of tasks Minaee et al. (2024), demonstrating impressive generalization and reasoning capabilities. As research continues to expand from unimodal understanding toward multimodal understanding and generation, growing attention has been devoted to unifying these tasks within a single framework, commonly referred to as omni-modal large language models (OLLMs). With the continued development of such omni models Fu et al. (2025); Xie & Wu (2024b); Team (2025); Li et al. (2025); Xie & Wu (2024a); Luo et al. (2025); Xu et al. (2025); AI et al. (2025), these systems are increasingly relevant to human-computer interaction and embodied intelligence.

In the area of human-computer interaction, there is a growing demand for OLLMs that can engage in natural and expressive interactions with humans. Human communication is inherently multimodal and further extends beyond verbal content alone. In face-to-face interaction, temporally coherent 3D facial animation synchronized with speech plays a crucial role in conveying non-verbal cues and enhancing interaction naturalness, particularly in applications such as virtual characters, digital avatars, and embodied agents. However, existing OLLMs primarily focus on linguistic, acoustic, or pixel-level visual outputs, while expressive non-verbal modalities such as 3D facial animation remain largely underexplored. Motivated by this gap, we investigate integrating 3D facial animation generation into OLLMs. A natural idea is to directly attach a blendshape decoder to an LLM and predict animation from its hidden representations. In practice, we found this design exposes a challenge: LLM hidden states are optimized for sparse, token-level semantics with weakly constrained temporal structure, whereas 3D facial animation requires dense and temporally smooth motion at a much finer time scale. Bridging these representations forces the decoder to infer fine-grained dynamics from coarse semantic features, resulting in an ill-conditioned mapping that typically demands substantially larger model capacity and more paired speech–face supervision for stable generation.

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.

In this paper, we propose Expressive Omni (Ex-Omni), an omni-modal framework that augments OLLMs with speech-accompanied 3D facial animation, where facial motion is represented using ARKit-52 blendshape coefficients Lewis et al. (2014) and generated in a non-autoregressive manner. Ex-Omni follows text/speech instructions to generate synchronized speech paired with facial animation in an end-to-end manner. We use *expressive* in the sense common to speech-driven facial animation: coherent articulation, phoneme-related mouth motion, and speech–face coordination in addition to the spoken content. To facilitate stable and temporally coherent facial animation learning from LLM semantics, Ex-Omni decouples semantic reasoning from temporal generation through a hierarchical temporal-generation design. Rather than directly predicting facial motion from LLM hidden states, Ex-Omni first employs a speech-unit generator with blendshape co-supervision, where discrete speech units provide explicit temporal scaffolding while the generator hidden states are encouraged to encode facially relevant cues. A blendshape decoder then predicts blendshape sequences conditioned on both signals. To better bridge high-level semantics and temporal motion generation, we further introduce a unified token-as-query gated fusion (TQGF) mechanism to selectively regulate how and when semantic information is injected into the speech and facial generation processes, simplifying optimization and improving temporal alignment.

We also construct InstructS2SF-1200K, a 1.2M-sample weakly supervised dataset for training OLLMs with speech-accompanied 3D facial animation. It contains two subsets tailored to different training stages: 1M Text-to-Speech-Face (TTSF) samples for speech-blendshape co-pretraining, and 200K Speech-to-Speech-Face (S2SF) QA samples for dialogue-based speech-blendshape co-pretraining. By combining large-scale weak supervision with speech-face data, InstructS2SF-1200K helps bridge the gap between limited real-world recordings and open-domain generalization. Experimental results show that Ex-Omni preserves substantial speech-task capability, improves over the evaluated non-teacher cascades in the controlled downstream setting, and remains close to the Audio2Face-3D teacher cascade while supporting native face generation.

Overall, the key contributions are:

- We propose Expressive Omni (Ex-Omni), an OLLM framework for unified instruction following and generation across text, speech, and speech-accompanied 3D facial animation.
- To reduce the difficulty of learning temporally coherent 3D facial animation from LLM semantics, Ex-Omni adopts a two-stage temporal-generation design. Discrete speech units provide the primary temporal scaffold, while TQGF bridges LLM semantics to speech generation and speech-generator representations to facial-animation generation.
- We construct InstructS2SF-1200K, a 1.2M-sample weakly supervised corpus with 1M TTSF synthesis samples and 200K dialogue-based S2SF samples. Experiments show that Ex-Omni approaches its teacher cascade and improves over the evaluated non-teacher cascades in controlled downstream comparisons while enabling efficient native face generation without an external S2F model at inference.

2 Related Work

Omni-modal Large Language Models. OLLMs Fu et al. (2025); Xie & Wu (2024b); Team (2025); Li et al. (2025) represent an important direction in multimodal large language models, as they integrate the capabilities of understanding and generating across multiple modalities, such as text, speech, and vision, within a unified framework. For example, Mini-Omni Xie & Wu (2024a) utilizes text-instructed speech generation and batch-parallel strategies to achieve seamless speech synthesis while preserving the model's text capabilities. OpenOmni Luo et al. (2025) introduces a two-stage training framework to achieve zero-shot cross-modal alignment from vision-language tasks to speech-language tasks. Qwen2.5-Omni Xu et al. (2025) introduce the Thinker-Talker architecture to integrate text, speech, and vision modalities into a unified end-to-end model. Ming-Omni AI et al. (2025) integrates visual generation capabilities into a unified omni-modal model, utilizing modality-specific routers to achieve understanding and generation across multiple modalities.

Facial Animation Generation. Facial animation generation is an important research area for improving system interactivity. Earlier methods Chen et al. (2019); Mittal & Wang (2020); Zhang et al. (2021); Hong et al. (2022) mainly focus on 2D facial animation generation, a field that has become

mature after years of research. In recent years, 3D facial animation generation Richard et al. (2021); Xing et al. (2023); Peng et al. (2023b;a); Fan et al. (2024); Peng et al. (2025) has gradually received more attention. These methods have generally focused on predictions based on mesh representations or parameterized models to enhance realism. For mesh-based methods, FaceFormer Fan et al. (2022) introduces a transformer-based approach for generating 3D facial animations, using autoregressive modeling to capture long-term audio context. CodeTalker Xing et al. (2023) utilizes discrete motion priors learned from real facial movements, applying a vector quantized autoencoder to reduce uncertainty in the audio-to-motion mapping process. For parameterized methods, ARKit-like blendshape models Lewis et al. (2014) are commonly used. EmoTalk Peng et al. (2023b) disentangles emotion and content from speech to generate expressive facial movements, while DuelTalk Peng et al. (2025) supports multi-round, dual-speaker interactions. UniTalker Fan et al. (2024) combines mesh-based and parameterized annotation styles to improve scalability, and xADA Taylor et al. (2025) generates controllable face, tongue, and head animation directly from input speech. These systems study the standalone audio-to-animation mapping. Ex-Omni instead studies how facial animation can be generated natively alongside a newly reasoned text and speech response inside an OLLM.

