HumanOmni-Speaker: Identifying Who said What and When

Detao Bai[1], Xihan Wei[1], and Zhiheng Ma[2,3,4⋆]

[1] Tongyi Lab Alibaba Group
[2] Shenzhen University of Advanced Technology
[3] Guangdong Provincial Key Laboratory of Computility Microelectronics
[4] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
{detao.bdt,xihan.wxh}@alibaba-inc.com, mazhiheng@suat-sz.edu.cn
https://github.com/HumanMLLM/HumanOmni-Speaker

Abstract. While Omni-modal Large Language Models have made strides in joint sensory processing, they fundamentally struggle with a cornerstone of human interaction: deciphering complex, multi-person conversational dynamics to accurately answer "Who said what and when." Current models suffer from an "illusion of competence"—they exploit visual biases in conventional benchmarks to bypass genuine cross-modal alignment, while relying on sparse, low-frame-rate visual sampling that destroys crucial high-frequency dynamics like lip movements. To address this limitation, we introduce Visual-Registered Speaker Diarization and Recognition (VR-SDR) and the HumanOmni-Speaker Benchmark. By strictly eliminating visual shortcuts, this rigorous paradigm demands true end-to-end spatio-temporal identity binding using only natural language queries. To overcome the underlying architectural perception gap, we propose HumanOmni-Speaker, powered by a Visual Delta Encoder. By sampling raw video at 25 fps and explicitly compressing inter-frame motion residuals into just 6 tokens per frame, it captures fine-grained visemes and speaker trajectories without triggering a catastrophic token explosion. Ultimately, HumanOmni-Speaker demonstrates strong multimodal synergy, natively enabling end-to-end lip-reading and high-precision spatial localization without intrusive cropping, and achieving superior performance across a wide spectrum of speaker-centric tasks.

Keywords: Multimodal Large Language Models · Audio-Visual Speaker Diarization · Audio-Visual Speech Recognition · Spatio-Temporal Alignment

1 Introduction

The rapid evolution of Large Language Models into Omni-modal architectures—exemplified by Gemini3 [13], Qwen3-Omni [38], Qwen2.5-Omni [37], LLaMA-Omni [15], OLA [23], and VITA [16]—has enabled the joint processing

⋆ Corresponding author: mazhiheng@suat-sz.edu.cn.

Fig. 1: Speaker-Centric examples in HumanOmni-Speaker benchmark. (top) Visual-Registered Speaker Diarization and Recognition (VR-SDR); (bottom) Four atomic subtasks: Speech Recognition (SR), Speaker Verification (SV), Speaker Identification (SI), and Speaker Localization (SL). Our model unifies visual, audio, and text cues in a single framework.

of visual, audio, and textual signals within a unified semantic space. Yet, as these models transition from isolated inputs to complex real-world environments, seamless human-centric interaction emerges as a critical frontier. At the core of this challenge is the ability to decipher multi-person conversational dynamics by accurately answering: "*Who said what and when*". The significance of this capability is twofold. Practically, it is the cornerstone for deploying autonomous embodied agents [17], meeting diarization systems [27] and wearable AI assistants [17, 18]. Scientifically, it serves as the ultimate stress test for multimodal synergy. It forces models to transcend static semantic matching (e.g., image captioning [34]) and achieve precise cross-modal spatio-temporal alignment—dynamically binding visual identities, continuous acoustic trajectories, and text under multi-speaker interference. While existing models excel at describing static frames [23, 29] or transcribing clear speech [9, 10], they severely struggle in these highly dynamic interactive scenarios.

Why has this fundamental weakness remained largely unaddressed? The answer lies in the deeply flawed evaluation paradigms currently in use. Existing speaker-related tasks [6, 26, 31, 39] and Omni benchmarks [21, 22, 35] isolate the "*Who*", "*When*", and "*What*" into siloed, unimodal atomic tasks. More importantly, they inadvertently allow models to bypass true cross-modal understanding through "shortcut learning" driven by strong visual biases [7]. For instance, a model might correctly identify a speaker simply because they are the central subject in a frame or holding a microphone. By relying on such single-modal visual cues, models achieve high scores without ever aligning the audio stream with subtle, continuous facial movements. This pseudo-multimodal alignment creates an illusion of competence, effectively masking the models' profound inability to perform genuine spatio-temporal identity binding.

To address this limitation, we must establish a rigorous evaluation paradigm that strictly eliminates these shortcuts. Therefore, we introduce **Visual-Registered Speaker Diarization and Recognition (VR-SDR)**, an uncompromising task that forces models to output structured records of identities, timestamps, and transcribed content based solely on audio-visual streams and natural language identity descriptions (as illustrated in Fig. 1). When existing Omni-models are subjected to this stringent standard, their true underlying bottleneck is immediately exposed: a severe architectural perception gap. Current models overwhelmingly rely on sparse visual sampling (typically 1-2 fps)—akin to processing only the static **I-frames** in video compression while discarding the dynamic **P-frames and B-frames**. This fundamentally destroys the high-frequency temporal dynamics, such as visemes and micro-expressions, essential for tracking speakers. Furthermore, simply increasing the frame rate of standard Vision Transformers (ViTs) to recover this data is catastrophic, as it triggers a quadratic token explosion and drowns subtle motion signals in redundant background noise.

To systematically address both the evaluation illusion and the architectural bottleneck, we propose **HumanOmni-Speaker**, a speaker-centric unified Omni-model. The main contributions of this study are reflected in the following three aspects:

1. **Redefining evaluation paradigms: bridging the omni-modal capability-assessment gap via the VR-SDR task.** We construct the HumanOmni-Speaker Benchmark, integrating VR-SDR with atomic diagnostic subtasks. By categorizing samples into Easy and Hard levels—where the Hard set explicitly removes visual biases—we rigorously evaluate cross-modal identity consistency, temporal role stability, and comprehensive multimodal understanding without allowing shortcut learning.
2. **Architectural Optimization: introducing Visual Delta Encoder to address the shortcomings in high-frequency dynamic perception.** Through the Structured Visual Tokenizer (SVT), the Visual Delta Encoder compresses inter-frame motion residuals into just 6 tokens per frame. This highly condensed representation enables the model to support high-frequency video sampling at 25 fps, allowing it to explicitly track speakers and extract fine-grained visual phonemes (Visemes) without the catastrophic token overhead.
3. **HumanOmni-Speaker System: an Omni model tailored for speaker-centric tasks.** Experiments demonstrate that the proposed model significantly outperforms existing open-source models on our benchmark and competes with closed-source models like Gemini3-Pro, effectively addressing the shortcomings in fine-grained motion modeling and cross-modal spatio-temporal alignment. Notably, HumanOmni-Speaker is the first Omni model capable of end-to-end lip-reading and high-precision speaker localization directly from raw video without intrusive lip-cropping pre-processing. We will release the benchmark dataset, code, and model checkpoints upon acceptance.