Speech Language Models. Recent advancements in speech language models Fang et al. (2025a;b); Chen et al. (2025a); Zhang et al. (2023); Hassid et al. (2023); Chu et al. (2024); Chen et al. (2025b); Xie & Wu (2024a) have significantly improved speech understanding and generation in an end-to-end manner, eliminating the need for cascaded ASR and TTS models. For example, SpeechT5 Ao et al. (2022) aligns text and speech representations into a shared semantic space using a unified encoder-decoder structure and cross-modal vector quantization methods. Moshi Défossez et al. (2024) addresses the issues of latency and information bottlenecks in traditional speech dialogue systems through its full-duplex speech-to-speech generation framework and the Inner Monologue design. SpeechGPT-Gen Zhang et al. (2024) introduces the Chain-of-Information Generation to decouple the modeling of semantic and perceptual information,thus making the speech generation process more efficient and precise. GLM-4-Voice Zeng et al. (2024) addresses the delay and error accumulation problems by adopting a 12.5Hz speech segmenter, streaming reasoning and large-scale speech-to-text pre-training.

3 METHOD

3.1 OVERVIEW

Figure 1 shows the overall pipeline of Ex-Omni. Given an instruction $\mathcal{I}$ comprising a text input x, a speech waveform a, or both, Ex-Omni performs instruction understanding and multimodal generation within an LLM-centered unified framework, where the LLM focuses on semantic reasoning rather than direct temporal generation. The model produces a discrete speech unit sequence u, which is decoded into waveform speech, as well as a sequence of 3D facial animation parameters y (*i.e.,* blendshape coefficients). At a high level, Ex-Omni adopts a structured decomposition that reduces the difficulty of learning temporally coherent generation by decoupling semantic reasoning from modality-specific temporal synthesis. Specifically, the LLM is responsible for instruction understanding and semantic reasoning, while speech units are used as an explicit temporal scaffolding to guide downstream speech and facial animation generation. We formulate the overall model as

$$(u, y) = \mathcal{F}(\mathcal{I}; \theta), \quad \theta = \{\theta_E, \theta_P, \theta_L, \theta_U, \theta_F\}, \tag{1}$$

where θ_E, θ_P, and θ_L correspond to the speech encoder, speech projector and LLM, respectively. θ_U denotes the speech-unit generator with blendshape co-supervision while θ_F represents the blendshape decoder.

3.2 UNIFIED SPEECH-TEXT REPRESENTATION

Given an instruction $\mathcal{I}$ containing a text input x, a speech waveform a, or both, we map each available input modality into the shared LLM token embedding space. When speech is present, it is first encoded into high-level representations and then projected as

$$X_s = P_{\theta_P}(E_{\theta_E}(a)) \in \mathbb{R}^{T_s \times d}, \tag{2}$$

while the available text tokens are embedded as

$$X_l = \mathrm{Emb}_{\theta_L}(x) \in \mathbb{R}^{T_l \times d}, \tag{3}$$

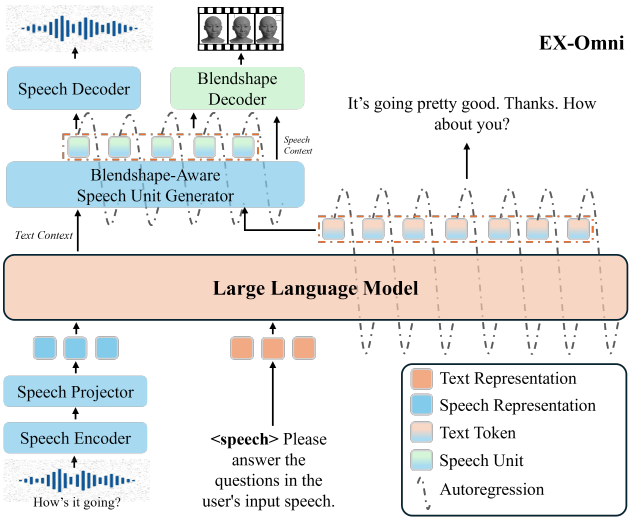

Figure 1: Model architecture of Ex-Omni.

where d denotes the LLM embedding dimension and Emb_{θ_L} is the input token embedding layer of Qwen3-8B (Yang et al., 2025). Depending on the available input modalities, the unified input representation is constructed as

$$X = \begin{cases} X_l, & \text{if } \mathcal{I} \text{ contains text only,} \\ X_s, & \text{if } \mathcal{I} \text{ contains speech only,} \\ [X_l; X_s], & \text{if } \mathcal{I} \text{ contains both modalities,} \end{cases} \qquad X \in \mathbb{R}^{T_X \times d}, \tag{4}$$

where T_X denotes the resulting input sequence length.

3.3 LLM-Centered Reasoning

The LLM serves as a semantic reasoner that focuses exclusively on instruction understanding and high-level reasoning. Given the unified input representation X, the LLM (Qwen3-8B) performs autoregressive generation to produce the text response $t_{1:T_{lr}}$. During this process, we extract the sequence of last hidden states corresponding to the generated response tokens:

$$H = [h_1, h_2, \ldots, h_{T_{lr}}] \in \mathbb{R}^{T_{lr} \times d}, \tag{5}$$

where h_i represents the features at step i containing high-level semantic reasoning information. The probability of the next token is predicted based on these states:

$$p_{\theta_L}(t_{i+1} \mid t_{1:i}, X) = \mathrm{Softmax}(W_o h_i). \tag{6}$$

3.4 Joint Speech and 3D Facial Animation Generation

Ex-Omni jointly generates speech and 3D facial animation, aiming to maintain semantic consistency and temporal alignment across modalities. We adopt token-as-query gated fusion (TQGF), a task-specific cross-attention module in which the target token sequence serves as the query and upstream semantic representations provide contextual key/value states. This design assigns temporal responsibility to the target token sequence and injects semantic cues through a learned gate, preserving the speech-unit temporal scaffold while avoiding direct dense-motion regression from LLM states.

Formally, let $Q \in \mathbb{R}^{M \times d}$ denote query tokens and $C \in \mathbb{R}^{N \times d_c}$ denote context tokens. We first project C to the module width d when needed. The gated fusion operation is

$$\mathrm{Fuse}(Q, C) = Q + \sigma(QW_g + b_g) \odot \mathrm{Attn}(Q, P_C(C)), \tag{7}$$

where $\mathrm{Attn}(Q, P_C(C))$ is cross-attention from Q to the projected context, W_g and b_g produce element-wise gates with the same hidden width as the attention output, and $\sigma(\cdot)$ is the sigmoid

function. In multi-head attention the gate is applied per token and broadcast over the corresponding head dimensions.