2 Related Work

2.1 Speaker-Centric Tasks and Omni-Benchmarks

Traditional speaker research decomposes complex interactions into isolated, unimodal subtasks like Automatic Speech Recognition [26], Speaker Diarization [39], and Active Speaker Detection [31]. This "siloed" paradigm severs the "Who", "When", and "What", masking models' inability to perform genuine cross-modal spatio-temporal binding. Furthermore, existing Omni-modal benchmarks [8, 21, 22, 35] lack comprehensive speaker-centric evaluation. V-STaR [8] and OmniBench [21] omit audio-visual speaker tasks, while StreamingBench [22] and OmniMMI [35] are susceptible to visual biases, allowing models to exploit single-modal shortcuts rather than exhibiting true multimodal synergy. To strictly evaluate end-to-end audio-visual alignment and eliminate this illusion of competence, we introduce the VR-SDR task and the comprehensive HumanOmni-Speaker Benchmark.

2.2 Omni-Modal Large Language Models

Recent Omni-models—including Gemini3 [13], Qwen3-Omni [38], Qwen2.5-Omni [37], LLaMA-Omni [15], OLA [23], and VITA [16]—enable end-to-end multimodal processing. However, they typically rely on sparse visual sampling (1-2 fps), capturing static spatial semantics but creating a severe *architectural perception gap* that discards high-frequency dynamics like visemes. While recent work [20] increases the frame rate to 16 fps with post-hoc token compression, its underlying static feature extractor remains fundamentally misaligned for continuous motion. Naively increasing sampling rates risks a quadratic *token explosion* and drowns subtle temporal signals in redundant spatial noise. To overcome this, our Visual Delta Encoder operates at 25 fps to explicitly encode inter-frame motion residuals, capturing fine-grained dynamics with minimal token overhead to bridge the audio-visual spatio-temporal gap.

3 HumanOmni-Speaker Benchmark

To systematically assess the comprehensive capabilities of Omni-models in complex speaker-centric scenarios, we construct the HumanOmni-Speaker Benchmark —a hierarchically structured evaluation framework.

At its core is the holistic **VR-SDR** task, designed to rigorously test the unified understanding of "Who said what and when" across all modalities. To complement this, we incorporate four atomic subtasks: Speech Recognition (SR), Speaker Verification (SV), Speaker Localization (SL), and Speaker Identification (SI), as depicted in Fig. 1. The inclusion of these atomic tasks serves a critical *diagnostic* purpose: when a model fails the complex VR-SDR task, this hierarchical structure allows us to precisely isolate the underlying bottleneck—determining whether the failure stems from fundamental acoustic perception, spatial visual tracking, or the final cross-modal fusion.

Beyond the task architecture, the integrity of the evaluation data is equally critical. Existing speaker evaluation datasets are frequently compromised by strong visual biases—such as extreme close-ups, handheld microphones, or single-person compositions (see Fig. 2a). These biases inadvertently provide "recognition shortcuts," enabling models to succeed using static visual heuristics rather than genuine audio-visual alignment. To strictly eliminate this shortcut learning in our visual-dependent tasks, we implemented a rigorous *manual filtering* process to categorize the multimodal evaluation into **Easy** and **Hard** levels. For the Hard set, human annotators meticulously removed samples containing obvious visual clues. Consequently, the Hard set exclusively comprises complex multi-person scenes and non-subject compositions (Fig. 2b) where speaker identity cannot be reliably deduced from any single static frame. By integrating diagnostic subtasks and enforcing manually curated difficulty levels, the HumanOmni-Speaker Benchmark guarantees an uncompromising, objective evaluation of true cross-modal spatio-temporal capabilities. Detailed task divisions are summarized in Table 2.

Fig. 2: Humanomni-Speaker benchmark sample characteristics and statistics.(a) Samples with strong visual biases which provide "recognition shortcuts". (b) Samples without visual biases that require both acoustic and visual information for speaker identification. (c) Benchmark sample statistics overview.

3.1 Holistic Evaluation: Visual-Registered Speaker Diarization and Recognition (VR-SDR)

To comprehensively evaluate Omni-models, we propose a novel paradigm for resolving "*Who said what and when*"—Visual-Registered Speaker Diarization and Recognition (VR-SDR).

Task Definition: Given an audio-visual input of a multi-person conversation and a set of visual identity registration queries in natural language (e.g., *Alice: A woman with long curly hair; Bob: A man in a black T-shirt*), the model must achieve end-to-end identity binding based solely on visual descriptions. It is required to output structured records containing identity labels, timestamps, and transcribed content (e.g., *Alice: our first topic is family... [0.00-6.39s]; Bob: who am i the most like... [6.43-8.05s]*).

Unlike traditional audio-only SDR systems [39] that rely on pre-extracted speaker embeddings for registration, VR-SDR requires only visual semantic descriptions to complete identity binding, completely eliminating the need for prior acoustic enrollment. This task tightly integrates the speakers' visual and

audio streams with textual identity cues, aligning perfectly with the language-centric, multimodal interaction logic of Omni-models.

Evaluation Metrics: We employ *Identity-Fixed* metrics to rigorously evaluate "*Who said what*" and "*Who said when*".