For speech generation, we model speech synthesis as an autoregressive prediction of discrete speech units using a speech-unit generator with blendshape co-supervision (Qwen3-0.6B). Given the generated text tokens $t_{1:T_{lr}}$, the speech-unit generator with blendshape co-supervision enriches the explicit token embeddings with the semantic hidden states H through:

$$\tilde{H} = \mathrm{Fuse}_{\theta_U}\left(\mathrm{Emb}_{\theta_U}(t_{1:T_{lr}}), P_U(H)\right), \tag{8}$$

where Emb_{θ_U} denotes the token embedding layer adopted from GLM-4-Voice, P_U maps LLM hidden states to the unit-generator width, and $\tilde{H} \in \mathbb{R}^{T_{lr} \times d}$ serves as the conditioning signal. Then, the speech-unit generator with blendshape co-supervision predicts a sequence of speech units $u_{1:T_u}$:

$$p_{\theta_U}(u_{1:T_u} \mid \tilde{H}) = \prod_{j=1}^{T_u} p_{\theta_U}(u_j \mid u_{<j}, \tilde{H}). \tag{9}$$

For 3D facial animation generation, we parameterize facial motion using ARKit-52 blendshape coefficients and formulate S2F generation as a non-autoregressive sequence prediction problem, where the model outputs the full blendshape coefficients $\hat{y}_{1:T_y} \in \mathbb{R}^{T_y \times 52}$ in parallel. Given discrete speech units $u_{1:T_u}$, we first obtain unit embeddings and align them to the target video frame rate by temporally resampling features to length T_y (linear interpolation in our implementation), yielding frame-level query representations $Q_y \in \mathbb{R}^{T_y \times d}$. In parallel, we project the hidden states of the speech-unit generator with blendshape co-supervision into the same space to obtain contextual key/value representations $S \in \mathbb{R}^{T_u \times d}$, which carry semantically rich information. We then inject speech semantics into the frame-level queries using the TQGF module:

$$H_f = \mathrm{Fuse}_{\theta_F}(Q_y, S) \in \mathbb{R}^{T_y \times d}. \tag{10}$$

We then refine $\mathbf{H}_f$ using a non-autoregressive Transformer decoder with rotary positional embeddings (Su et al., 2024) in its self-attention layers to obtain the predicted sequence of 3D facial parameters:

$$\hat{y}_{1:T_y} = D_{\theta_F}(H_f), \quad \hat{y}_t \in \mathbb{R}^{52}. \tag{11}$$

3.5 Training Objectives

Autoregressive Objectives for Text and Speech. For text tokens and discrete speech units, we adopt standard autoregressive modeling. Given a token sequence $z = (z_1, \ldots, z_T)$, the objective is $\mathcal{L}_{\text{ar}} = -\sum_{t=1}^{T} \log p(z_t \mid z_{<t})$, where z_t denotes either a text token or a speech unit, depending on the supervision available for a given sample. In practice, the text-token objective supervises ASR transcripts and response text in S2SF samples, while the speech-unit objective is applied to both TTSF and S2SF samples.

3D Facial Animation Generation. For 3D facial animation, the blendshape decoder is trained with a frame-wise regression loss. Let $\hat{y}_t \in \mathbb{R}^K$ and $y_t \in \mathbb{R}^K$ denote the predicted and ground-truth blendshape annotations at frame t, respectively. We define the facial loss as $\mathcal{L}_{\text{bs}} = \frac{1}{B}\sum_{i=1}^{B} \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \left\| \hat{y}_t^{(i)} - y_t^{(i)} \right\|_2^2$, where $\mathcal{T}_i$ denotes the valid temporal range determined by the target sequence length.

3.6 Training Strategy

Stage I (Speech-to-Text Pretraining). We train the speech projector on ASR data while freezing all other components. This stage aligns speech representations with the semantic space of the base LLM.

Stage II (Speech-Blendshape Co-pretraining). We pre-train the speech-unit generator and the blendshape decoder on 1M TTSF samples while freezing the base LLM and unrelated modules. This stage introduces joint speech and blendshape supervision to establish speech-blendshape alignment; the unit generator becomes blendshape-aware through gradients from the jointly trained decoder rather than through a separate explicit blendshape target.

Stage III (Dialogue-based Speech-Blendshape Co-pretraining). The LLM, speech projector, speech-unit generator with blendshape co-supervision, and blendshape decoder are jointly optimized on a mixture of ASR, TTSF, and S2SF data, while the speech encoder and speech decoder remain frozen. Text autoregression, speech-unit autoregression, and blendshape regression are computed in one masked graph and applied only when the corresponding labels are present. Gradients update the blendshape decoder and speech-unit generator with blendshape co-supervision and, through their conditioning states, the unfrozen LLM and speech projector at substantially smaller learning rates. Ex-Omni predicts speech units and blendshapes; the frozen waveform decoder is not trained. This stage adapts the model to dialogue-oriented speech-blendshape generation.

4 Data Construction

The data statistics of the full training corpus used by Ex-Omni are summarized in Table 1. Overall, the training data consist of an external ASR corpus together with our proposed InstructS2SF-1200K. Large-scale paired text, speech, and high-quality 3D facial motion for open-domain OLLM training is difficult to collect, so we use Audio2Face-3D (Chung et al., 2025) pseudo-labels as weak supervision. This follows the broader use of synthesized targets as a practical scaling strategy in OLLM training Fang et al. (2025a); Luo et al. (2025). Because speech-driven facial motion is inherently one-to-many, these pseudo-labels provide a consistent temporal supervision signal rather than a unique facial-motion ground truth. More details can be found in Appendix A.

Table 1: Statistics of the three-stage training corpus.

Stage	Type	Duration (Hour / Second)					
		Total	Mean	Median	Min	Max	Std
I	ASR	2113.38	10.47	10.46	0.83	30.00	4.81
II	TTSF	2814.57	10.13	8.68	0.40	161.80	5.71
III	S2SF (Prompt)	782.73	14.09	12.68	1.16	75.36	6.61
	S2SF (Response)	1434.67	25.82	23.04	0.40	296.40	14.31
	ASR	25.91	9.33	8.07	3.00	30.00	5.01
	TTSF	28.38	10.22	8.72	0.64	52.20	5.76

5 Experiments

5.1 Evaluation

Speech-/Text-to-Face Evaluation. For S2F, evaluations are conducted on AlpacaEval Li et al. (2023) and CommonEval subsets of VoiceBench (Chen et al., 2024). We use Sync-C and Sync-D from SyncNet Chung & Zisserman (2016); Prajwal et al. (2020) (higher Sync-C and lower Sync-D indicate better alignment) for lip-speech synchronization. SyncNet is a widely used third-party reference model for audio-visual synchronization evaluation and is independent of the Audio2Face-3D teacher used for pseudo-label generation. Lip Vertex Error (LVE) requires a paired ground-truth facial-motion sequence, which open-domain generated dialogue does not provide; computing LVE against a teacher output would turn the pseudo-label source into the metric target. We therefore combine SyncNet with a controlled downstream comparison, a human preference study, temporal motion diagnostics, and qualitative videos. T2F evaluation follows the same protocol as S2F evaluation, except that the input is text rather than speech.