Identity-Fixed SA-WER (Who said what): Unlike the traditional concatenated minimum-permutation word error rate (cpWER), which allows for optimal label re-alignment, we employ Speaker-Attributed Word Error Rate (SA-WER). This is a strictly harsher metric requiring the model's output to precisely correspond to the vision-registered ID. A word is counted as correct only if both the text and the assigned Speaker ID are accurate. This directly measures the model's capacity to bind visual descriptions with acoustic transcriptions. Formally, given a set of registered speakers $\mathcal{S}$, it is calculated as:

$$\text{SA-WER} = \frac{\sum_{s \in \mathcal{S}} \text{EditDistance}(\text{Ref}_s, \text{Hyp}_s)}{\sum_{s \in \mathcal{S}} |\text{Ref}_s|} \tag{1}$$

where Ref_s is the reference transcript for speaker s, and Hyp_s is the hypothesized transcript that the model attributes to speaker s.

Identity-Fixed IER (Who said when): The Identification Error Rate (IER) evaluates the model's capability in Identity-Fixed speaker diarization, simultaneously assessing identity binding accuracy and temporal alignment precision. Compared to the traditional Speaker Diarization Error Rate (DER), which allows for label permutation to minimize penalties, IER enforces an absolute mapping between the model output and the vision-registered identities $\mathcal{S}$. Formally, it is defined as:

$$\text{IER} = \frac{\sum_{s \in \mathcal{S}} (\text{Miss}_s + \text{FA}_s + \text{Conf}_s)}{\sum_{s \in \mathcal{S}} \text{Dur}_s} \tag{2}$$

where Miss_s, FA_s, and Conf_s represent the durations of missed speech, false alarms, and speaker confusion attributed to speaker s, respectively, and Dur_s denotes the total ground-truth duration for speaker s.

3.2 Diagnostic Evaluation and Construction Details

To systematically analyze performance bottlenecks in the holistic VR-SDR task, we designed a progressive diagnostic chain of four atomic subtasks—ranging from basic acoustic competence to full cross-modal identity binding. To construct this benchmark, we implemented a rigorous annotation pipeline. After filtering source datasets [6, 11, 14, 19, 26], we utilized Qwen-VL [29] to generate foundational appearance descriptions and QA pairs. A team of 10 English-proficient annotators (TEM-8 certified or equivalent) and three senior inspectors then performed meticulous calibration to guarantee high-fidelity annotations (Fleiss' $\kappa = 0.82$ on $\sim$1.5K samples; 18% disagreements resolved by inspector adjudication). For the Easy/Hard partitioning, a sample is labeled *Easy* if it exhibits any salient visual shortcut (close-up, visible microphone, single-person composition, or centrally lit speaker), and *Hard* otherwise. The resulting tasks are structured as follows:

- **Speech Recognition (SR):** Underpins the "What" dimension by evaluating basic acoustic content transcription. Evaluated using LibriSpeech-Test-Clean [26].
- **Speaker Verification (SV):** Probes acoustic identity discrimination (the "Who" dimension) without visual cues. Built from the VoxCeleb2 [11] test set, it comprises 486 binary QA pairs (1:1 positive-negative ratio).
- **Speaker Localization (SL):** Evaluates audio-visual spatial perception by requiring end-to-end localization of the active speaker. Extracted from Columbia-ASD [6] (500 multi-person clips), it forces models to directly output face coordinates from raw video, bypassing traditional cascaded "detection–cropping–classification" pipelines.
- **Speaker Identification (SI):** The atomic task closest to VR-SDR. It tests cross-modal identity binding by requiring the model to match a natural language visual description to the active speaker under visual interference. Sourced from AVSpeech [14] using 4-option multiple-choice questions.
- **VR-SDR (Holistic):** The ultimate stress test. Sourced from VoxMM [19], it comprises ∼464 multi-person conversation segments, with descriptions iteratively refined via human-LLM collaboration to ensure unique identity binding.

Within this framework, SR and SV serve as audio-only baselines, whereas VR-SDR, SL, and SI demand audio-visual processing of multi-person scenarios. To ensure visually dependent tasks are not confounded by spurious heuristics, we partitioned the SL and SI samples into Easy and Hard subsets. For the Hard sets, we manually filtered out samples exhibiting significant visual biases, strictly mandating genuine temporal audio-visual alignment. Evaluation metrics and sample distributions are summarized in Table 1 and Fig. 2(c), respectively.

Table 1: The Hierarchical Evaluation Matrix and Benchmark Statistics

.

	Evaluation Framework				Benchmark Statistics			
Level	**Speaker-Task**	**Modality**	**Core Competency**	**Metric**	**Samples**	**Duration**	**Avg Pers.**	**Avg Spk.**
Holistic	Vision-Registered Speaker Diarization Recognition (VR-SDR)	**A-V-T**	Omni-modal comprehension across audio, text, and video	SA-WER(%) ↓ IER (%) ↓	464	1.2h	3.4	2.1
Atomic	Speech Recognition (SR)	**A**	Acoustic Content Transcription	WER (%) ↓	2620	5.4h	1	1
	Speaker Verification (SV)	**A**	Acoustic Identity Discrimination	Error (%) ↓	486	1.0h	1.5	1.5
	Speaker Localization (SL)	A-**V**	Audio-Visual Spatial Perception	Miss Rate (%)↓	500	0.62h	2.6	1
	Speaker Identification (SI)	A-**V-T**	Audio-Visual Identity Binding	Error (%) ↓	942	1.3h	3.2	1

4 Architecture

4.1 Overview

HumanOmni-Speaker is an Omni model specifically built for human-centric speaking scenarios. As shown in Fig.3, the architecture serializes visual spatial, visual temporal, audio, and text features into tokens via the Visual Base Encoder,

Visual Delta Encoder, Audio Encoder, and Text Tokenizer. These tokens are aligned and fed into an LLM for decoding. Multimodal information is fused through a cross-attention mechanism to achieve a deep understanding of "Who", "When" and "What" in complex interactions.

Regarding visual representation, HumanOmni-Speaker utilizes a dual-stream path to explicitly deconstruct spatial semantics and temporal dynamics. The first stream, the **Visual Base Encoder**, shares its backbone with Qwen2.5-Omni and extracts stable identity features and environmental context through low-frequency sampling (1-2 fps). The second stream is the **Visual Delta Encoder**, our core innovation, explicitly models inter-frame residuals via 25 fps high-frequency sampling, capturing temporal details and high-frequency features that are ignored by Visual Base Encoder due to sparse sampling. This design ensures high-precision perception of speaker-centric scenes while maintaining a low computational load through functional specialization. The audio and text encoders remain consistent with Qwen2.5-Omni.