Speech QA Evaluation. We evaluate speech QA on VoiceBench Chen et al. (2024), which covers a diverse set of speech-based tasks, including open-ended question answering, reference-based QA, multiple-choice QA, reasoning, instruction following and safety. Open-ended QA is evaluated using GPT-based scores (scores from 1-5), while other tasks are evaluated using accuracy-based metrics. All the evaluations were conducted using the open-source code of VoiceBench to ensure consistency.

Speech Generation Evaluation. For speech generation evaluation, we assess both response-level speech quality and speech-text consistency on the AlpacaEval and CommonEval speech QA sets. Specifically, we compute UTMOS on the generated speech, and transcribe the generated waveform using Whisper-Large-V3 Radford et al. (2023). The resulting ASR transcript is then compared with the model's textual response using Word Error Rate (WER), which measures consistency between spoken and textual outputs.

Baselines For S2F evaluation, we compare Ex-Omni with two recent S2F methods, EmoTalk Peng et al. (2023b) and UniTalker Fan et al. (2024), both of which support direct prediction of facial blend-

Table 2: End-to-end comparison with non-teacher cascaded OLLM+S2F systems. Cascaded methods pass generated speech to an external S2F model, whereas Ex-Omni jointly generates speech and facial animation; Table 3 isolates the downstream module under fixed speech. The best result in each column is shown in **bold**.

Method	Speech-to-Face (Sync-D ↓ / Sync-C ↑)		Text-to-Face (Sync-D ↓ / Sync-C ↑)	
	AlpacaEval	CommonEval	AlpacaEval	CommonEval
Qwen2.5-Omni-3B + EmoTalk	10.538 / 3.305	10.750 / 3.203	10.643 / 3.166	10.830 / 3.106
Qwen2.5-Omni-3B + UniTalker-B-D3	10.449 / 3.222	10.632 / 3.156	10.544 / 3.153	10.494 / 3.419
Qwen2.5-Omni-3B + UniTalker-B-D6	9.873 / 3.738	10.088 / 3.600	9.945 / 3.629	10.302 / 3.504
Qwen2.5-Omni-7B + EmoTalk	10.554 / 3.273	11.215 / 3.628	10.513 / 3.381	11.180 / 3.641
Qwen2.5-Omni-7B + UniTalker-B-D3	10.494 / 3.200	10.854 / 3.767	10.520 / 3.319	10.883 / 3.675
Qwen2.5-Omni-7B + UniTalker-B-D6	10.176 / 3.405	10.856 / 4.012	10.115 / 3.641	10.833 / 3.980
Ex-Omni	**9.233 / 5.385**	**9.212 / 5.363**	**9.313 / 5.236**	**9.239 / 5.403**

Table 3: Controlled downstream comparison with a fixed Ex-Omni response. All four cascades receive identical Ex-Omni-generated speech; Ex-Omni uses its jointly trained facial-animation branch. Audio2Face-3D is the pseudo-label teacher. Best and second-best results are shown in **bold** and underline, respectively.

Method	Speech-to-Face (Sync-D ↓ / Sync-C ↑)		Text-to-Face (Sync-D ↓ / Sync-C ↑)	
	AlpacaEval	CommonEval	AlpacaEval	CommonEval
Ex-Omni + Audio2Face-3D	**8.928 / 5.775**	**8.921 / 5.699**	**8.913 / 5.669**	**8.896 / 5.741**
Ex-Omni + EmoTalk	10.758 / 4.321	10.416 / 3.710	10.534 / 4.319	10.418 / 3.490
Ex-Omni + UniTalker-B-D3	10.946 / 3.882	11.117 / 2.900	10.893 / 3.708	10.792 / 3.130
Ex-Omni + UniTalker-B-D6	10.701 / 4.343	10.668 / 3.516	10.786 / 4.049	10.334 / 3.771
Ex-Omni	9.233 / 5.385	9.212 / 5.363	9.313 / 5.236	9.239 / 5.403

shape coefficients. We additionally include Ex-Omni+Audio2Face-3D as a teacher-cascade baseline because Audio2Face-3D supplies the weak animation supervision used to train Ex-Omni. All Ex-Omni cascades receive exactly the same generated text and speech as Ex-Omni, isolating the facial animation module. For speech QA evaluation, we compare Ex-Omni with several representative OLLMs and speech large language models. Specifically, the baselines include Qwen2.5-Omni Xu et al. (2025), VITA-1.0 Fu et al. (2024), VITA-1.5 Fu et al. (2025), Mini-Omni Xie & Wu (2024a), Mini-Omni2 Xie & Wu (2024b), Moshi Défossez et al. (2024), SLAM-omni Chen et al. (2025b), and LLaMA-Omni Fang et al. (2025a). For speech generation evaluation, we compare Ex-Omni with Qwen2.5-Omni under the same response-generation setting, and report both UTMOS (Saeki et al., 2022) and speech-text consistency measured by WER.

5.2 Experimental Results and Analysis

3D Facial Animation Generation Results. Table 2 first compares complete Ex-Omni outputs with non-teacher cascaded OLLM+S2F systems. Because those cascades use a different upstream speech generator, this table characterizes end-to-end pipeline behavior rather than isolating the facial-animation module. The controlled comparison in Table 3, which fixes the upstream Ex-Omni response, is the primary evidence for the native branch. The Audio2Face-3D cascade is expected to rank first because it directly invokes the model that supplies our pseudo-labels. Ex-Omni consistently ranks second and is substantially closer to this teacher cascade than to the other external S2F models. On CommonEval S2F, for example, Ex-Omni differs from the teacher by only 0.291 Sync-D and 0.336 Sync-C; relative to the strongest non-teacher cascade, it improves Sync-D by 1.204 and Sync-C by 1.653. The result indicates that pseudo-label supervision transfers much of the teacher's synchronization behavior into Ex-Omni's native branch, without requiring an external S2F model at inference. The human study in Table 4 supports the same conclusion: the teacher cascade receives 55.8% of first-place preference, Ex-Omni receives 41.8%, and the three non-teacher cascades together receive only 2.4%. Combined with the face-latency results in Table 9, Ex-Omni

Table 4: Human preference on CommonEval S2F. Tied first-place votes are split evenly; tie rates are reported separately.

Method	**Preference ↑**	**Tie**
Ex-Omni + Audio2Face-3D	**55.8%**	21.1%
Ex-Omni	41.8%	21.6%
Ex-Omni + EmoTalk	1.9%	1.8%
Ex-Omni + UniTalker-B-D3	0.3%	1.3%
Ex-Omni + UniTalker-B-D6	0.2%	1.1%

approaches teacher-cascade synchronization while avoiding the need for an external S2F model at inference.

Table 5: Speech QA comparison on VoiceBench under its standard speech-input setting. ↑ means higher is better. ∗ denotes results reproduced using open-source code.