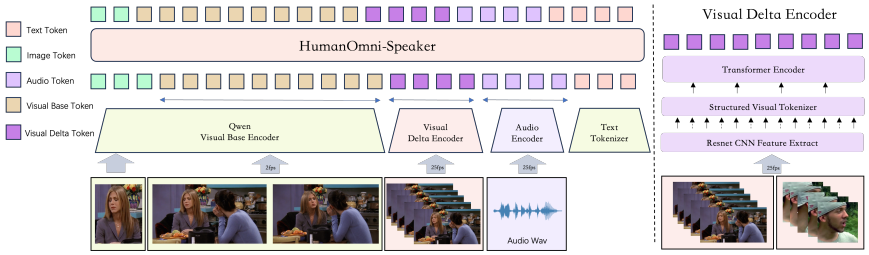

Fig. 3: Overview of HumanOmni-Speaker architecture for human-centric speaking scenarios. It integrates text, audio, and visual inputs through Text Tokenizer, Audio Encoder, Visual Base Encoder (1-2 fps), and Visual Delta Encoder (25 fps). The Visual Delta Encoder (right) employs a ResNet-18 backbone, Spatio-Temporal Vision Transformer (SVT), and Transformer encoder, producing only 6 structured tokens per frame. All modality-specific tokens are aligned in a shared LLM decoder, enabling the model to reason about "Who", "When" and "What" in complex interactions.

4.2 Visual Delta Encoder

Unlike traditional image encoders, the Visual Delta Encoder explicitly extracts fine-grained, inter-frame temporal dynamics. As depicted in Fig. 3 (right), it employs a three-stage architecture: **Local Feature Perception** utilizes a lightweight ResNet-18 backbone to balance computational efficiency with local motion capture at 25 fps. **Structured Visual Tokenizer (SVT)** applies hierarchical spatial (7×7) and large-receptive-field temporal ($k = 63$) convolutions to compress dense CNN features into just 6 structured tokens per frame. This mitigates the sequence load of high-frequency sampling while preserving high-fidelity lip visemes and motion trajectories. **Global Context Encoding** leverages a Transformer Encoder to process these tokens across frames, integrating discrete temporal increments into coherent behavioral semantics.

This architecture empowers the Visual Delta Encoder with two critical capabilities:

- **Spatio-Temporal Speaker Tracking:** Empowered by the 25 fps sampling rate, the encoder extracts highly continuous spatial trajectories. Grad-CAM visualizations (Fig. 4) confirm that its shallow CNN layers consistently lock onto the speaker's face and mouth, providing the underlying motion primitives required for VR-SDR.
- **Fine-Grained Viseme Modeling:** The module directly extracts visual phonemes (visemes) from raw video streams, bypassing the need for intrusive face alignment or lip-cropping. As evidenced in Table 3, this inter-frame residual representation successfully overcomes the historical limitations of Omni-models in Visual Speech Recognition (VSR).

Through the synergy of trajectory tracking and viseme modeling, the Visual Delta Encoder serves as a vital "high-frequency cross-modal bridge." By outputting dense spatio-temporal tokens, it seamlessly spans the gap between high-frequency audio streams and sparse, high-level visual semantics. Ultimately, it transcends basic feature extraction, directly enabling the Omni-model to spatially and temporally align multi-speaker dynamics and reliably resolve "Who said what and when."

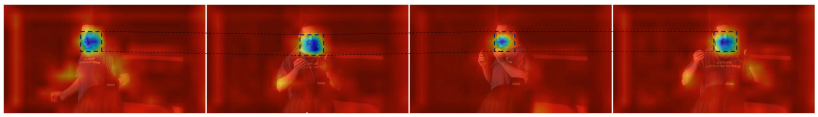

Fig. 4: The attention maps generated by Grad-CAM show that Visual Delta Encoder successfully localizes and tracks the speaker's mouth.

5 Training

We employ a progressive three-stage training paradigm to systematically equip the model with high-frequency dynamic perception and cross-modal alignment capabilities (Fig. 5).

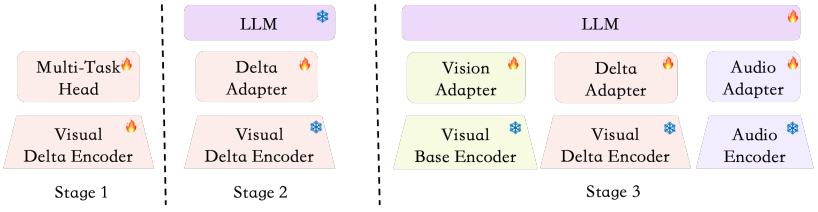

Fig. 5: The Progressive Training Pipeline of HumanOmni-Speaker.

Stage 1: Visual Delta Encoder Pre-training. To master high-frequency motion primitives, we pre-train the Visual Delta Encoder on a curated large-scale speaker-centric dataset aggregating AVSpeech, LRS2, LRS3, VoxCeleb2, and AVA-ASD [1, 11, 12, 14, 31]. We formulate this as a multi-task supervised learning framework. Using a shared encoder backbone with task-specific heads, we simultaneously optimize for spatial tracking (via face center-point regression) and fine-grained lip-reading (using VSR and AVSR objectives from CoGenAV [3]). This simultaneous optimization endows the encoder with the fundamental capability to continuously track spatial displacements and extract robust viseme features in the temporal domain.

Stage 2: Modality Alignment. Next, we align the visual delta features with the latent semantic space of the LLM. Using the pre-training dataset, we freeze both the Visual Delta Encoder and the LLM decoder, optimizing only a lightweight MLP projector. Crucially, we enforce *modality isolation* during this phase: the original Visual Base Encoder and Audio Encoder from Qwen2.5-Omni are bypassed, and the LLM receives features exclusively from the Visual Delta Encoder. This strict isolation forces the LLM to effectively translate high-fidelity viseme details and motion trajectories into interpretable behavioral semantics, preventing it from relying on representational shortcuts from established modalities.