Model	AlpacaEval ↑	CommonEval ↑	WildVoice ↑	SD-QA ↑	MMSU ↑	OBQA ↑	BBH ↑	IFEval ↑	AdvBench ↑	Overall ↑
Qwen2.5-Omni-7B	**4.49**	**3.93**	2.71*	**55.71**	**61.32**	**81.10**	60.80*	52.87	**99.42**	**70.42***
Moshi	2.01	1.60	1.30	15.64	24.04	25.93	47.40	10.12	44.23	29.51
VITA-1.0	3.38	2.15	1.87	27.94	25.70	29.01	47.70	22.82	26.73	56.48
VITA-1.5	4.21	3.66	3.48	38.88	52.15	71.65	55.30	38.14	97.69	64.53
LLaMA-Omni	3.70	3.46	2.92	39.69	25.93	27.47	49.20	14.87	11.35	41.12
Mini-Omni	1.95	2.02	1.61	13.92	24.69	26.59	46.30	13.58	37.12	30.42
Mini-Omni2	2.32	2.18	1.79	9.31	24.27	26.59	46.40	11.56	57.50	33.49
SLAM-Omni	1.90	1.79	1.60	4.16	26.06	25.27	48.80	13.38	94.23	35.30
Ex-Omni	4.31	3.82	**3.49**	50.71	46.03	56.70	**61.10**	**55.72**	87.12	65.53

Speech QA Results. Table 5 evaluates whether Ex-Omni retains speech QA capability after introducing native speech–face generation. Ex-Omni achieves the second-best overall score of 65.53, outperforming all compared systems except Qwen2.5-Omni-7B, and ranks first on WildVoice, BBH, and IFEval. This performance is notable because Ex-Omni additionally generates speech-synchronized 3D facial animation, a capability absent from the compared systems. Its remaining gaps on MMSU, OBQA, and AdvBench suggest room for broader knowledge- and safety-oriented instruction tuning, whereas the current 200K-sample S2SF training stage primarily focuses on introducing the new generation modality. Overall, Ex-Omni adds native non-verbal generation while retaining competitive speech understanding and reasoning capabilities.

Speech Generation Results. Table 6 evaluates the quality of generated speech and its consistency with the corresponding textual response. Despite jointly learning speech and facial-animation generation, Ex-Omni achieves the best UTMOS on AlpacaEval (4.523) and comparable quality on CommonEval (4.491). Its S-T WER is competitive on AlpacaEval and substantially lower on CommonEval (3.54 versus 20.10–21.72), indicating reliable response-level speech–text alignment. This CommonEval gap is partly affected by response length: in our obversation, Qwen2.5-Omni often produces longer responses, some lasting approximately 90 seconds, while speech beyond roughly 60 seconds is more susceptible to truncation or late-stage errors. Overall, Ex-Omni maintains strong speech quality and speech–text fidelity while additionally generating synchronized facial animation.

Table 6: Speech response quality and speech–text consistency (S-T Consis.) on AlpacaEval and CommonEval. UTMOS is computed on generated speech, while S-T Consis. is measured by WER after transcribing the generated audio with Whisper-Large-V3 and comparing it with the corresponding textual response.

Method	AlpacaEval		CommonEval	
	UTMOS ↑	S-T Consis. ↓	UTMOS ↑	S-T Consis. ↓
Qwen2.5-Omni-3B	4.174	**33.53**	4.495	20.10
Qwen2.5-Omni-7B	4.502	55.11	**4.525**	21.72
Ex-Omni	**4.523**	34.22	4.491	**3.54**

Effect of TQGF. Table 7 evaluates the TQGF interface against single-source conditioning and direct feature concatenation. The full model performs best across all metrics. Using only speech tokens reduces Sync-C by 0.559, showing that the temporal scaffold alone lacks sufficient facial context, while using only hidden features reduces Sync-C by 0.617, indicating that contextual representations cannot replace explicit speech-unit timing. Direct concatenation performs worst, increasing Sync-D by 0.976 and reducing Sync-C by 1.342. These results demonstrate that speech tokens and hidden features provide complementary temporal and contextual cues, which TQGF integrates more effectively while preserving response and speech quality. In our training runs, we additionally ob-

Table 7: Effect of TQGF on AlpacaEval. QA Score denotes the GPT-based open-ended response score on a 1–5 scale. ↑/↓ indicates that higher/lower is better.

Method	Sync-D↓ / Sync-C↑	QA Score↑	UTMOS↑
Ex-Omni	**9.233 / 5.385**	**4.31**	**4.523**
w/o TQGF (only speech tokens)	9.591 / 4.826	4.29	4.161
w/o TQGF (only hidden features)	9.475 / 4.768	4.23	3.968
w/o TQGF (concatenation)	10.209 / 4.043	4.27	4.136

served slightly faster loss convergence when using the learned gate than with the otherwise identical ungated fusion module.

Ablation Study on Input Modality and Knowledge Retention. As shown in Table 8, the text-to-text comparison isolates knowledge retention after multimodal training: Ex-Omni decreases by only 0.12 on both AlpacaEval and CommonEval relative to its Qwen3-8B initialization. The additional gap under speech input also includes speech encoding and cross-modal alignment, and therefore should not be interpreted as degradation of the LLM's text reasoning alone. This ablation shows that Ex-Omni retains most of the base model's text capability while acquiring speech understanding and native facial-animation generation.

Table 8: Input-modality ablation on the open-ended QA subsets. ↑ means higher is better.

Model	Input	AlpacaEval ↑	CommonEval ↑
Qwen3-8B	Text	**4.82**	**4.51**
Ex-Omni	Text	<u>4.70</u>	<u>4.39</u>
Ex-Omni	Speech	4.31	3.82

Table 9: Results of latency analysis.

Metric	Text input	Speech input
Thinker TTFT (s) ↓	0.019	0.038
Talker first unit from input (s) ↓	0.478	0.451
Codec decoding (s) ↓	0.081	0.082
Rendering TTFC (s) ↓	0.303	0.305
Text TPS (tokens/s) ↑	101.98	101.27
Speech TPS (units/s) ↑	179.71	179.44
Video FPS (frames/s) ↑	143.69	142.80

Latency Analysis. Table 9 reports average latency and throughput using our adaptation to vLLM-Omni on a single GPU on CommonEval. Measured from when the complete input is available, Thinker TTFT is 0.019s for text input and 0.038s for speech input; the first speech unit is produced after 0.478s and 0.451s, respectively. Decoding the first 10 speech units takes 0.081s and 0.082s, while rendering, encoding, and muxing the first video segment takes 0.303s and 0.305s. Text generation reaches 101.98 and 101.27 tokens/s, and speech generation reaches 179.71 and 179.44 units/s. Video rendering and encoding throughput is 143.69 and 142.80 frames/s, respectively. These results demonstrate that the native facial-animation branch can be executed efficiently within the integrated generation pipeline.

6 Conclusion

In this paper, we introduced Ex-Omni, a framework that extends OLLMs with native speech-accompanied 3D facial animation generation. To address the mismatch between token-level semantic reasoning and fine-grained facial motion, Ex-Omni decouples high-level understanding from modality-specific temporal synthesis through discrete speech-unit scaffolding and a task-specific token-as-query gated fusion interface. We further constructed InstructS2SF-1200K to provide large-scale weak speech–face supervision for both synthesis and dialogue-oriented scenarios. Experiments show that Ex-Omni preserves substantial speech understanding and generation capability, improves over the evaluated non-teacher cascades in controlled downstream comparisons, and approaches the Audio2Face-3D teacher cascade in both SyncNet and human preference while avoiding an external S2F model at inference.