Stage 3: End-to-End Joint Fine-Tuning. Finally, we construct a comprehensive dataset of approximately 1.5 million audio-visual instructions tailored for speaker-centric scenarios, encompassing ASR, VSR, Speaker Localization, Verification, Identification, and VR-SDR. We standardize all unimodal and multimodal inputs (raw video frames, continuous audio waveforms, and text prompts) into a unified instruction format.

During this joint optimization phase, all structural encoders remain strictly frozen while their respective MLP projectors are updated. For the LLM decoder, we employ a hybrid parameter-efficient tuning strategy: the initial layers undergo full fine-tuning to promote deep, early-stage cross-modal fusion, while the deeper layers are adapted via LoRA. This strategy maximizes multimodal spatio-temporal synergy while preserving the foundational generation quality of the underlying LLM.

6 Experiment

6.1 HumanOmni-Speaker Benchmark

Table 2 reports the optimal prompt-tuned performance of various models across the HumanOmni-Speaker Benchmark.

Acoustic Perception: Speech Transcription and Voiceprint Analysis. While most Omni-models achieve near-ceiling performance in basic Speech Recognition (ASR), the Speaker Verification task exposes a critical flaw in open-source voiceprint extraction. Models like OLA and VITA1.5 perform near random chance, and the Qwen-Omni series exhibits severe distribution bias (over-predicting distinct speakers). In contrast, the closed-source Gemini3-Pro demonstrates robust

Table 2: Results on HumanOmni-Speaker Benchmarks. The best results are highlighted.

Method	Model Size	Speech Recognition	Speaker Verification	Speaker Localization	Speaker Identification		Atomic AVG	VR-SDR What	VR-SDR When	Holistic AVG
		easy(↓)	easy(↓)	easy(↓)	easy(↓)	hard(↓)	(↓)	hard(↓)	hard(↓)	(↓)
Closed-source Omni Model										
Gemini3-Pro	-	1.39	**5.2**	12.8	5.5	30.5	11.1	**36.6**	36.3	**36.5**
Qwen3-Omni-flash	-	**1.22**	43.9	2.8	3.6	43.5	19.0	82.9	47.2	65.0
Open-source Omni Model										
OLA [23]	7B	1.9	51.1	20.6	12.4	63.2	29.8	95.4	56.85	76.1
VITA1.5 [16]	7B	3.4	51.4	20.2	10.3	56.6	28.4	93.6	54.40	74
Qwen2.5-Omni [37]	3B	2.2	44.2	7.4	6.6	51.5	22.4	84.6	50.9	67.8
Qwen2.5-Omni [37]	7B	1.8	37.1	7.4	4.0	54.5	20.9	83.6	49.4	66.5
HumanOmni-Speaker Model										
Qwen2.5-Omni-SFT	3B	2.0	13.4	3.0	1.7	33.2	10.7	52.1	31.5	41.8
HumanOmni-Speaker	3B	1.9	13.2	**0.8**	**1.0**	**21.1**	**7.6**	47.1	**28.5**	37.8

discriminative capability with a 5.2% error rate. HumanOmni-Speaker successfully bridges this open-source gap, achieving a 13.2% error rate and significantly outperforming its base Qwen2.5-Omni-3B model.

Visual Perception: High-Precision Speaker Localization and Identity Binding. HumanOmni-Speaker dominates the Speaker Localization task with a mere 0.8% error rate, drastically outperforming both Qwen3-Omni (2.8%) and Gemini3-Pro (12.8%). This stark contrast underscores a direct correlation between spatial accuracy and visual sampling frequency. While Gemini and Qwen-Omni rely on sparse 1–2 fps sampling, our dual-rate architecture (incorporating the 25 fps Visual Delta Encoder) successfully captures the high-frequency dynamics necessary for precise speaker tracking.

Furthermore, the Speaker Identification task reveals a profound performance gap between the Easy and Hard sets. In the Easy set, all models exhibit low error rates (mostly $< 10\%$), indicating a baseline ability to associate captions with clear visual subjects. However, in the Hard set—which features extreme multi-speaker interference and removes visual shortcuts—all baseline models suffer catastrophic performance degradation. This confirms that existing Omni-models heavily rely on static visual heuristics, whereas robust cross-modal identity binding in dynamic environments remains a critical frontier that our architecture begins to address.

Multimodal Understanding: Vision-Registered Speaker Diarization and Recognition. As a comprehensive stress test, the VR-SDR task demands the seamless integration of textual queries, visual streams, and audio signals to jointly resolve "Who", "When", and "What". Our findings reveal that most open-source Omni-models—such as the Qwen-Omni series, OLA, and VITA1.5—falter under the strict demands of full-modality fusion. In contrast, Gemini3-Pro establishes a robust closed-source baseline, achieving 36.6% and 36.3% on SA-WER and IER, respectively. This profound performance gap underscores that current open-source models largely fail to achieve genuine spatio-temporal binding between visual identities and acoustic dynamics, indicating that true Omni-modal collaborative understanding remains in its infancy.

Holistic Analysis: HumanOmni-Speaker Performance and Remaining Challenges. Comparative analysis against a Qwen2.5-Omni-SFT baseline (lacking the Visual Delta Encoder) highlights the critical necessity of our proposed

module in deciphering complex, dynamic multi-speaker environments. Integrating the Visual Delta Encoder yielded a dramatic reduction in Speaker Localization error, dropping from 3.0% to a remarkable 0.8%. Concurrently, it drove a substantial decrease in the Speaker Identification Hard Set error rate, lowering it from 33.2% to 21.1%. Beyond spatial localization, this high-frequency motion encoding catalyzes true full-modality synergy. In the demanding VR-SDR task, the module optimized the "Who said what" and "Who said when" metrics from 52.1% and 31.65% down to 47.1% and 28.5%, respectively. Decomposing the residual 47.1% SA-WER reveals that missed segments account for ∼6%, identity misattribution for ∼25% (consistent with SI-Hard error), and transcription errors for the remainder (speaker-blind WER = 21.0%)—confirming that cross-modal identity binding is the primary bottleneck. Moreover, as speaker count grows from 2 to 3, both SA-WER and IER degrade substantially; Gemini3-Pro exhibits the same trend, indicating that multi-party overlapping speech remains a systematic challenge for current Omni architectures regardless of model scale.These gains conclusively validate the Visual Delta Encoder as an essential mechanism for enforcing robust spatio-temporal alignment and achieving the deep cross-modal binding required to definitively resolve "Who said what and when".