Limitations

Ex-Omni has four main limitations. First, long speech responses produce lengthy frame-level facial sequences, increasing computational and memory costs and challenging long-range consistency. Second, the overall real-time factor remains above 1.0, and the latency advantage is limited to the facial-animation branch rather than the complete pipeline. Third, the compact ARKit-52 representation may underrepresent subtle affective and person-specific facial motions Chung et al. (2025). Finally, reliance on Audio2Face-3D pseudo-labels may introduce teacher-specific biases.

AI USE STATEMENT

Generative AI tools were used for source-code review and language polishing. All AI-assisted suggestions were independently checked against the implementation and experimental results. The authors take full responsibility for the content.

ETHICS STATEMENT

Ex-Omni generates speech and speech-synchronized 3D facial animation, which could be misused to create misleading synthetic avatar content. Deployment should clearly disclose synthetic outputs, obtain consent when identifiable voices or avatars are involved, and adopt appropriate provenance and access controls. Potential biases across languages, accents, speaking styles, and facial templates should also be evaluated before deployment.

REPRODUCIBILITY STATEMENT

The paper and appendix specify the model architecture, training objectives, data construction, component provenance, hyperparameters, and evaluation protocols. The supplementary material provides code to facilitate reproduction. Reconstructing the training corpus additionally requires the external synthesis and pseudo-labeling tools described in the appendix. Code is also provided in the supplementary material to facilitate reproduction.

REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, Guangming Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Libin Wang, Lixiang Ru, Lele Xie, Longhua Tan, Lyuxin Xue, Lan Wang, Mochen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qinglong Zhang, Qiang Xu, Rui Liu, Ruijie Xiong, Sirui Gao, Tinghao Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yunxiao Sun, Yipeng Chen, Yifei Wu, Yongjie Lyu, Ziping Ma, Zipeng Feng, Zhijiang Fang, Zhihao Qiu, Ziyuan Huang, and Zhengyu He. Ming-omni: A unified multimodal model for perception and generation. *CoRR*, abs/2506.09344, 2025.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5723–5738, 2022.

Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7832–7841, 2019.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, Yabin Li, Xiang Lv, Jiaqing Liu, Haoneng Luo, Bin Ma, Chongjia Ni, Xian Shi, Jialong Tang, Hui Wang, Hao Wang, Wen Wang, Yuxuan Wang, Yunlan Xu, Fan Yu, Zhijie Yan, Yexin Yang, Baosong Yang, Xian Yang, Guanrou Yang, Tianyu Zhao, Qinglin Zhang, Shiliang Zhang, Nan Zhao, Pei Zhang, Chong Zhang, and Jinren Zhou. Minmo: A multimodal large language model for seamless voice interaction. *CoRR*, abs/2501.06282, 2025a. doi: 10.48550/ARXIV.2501.06282.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, Kai Yu, Yuxuan Hu, Jinyu Li, Yan Lu, Shujie Liu, and Xie Chen. Slam-omni: Timbre-controllable voice interaction system with single-stage training. In *Findings of the Association for Computational Linguistics*, pp. 2262–2282, 2025b.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *CoRR*, abs/2410.17196, 2024.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *CoRR*, abs/2407.10759, 2024.

Chaeyeon Chung, Ilya Fedorov, Michael Huang, Aleksey Karmanov, Dmitry Korobchenko, Roger Blanco i Ribera, and Yeongho Seol. Audio2face-3d: Audio-driven realistic facial animation for digital avatars. *CoRR*, abs/2508.16401, 2025.

J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

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.

Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through A unified model. In *Computer Vision - ECCV 2024 - 18th European Conference*, volume 15099, pp. 204–221, 2024.

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*, pp. 18749–18758, 2022.

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*. OpenReview.net, 2025a.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. Llama-omni 2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 18617–18629. Association for Computational Linguistics, 2025b.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. VITA: towards open-source interactive omni multimodal LLM. *CoRR*, abs/2408.05211, 2024.

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.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. Textually pretrained speech language models. In *The Thirty-Seventh Annual Conference on Neural Information Processing Systems*, 2023.

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 *Proc. of SLT*, 2024.

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: A large-scale, extensive, multilingual, and diverse dataset for speech generation. In *arXiv:2501.15907*, 2025.

Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3387–3396, 2022.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

John P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frédéric H. Pighin, and Zhigang Deng. Practice and theory of blendshape facial models. In *35th Annual Conference of the European Association for Computer Graphics*, pp. 199–218, 2014.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models, 2023.

Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, et al. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data. *arXiv preprint arXiv:2511.12609*, 2025.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, Xiaobo Xia, Hamid Alinejad-Rokny, Fei Huang, and Min Yang. Openomni: Advancing open-source omnimodal large language models with progressive multimodal alignment and real-time emotional speech synthesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 3279–3287, 2020.

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*, pp. 5206–5210, 2015.

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

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *IEEE/CVF International Conference on Computer Vision*, pp. 20630–20640, 2023b.

Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21055–21064, 2025.

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 *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 484–492, 2020. ISBN 9781450379885.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, 2023.

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *2021 IEEE/CVF International Conference on Computer Vision*, pp. 1153–1162, 2021.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In Hanseok Ko and John H. L. Hansen (eds.), *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pp. 4521–4525. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-439. URL https://doi.org/10.21437/Interspeech.2022-439.

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

Sarah Taylor, Salvador Medina, Jonathan Windle, Erica Alcusa Sáez, and Iain Matthews. xada: Controllable and expressive audio-driven animation. In *Proceedings of the ACM SIGGRAPH 2025 Conference Papers*, 2025. doi: 10.1145/3721238.3730711.

Meituan LongCat Team. Longcat-flash-omni technical report. *CoRR*, abs/2511.00279, 2025.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming. *CoRR*, abs/2408.16725, 2024a.

Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *CoRR*, abs/2410.11190, 2024b.

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

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.

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.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. WENETSPEECH: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022*, pp. 6182–6186, 2022.

Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. FACIAL: synthesizing dynamic talking face with implicit attribute learning. In *2021 IEEE/CVF International Conference on Computer Vision*, pp. 3847–3856, 2021.

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 *Findings of the Association for Computational Linguistics*, pp. 15757–15773, 2023.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechgpt-gen: Scaling chain-of-information speech generation. *CoRR*, abs/2401.13527, 2024. doi: 10.48550/ARXIV.2401.13527.

A Implementation Details

Data Construction. In Stage I, about 720K ASR samples are primarily sampled from the Emilia (He et al., 2025) corpus. We further incorporate the train-clean-100, train-clean-360, and train-other-500 subsets of LibriSpeech Panayotov et al. (2015), together with WenetSpeech Zhang et al. (2022) training data whose confidence scores exceed 0.95.