6.2 Visual and Audio-Visual Speech Recognition

Due to the fine-grained dynamic representations provided by Visual Delta Encoder, HumanOmni-Speaker excels in visual speech recognition (VSR) and audio-video speech recognition (AVSR). We validated the model's performance on LRS2 and LRS3, and the results are shown in Table 3.

Compared to Specific VSR models: No preprocessing required. Traditional specific VSR models such as Av-HuBERT and Auto-AVSR rely on complex pre-processing including face alignment and lip region of interest (ROI) cropping. HumanOmni-Speaker achieves comparable performance to Auto-AVSR on the VSR task of both the LRS3 (33.4% vs 33.0%) and LRS2 (27.9% vs 29.8%) without any pre-processing of the raw dataset.

Compared to LLM-based AVSR models: Stronger audio-video fusion performance. Compared to recent LLM-based AVSR models such as Llama-AVSR and Whisper-flamingo, HumanOmni-Speaker delivers exceptional performance in core AVSR metrics. Our model delivers competitive AVSR performance, yielding 0.76% WER on LRS3 and 1.36% on LRS2, outperforming other LLM-based models.

Compared to Omni models: Filling the Lip-Reading Gap. Because existing Omni models focus on global visual and audio understanding and lack mechanisms for capturing high-frequency dynamic features, they are often ineffective when faced with VSR and AVSR tasks. HumanOmni-Speaker is the first Omni model to support end-to-end lip-reading natively, filling the gap in fine-grained motion modeling of Omni models.

Comparing the performance of the HumanOmni-Speaker model on ASR and AVSR. we observe that adding visual information significantly reduces the WER from 3.63% to 0.76% on LRS3, and from 3.47% to 1.36% on

Table 3: Results on VSR and AVSR.

Method	Preprocessing	LRS2			LRS3		
		vsr	*asr*	*avsr*	*vsr*	*asr*	*avsr*
Specific VSR Model							
CTC/Attention [28]	Lip crop & align	63.5	-	7.0	-		
AV-HuBERT Base [33] [41]	Lip crop & align	31.2	-	-	34.8	-	-
AutoAVSR [24]	Lip crop & align	27.9	-	1.5	33.0	-	0.9
LLM-base AVSR Model							
Llama-SMoP [5]	Lip crop & align	-			-	-	0.96
Llama-AVSR [4]	Lip crop & align	-			24.0	0.79	0.77
Whisper-flamingo [32]	Lip crop & align	-	-	1.4	-	-	0.76
Omni Model							
OLA [23]	Raw video	-	5.5	-	-	4.7	-
Qwen2.5-Omni [37]	Raw video	-	3.47	-	-	3.63	-
HumanOmni-Speaker	Raw video	29.8	3.47	**1.36**	33.4	3.63	**0.76**

Table 4: Results on ASR.

Method	Model Size	Librispeech-dev		Librispeech-test	
		dev-clean	*dev-other*	*test-clean*	*test-other*
Audio Models					
Whisper-small [30]	0.3B	4.4	10.1	4.6	10.3
SenseVoice-L [2]	1.6B	-		2.6	4.3
Qwen-Audio [10]	7B	1.8	4.0	2.0	4.2
Qwen2-Audio [9]	7B	1.3	3.4	1.6	3.6
Omni Models					
HumanOmni [40]	7B	3.8	7.5	3.7	8.0
VITA [16]	7B	7.6	16.6	8.1	18.4
Mini-Omni2 [36]	7B	4.7	9.4	4.8	9.8
OLA [23]	7B	1.9	4.4	1.9	4.2
Qwen2.5-Omni [37]	3B	2.0	4.1	2.2	4.5
HumanOmni-Speaker	3B	1.91	4.16	1.88	4.56

LRS2. This quantitatively demonstrates that lip movement features extracted by the Visual Delta Encoder strongly enhance the audio signal and effectively serve as a semantic bridge across modalities.

6.3 Automatic Speech Recognition

To evaluate HumanOmni-Speaker on ASR task, we compared its performance with Audio-LLMs and omni models on LibriSpeech. As shown in Table 4.

HumanOmni-Speaker (3B) outperforms existing open-source omni models on LibriSpeech, including HumanOmni, VITA, Mini-Omni2, and OLA. It achieves comparable performance to the base model Qwen2.5-Omni (3B), and even slightly outperforms it on clean subsets, with 1.91% and 1.88% WER (vs. 2.0% and 2.2%) on dev-clean and test-clean, respectively. Compared to dedicated Audio-LLMs, it also achieves comparable performance with a smaller parameter set.

This indicates that HumanOmni-Speaker retains the strong ASR capabilities of Qwen2.5-Omni. We suggest that our design of the model structure and training method is crucial. In terms of model structure, while incorporating new speaker-aware modules Visual Delta Encoder to enhance visual modeling, it retains the original audio encoder and language decoder backbone. Regarding training methods, the three-stage training strategy allows the model to effectively integrate dynamic speaker-related visual information without interfering with the original speech recognition pathway.

7 Ablation Study

We conduct a series of ablation experiments to validate the design choices of the Visual Delta Encoder from four perspectives: component effectiveness, end-to-end design advantages, hyperparameter sensitivity, and computational cost.

Component Effectiveness. As shown in Table 5, using only the Visual Delta Encoder (without Visual Base Encoder) on the Speaker Localization task achieves a 1.2% error rate, already significantly better than the base model (3.0%), indicating that the Visual Delta Encoder alone exhibits strong spatial awareness. Combining both encoders further reduces the error to 0.8%. On the more complex VR-SDR task, integrating the Visual Delta Encoder reduces SA-WER from 52.1% to 47.1% and IER from 31.5% to 28.5%.This indicates that the Visual Delta

Table 5: Ablation Study. Percentages in parentheses denote the relative improvement compared to the baseline (visual base only).

	Feature-type *visual base* | *visual delta* | *audio*	Speaker **Loc.** (↓)	VR-SDR **What** (↓)	VR-SDR **When** (↓)
pipeline-based	Yes | No | Yes	-	64.2	39.3
HumanOmni Speaker	Yes | No | Yes	3.0	52.1	31.5
	No | Yes | Yes	1.2	-	-
	Yes | Yes | Yes	**0.8** (-73%)	**47.1** (-9.6%)	**28.5** (-9.5%)

Encoder, by enhancing visual awareness of speaker location, acts as an effective semantic bridge across modalities, aligning high-frequency audio with sparse visual information and enabling deeper binding of "Who said what and when".

End-to-End vs. Alternative Pipelines. We compare our end-to-end architecture against two common alternative paradigms. First, a pipeline-based approach that applies Whisper-diarization [25] for audio segmentation followed by Qwen3-VL [29] for visual speaker identification yields significantly higher error rates (64.2% SA-WER and 39.3% IER), underscoring the advantage of tightly coupling audio-visual semantics within a single model. Second, replacing raw full-frame input with face-cropped regions from an off-the-shelf detector causes the SI-Hard error rate to rise sharply from 21% to 38%. This degradation stems from cascaded detection errors in multi-person scenes and the loss of global contextual cues (head pose, relative position, gesture) that are critical for cross-modal disambiguation. Together, these comparisons validate our end-to-end design, where the Visual Delta Encoder attends to the active speaker's mouth while retaining surrounding scene context for robust identity resolution.

Hyperparameter Sensitivity. We further analyze two key design parameters of the Visual Delta Encoder: the number of structured tokens per frame and the input frame rate. As shown in Fig. 6(a), reducing the token count below 6 leads to significant degradation in speaker localization accuracy due to insufficient spatial representation, while increasing it beyond 6 yields minimal gains—indicating that 6 tokens per frame strikes an optimal balance. Fig. 6(b) shows that an excessively low frame rate (e.g., 2 FPS) prevents effective lip reading; 16 FPS enables basic capability, while 25 FPS achieves optimal performance for fine-grained viseme modeling.

Computational Cost. Building on the above hyperparameter choices (6 tokens, 25 fps), we measure the resulting overhead on a 5.7 s clip (420×756, Qwen2.5-Omni-3B, V100). Incorporating Visual Delta tokens increases the token count from 2,257 to 3,111, inference time from 2.62 s to 3.13 s, peak GPU memory from 13.2 GB to 17.2 GB, and FLOPs from 26.9 T to 38.2 T. Compared to a naive 25 fps ViT approach that incurs $>10\times$ token blow-up, our structured design caps the overhead at only +38% additional tokens, demonstrating a favorable trade-off between high-frequency temporal perception and computational efficiency.

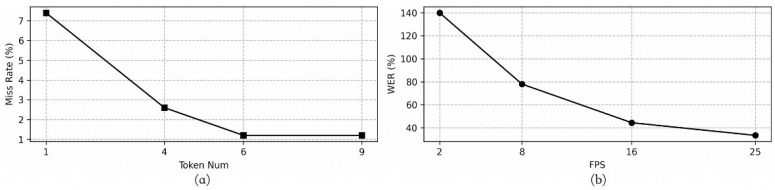

Fig. 6: (a) Effect of token number on SL task. (b) Effect of FPS on VSR task.

8 Conclusion

In this work, we presented HumanOmni-Speaker, a unified Omni-modal LLM specifically designed for complex, speaker-centric interactions. By introducing a high-frame-rate Visual Delta Encoder, we addressed the architectural perception gap, facilitating genuine end-to-end joint modeling of "*Who said what and when.*" Furthermore, to systematically mitigate visual shortcuts and overcome current evaluation limitations, we proposed the rigorous VR-SDR paradigm and the comprehensive HumanOmni-Speaker Benchmark. Extensive experiments demonstrate that our model significantly outperforms existing open-source baselines across a diverse spectrum of tasks—ranging from fine-grained lip-reading and spatial localization to holistic speaker diarization and recognition.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant 62576224, the Guangdong Provincial Key Laboratory of Computility Microelectronics 2024B1212010007, the Guangdong Basic and Applied Basic Research Foundation Project 2026A1515011733, the Shenzhen Key Technical Projects under Grant CJGJZD20220517141605013, JCYJ20220818101406014 and JSGG20220831105801004.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition (2018), https://arxiv.org/abs/1809.00496
2. An, K., Chen, Q., Deng, C., Du, Z., Gao, C., Gao, Z., Gu, Y., He, T., Hu, H., Hu, K., et al.: Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. arXiv preprint arXiv:2407.04051 (2024)
3. Bai, D., Ma, Z., Wei, X., Bo, L.: Cogenav: Versatile audio-visual representation learning via contrastive-generative synchronization (2025), https://arxiv.org/abs/2505.03186
4. Cappellazzo, U., Kim, M., Chen, H., Ma, P., Petridis, S., Falavigna, D., Brutti, A., Pantic, M.: Large language models are strong audio-visual speech recognition learners (2025), https://arxiv.org/abs/2409.12319
5. Cappellazzo, U., Kim, M., Petridis, S., Falavigna, D., Brutti, A.: Scaling and enhancing llm-based avsr: A sparse mixture of projectors approach (2025), https://arxiv.org/abs/2505.14336
6. Chakravarty, P., Tuytelaars, T.: Cross-modal supervision for learning active speaker detection in video (2016), https://arxiv.org/abs/1603.08907
7. Chen, L., Yue, Z., Xu, B., Jin, Q.: Unveiling visual biases in audio-visual localization benchmarks (2024), https://arxiv.org/abs/2409.06709
8. Cheng, Z., Hu, J., Liu, Z., Si, C., Li, W., Gong, S.: V-star: Benchmarking video-llms on video spatio-temporal reasoning (2025), https://arxiv.org/abs/2503.11495
9. Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., Zhou, J.: Qwen2-audio technical report (2024), https://arxiv.org/abs/2407.10759
10. Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., Zhou, J.: Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models (2023), https://arxiv.org/abs/2311.07919
11. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: Interspeech 2018. interspeech-2018, ISCA (Sep 2018). https://doi.org/10.21437/interspeech.2018.1929, http://dx.doi.org/10.21437/Interspeech.2018-1929
12. Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jul 2017). https://doi.org/10.1109/cvpr.2017.367, http://dx.doi.org/10.1109/CVPR.2017.367
13. DeepMind, G.: https://deepmind.google/models/gemini/ (2025), https://deepmind.google/models/gemini/
14. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics **37**(4), 1–11 (Jul 2018). https://doi.org/10.1145/3197517.3201357, http://dx.doi.org/10.1145/3197517.3201357
15. Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., Feng, Y.: Llama-omni: Seamless speech interaction with large language models (2025), https://arxiv.org/abs/2409.06666
16. Fu, C., Lin, H., Wang, X., Zhang, Y.F., Shen, Y., Liu, X., Cao, H., Long, Z., Gao, H., Li, K., Ma, L., Zheng, X., Ji, R., Sun, X., Shan, C., He, R.: Vita-1.5: Towards gpt-4o level real-time vision and speech interaction (2025), https://arxiv.org/abs/2501.01957