In Stage II, we construct 1M TTSF samples. Specifically, text prompts sampled from Emilia He et al. (2024; 2025) are converted into speech using Qwen3-TTS Hu et al. (2026) with a single, unified speaker identity. High-quality motion-capture facial animation data are scarce, and existing public datasets typically contain only a few thousand samples with limited coverage of speech content. We therefore adopt NVIDIA Audio2Face-3D Chung et al. (2025), an open-weights model trained on large-scale professionally captured facial motion, as a teacher to generate blendshape pseudo-labels from the synthesized speech. Concretely, we first run the off-the-shelf NVIDIA Audio2Emotion estimator Chung et al. (2025) on the synthesized audio to obtain a 5D compound emotion vector over anger, disgust, sadness, joy, and fear, where an all-zero vector is treated as neutral. Audio2Emotion is used only for dataset construction and is not trained by us. This emotion vector is then used as the conditional input to Audio2Face-3D to generate the corresponding blendshape sequence. These teacher-generated blendshape sequences are used as weak supervision signals for speech-blendshape co-pretraining, rather than being treated as deterministic facial ground truth.

In Stage III, we construct 200K dialogue-based S2SF samples based on InstructS2S-200K Fang et al. (2025a). To improve speech quality and consistency, we reconstruct the target speech using Qwen3-TTS, and then use Audio2Face-3D Chung et al. (2025) to generate corresponding blendshape pseudo-labels. As in Stage II, we first estimate a 5D compound emotion vector from the audio with Audio2Emotion and use it as the conditional input to Audio2Face-3D. In practice, we observe that most speech responses in QA-style dialogue are close to neutral under this pipeline, so the resulting pseudo-labels mainly emphasize accurate mouth articulation and lip-speech synchronization rather than exaggerated facial expressions. As in Stage II, these teacher-generated blendshape sequences serve as weak supervision signals, which is a more appropriate formulation given the one-to-many nature of speech-to-blendshape mapping. To preserve capabilities acquired in earlier stages, we additionally include 10K ASR replay data and 10K TTSF replay data during this stage. Standalone audio-to-animation systems such as xADA Taylor et al. (2025) address a different optimization setting from Ex-Omni, which jointly learns text generation, speech-unit generation, and facial regression from heterogeneous supervision. In reduced-data pilot experiments, sparsifying the animation-supervised subset caused larger loss fluctuations and unstable joint optimization. We therefore retain dense pseudo-label coverage during speech–blendshape co-pretraining; this choice concerns supervision scale rather than treating the pseudo-labels as unique ground truth.

Hyperparameters. All experiments are conducted on a machine equipped with 8 NVIDIA H20 GPUs, each with 96 GB of memory. We use CUDA 12.6, PyTorch 2.7.0 and Python 3.10 for model training and evaluation. The detailed hyperparameters of the three-stage training schedule are shown in Table 10.

Component Provenance. The speech encoder is Whisper-Large-V3 (Radford et al., 2023) and remains frozen throughout training. A two-layer MLP speech projector maps its downsampled features into the Qwen3-8B (Yang et al., 2025) embedding space. The speech-unit generator with blendshape co-supervision uses a Qwen3-0.6B (Yang et al., 2025) backbone, while discrete unit tokenization and waveform decoding follow GLM-4-Voice. The GLM-4-Voice (Zeng et al., 2024) waveform decoder remains frozen, and no waveform-level decoder is trained by Ex-Omni. As clarified in the main text, Emb_{θ_L} and Emb_{θ_U} are distinct input embedding layers associated with the Qwen3-8B LLM and the GLM-4-Voice-derived unit-generation pathway, respectively.

Convergence and Early Stopping. Stage II contains 1M samples totaling 2814.57 hours, so two epochs already correspond to large-scale co-pretraining. Its training loss decreases smoothly and reaches a plateau. Stage III uses a much smaller LLM learning rate together with ASR and TTSF replay. Its mixed-objective loss fluctuates because different batches expose different supervision masks, but the fluctuation amplitude decreases and the loss stabilizes. We therefore stop Stage III

Table 10: The detailed training setup for Ex-Omni across the three training stages.

Hyperparameter	I	II	III
epoch	1	2	2
effective batch size	128	128	32
optimizer	AdamW	AdamW	AdamW
warmup ratio	0.3	0.1	0.1
Gradient accumulation	1	2	4
lr of Speech Encoder	0	0	0
lr of Speech Projector	1×10^{-3}	0	2×10^{-5}
lr of LLM	0	0	2×10^{-6}
lr of Blendshape-Aware Speech Unit Generator	0	1×10^{-4}	2×10^{-5}
lr of blendshape decoder	0	1×10^{-4}	2×10^{-5}
lr of Speech Decoder	0	0	0
freeze Speech Encoder	✓	✓	✓
freeze Speech Projector	✗	✓	✗
freeze LLM	✓	✓	✗
freeze Blendshape-Aware Speech Unit Generator	✓	✗	✗
freeze blendshape decoder	✓	✗	✗
freeze Speech Decoder	✓	✓	✓

after two epochs, as further joint adaptation begins to degrade capabilities retained from the base LLM, consistent with the text-only QA trade-off in Table 8.

Human Evaluation Protocol. We use 200 CommonEval cases and ask 5 participants to compare five anonymized animations generated from the same speech response. Method order is randomized for each case. Evaluators are asked to select the animation with the best overall speech–face synchronization and naturalness, with ties allowed. Each case is evaluated by 5 raters. Tied first-place votes are split evenly when computing preference rates.

B Additional Experimental Results

B.1 Direct Facial Animation Generation from LLM Features

To examine the role of the speech-unit intermediate representation, we compare Ex-Omni with a *Direct LLM-to-Face* variant that predicts facial animation directly from LLM hidden features, bypassing the speech-unit intermediate representation in the facial-generation pathway. Under the same training-time budget, we observed difficulty in achieving convergence for this direct variant, resulting in poor lip-speech synchronization. As shown in Table 11, Ex-Omni achieves substantially lower Sync-D and higher Sync-C on CommonEval under both speech and text inputs. These results suggest that the speech-unit intermediate representation facilitates optimization and helps bridge token-level LLM semantics and fine-grained facial motion.

Table 11: Direct LLM-to-Face ablation on CommonEval under the same training-time budget. Each cell reports Sync-D ↓ / Sync-C ↑. Best results are shown in bold.

Method	Speech-to-Face	Text-to-Face
Direct LLM-to-Face	16.395 / 0.292	16.374 / 0.323
Ex-Omni	**9.212 / 5.363**	**9.239 / 5.403**

Temporal Motion Diagnostics. To supplement SyncNet and human preference, Table 12 reports first- and second-order temporal differences of the generated blendshape sequences. For blendshape vector y_t, velocity and acceleration are defined as $v_t = y_t - y_{t-1}$ and $a_t = v_t - v_{t-1}$. The ABS columns average coefficient-wise absolute magnitudes, and the L_2 columns average vector norms. These values describe motion scale and temporal variation; a very small value may indicate either smooth motion or under-animation. Ex-Omni exhibits larger acceleration magnitudes than the evaluated cascades, indicating greater frame-to-frame variation in the predicted coefficients.

Table 12: Velocity and acceleration diagnostics for generated ARKit-52 blendshape sequences. These statistics characterize motion dynamics and should be interpreted together with SyncNet, human preference, and qualitative results.

Method	Vel-Abs	Vel-L_2	Acc-Abs	Acc-L_2
Ex-Omni + UniTalker-B-D3	0.006834	0.074958	0.005849	0.061824
Ex-Omni + UniTalker-B-D6	0.006403	0.078124	0.005402	0.062921
Ex-Omni + EmoTalk	0.012511	0.141280	0.010173	0.110909
Ex-Omni + Audio2Face-3D	0.005135	0.080290	0.002846	0.043408
Ex-Omni	0.009687	0.138971	0.009532	0.131706

Case Study of 3D Facial Animation Generation. We provide qualitative examples in Appendix Figure 2, where three representative cases are rendered with two different templates: the NVIDIA-provided Claire template and a commercially purchased template with realistic skin and hair materials. We compare Ex-Omni with three cascaded variants, namely Ex-Omni+EmoTalk, Ex-Omni+UniTalker-B-D3, and Ex-Omni+UniTalker-B-D6. To isolate the effect of the facial animation module, all cascaded variants use the same text and speech responses generated by Ex-Omni as input, ensuring that the linguistic content, speech duration, and acoustic realization are controlled across methods. Therefore, the visual differences mainly reflect the quality of the predicted facial animation rather than differences in upstream response generation. As shown in the figure, Ex-Omni captures finer articulatory details more consistently than the cascaded alternatives. In Case 1, when pronouncing "quote", the mouth generated by Ex-Omni moves toward a rounded shape, while the other methods fail to capture this phoneme-related transition. In Case 2, after the speech ends, Ex-Omni closes the mouth naturally, whereas the cascaded methods tend to leave the mouth partially open. In Case 3, both Ex-Omni and Ex-Omni+UniTalker-B-D6 produce a large mouth opening for the pronunciation of "hard", while the other cascaded variants miss this articulation cue. These differences are more evident in the supplementary videos. Overall, the examples illustrate Ex-Omni's ability to capture phoneme-related mouth shapes and utterance-final mouth closure, complementing the quantitative lip-speech synchronization results.

Beyond the three representative examples in Figure 2, the supplementary material contains 20 additional CommonEval examples sampled randomly without manual filtering. Each example compares Ex-Omni with Ex-Omni+EmoTalk, Ex-Omni+UniTalker-B-D3, and Ex-Omni+UniTalker-B-D6. All methods use the same Ex-Omni-generated text and speech responses, so the comparisons isolate facial motion rather than response content, duration, or acoustic realization. The examples are rendered on different face templates to demonstrate that the predicted ARKit-52 coefficients are not tied to one visual identity.

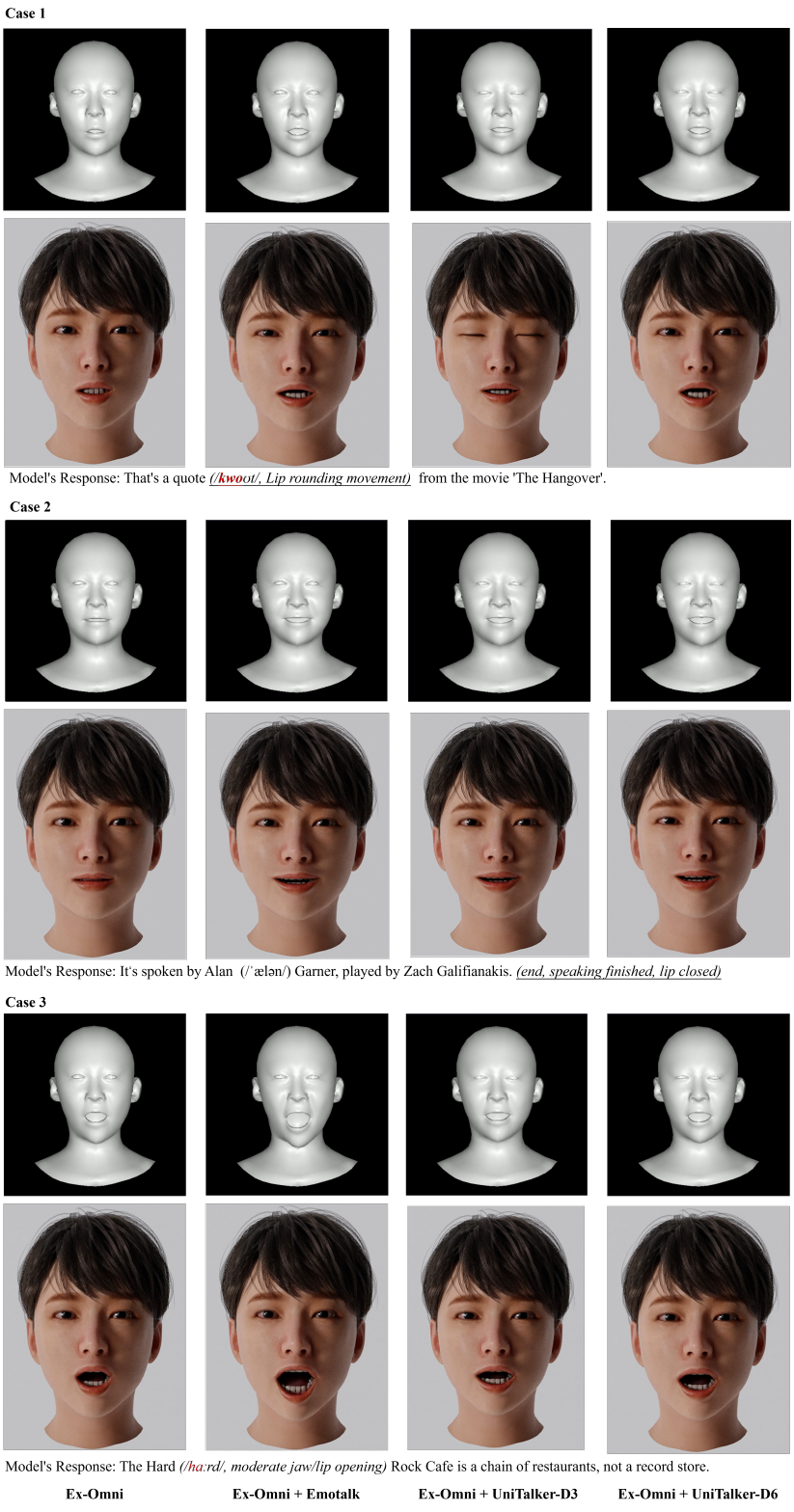

Figure 2: Case study on 3D facial animation generation. Three representative cases are rendered using both the NVIDIA-provided Claire template and a commercially purchased template with realistic skin and hair materials. All cascaded baselines use the same Ex-Omni-generated text and speech responses as input.