17. Fung, P., Bachrach, Y., Celikyilmaz, A., Chaudhuri, K., Chen, D., Chung, W., Dupoux, E., Gong, H., Jégou, H., Lazaric, A., Majumdar, A., Madotto, A., Meier, F., Metze, F., Morency, L.P., Moutakanni, T., Pino, J., Terver, B., Tighe, J., Tomasello, P., Malik, J.: Embodied ai agents: Modeling the world (2025), https://arxiv.org/abs/2506.22355
18. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., Gonzalez, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video (2022), https://arxiv.org/abs/2110.07058
19. Kwak, D., Jung, J., Nam, K., Jang, Y., Jung, J.W., Watanabe, S., Chung, J.S.: Voxmm: Rich transcription of conversations in the wild. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 12551–12555 (2024). https://doi.org/10.1109/ICASSP48485.2024.10446300
20. Li, Y., Tang, C., Zhuang, J., Yang, Y., Sun, G., Li, W., Ma, Z., Zhang, C.: Improving llm video understanding with 16 frames per second (2025), https://arxiv.org/abs/2503.13956
21. Li, Y., Zhang, G., Ma, Y., Yuan, R., Zhu, K., Guo, H., Liang, Y., Liu, J., Wang, Z., Yang, J., Wu, S., Qu, X., Shi, J., Zhang, X., Yang, Z., Wang, X., Zhang, Z., Liu, Z., Benetos, E., Huang, W., Lin, C.: Omnibench: Towards the future of universal omni-language models (2025), https://arxiv.org/abs/2409.15272
22. Lin, J., Fang, Z., Chen, C., Wan, Z., Luo, F., Li, P., Liu, Y., Sun, M.: Streamingbench: Assessing the gap for mllms to achieve streaming video understanding (2024), https://arxiv.org/abs/2411.03628
23. Liu, Z., Dong, Y., Wang, J., Liu, Z., Hu, W., Lu, J., Rao, Y.: Ola: Pushing the frontiers of omni-modal language model (2025), https://arxiv.org/abs/2502.04328
24. Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., Pantic, M.: Auto-avsr: Audio-visual speech recognition with automatic labels. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
25. Mahmoud, A.: Whisper diarization: Speaker diarization using openai whisper (2024), gitHub repository
26. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210 (2015). https://doi.org/10.1109/ICASSP.2015.7178964
27. Park, T.J., Kanda, N., Dimitriadis, D., Han, K.J., Watanabe, S., Narayanan, S.: A review of speaker diarization: Recent advances with deep learning (2021), https://arxiv.org/abs/2101.09624

28. Petridis, S., Stafylakis, T., Ma, P., Tzimiropoulos, G., Pantic, M.: Audio-visual speech recognition with a hybrid ctc/attention architecture (2018), https://arxiv.org/abs/1810.00108
29. QwenTeam, A.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
30. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023)
31. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., Pantofaru, C.: Ava-activespeaker: An audio-visual dataset for active speaker detection (2019), https://arxiv.org/abs/1901.01342
32. Rouditchenko, A., Gong, Y., Thomas, S., Karlinsky, L., Kuehne, H., Feris, R., Glass, J.: Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation. arXiv preprint arXiv:2406.10082 (2024)
33. Shi, B., Hsu, W.N., Lakhotia, K., Mohamed, A.: Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184 (2022)
34. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator (2015), https://arxiv.org/abs/1411.4555
35. Wang, Y., Wang, Y., Chen, B., Wu, T., Zhao, D., Zheng, Z.: Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts (2025), https://arxiv.org/abs/2503.22952
36. Xie, Z., Wu, C.: Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities (2024), https://arxiv.org/abs/2410.11190
37. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., Lin, J.: Qwen2.5-omni technical report (2025), https://arxiv.org/abs/2503.20215
38. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., Lin, J.: Qwen3-omni technical report (2025), https://arxiv.org/abs/2509.17765
39. Yin, H., Chen, Y., Deng, C., Cheng, L., Wang, H., Tan, C.H., Chen, Q., Wang, W., Li, X.: Speakerlm: End-to-end versatile speaker diarization and recognition with multimodal large language models (2026), https://arxiv.org/abs/2508.06372
40. Zhao, J., Yang, Q., Peng, Y., Bai, D., Yao, S., Sun, B., Chen, X., Fu, S., chen, W., Wei, X., Bo, L.: Humanomni: A large vision-speech language model for human-centric video understanding (2025), https://arxiv.org/abs/2501.15111
41. Zhu, Q., Zhou, L., Zhang, Z., Liu, S., Jiao, B., Zhang, J., Dai, L., Jiang, D., Li, J., Wei, F.: Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. IEEE Transactions on Multimedia **26**, 1055–1064 (2024). https://doi.org/10.1109/tmm.2023.3275873, http://dx.doi.org/10.1109/TMM.2023.3275873