EYES OVER EARS: MITIGATING VISUALLY INDUCED ACOUSTIC HALLUCINATIONS IN OMNI-MODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal LLMs jointly process language, vision, and audio. In practice, audio is incorporated only after a vision-language backbone has been trained. We find that established visual priors in such backbones can override acoustic evidence on audio-related tasks, giving rise to *Visually Induced Acoustic Hallucinations*. To diagnose and mitigate these hallucinations, we develop a diagnostic benchmark and synthesize training data targeting these failure modes. Directly fine-tuning existing open-source omni-modal LLMs on these data yields limited changes to their learned modality-dependence patterns, motivating intervention earlier in training. We therefore introduce Audite-Omni-4B, a compact model trained through progressive audio alignment, Mid-Training, and SFT. Introducing targeted supervision from Mid-Training onward proves more effective, shifting multimodal attention toward audio tokens during audio-related reasoning and substantially improving performance on our diagnostic benchmark. The improvements also extend to open-source omni QA and captioning tasks, indicating generalization beyond the targeted failure mode. We further train task-specialized GRPO experts and consolidate their respective strengths into a single model through MOPD. Audite-Omni-4B scores 89.9 on our diagnostic suite, compared with 72.7 for the strongest closed-source baseline Qwen3.8-Omni-Flash. It also surpasses Qwen3-Omni-30B-A3B on 10 of 11 public omni QA and captioning benchmarks, establishing a new open-source SOTA.

1 INTRODUCTION

Omni-modal large language models (LLMs) jointly process text, images, video, speech, and environmental sound within a unified reasoning framework. Systems such as Gemini (Gemini Team, 2023), GPT-4o (OpenAI, 2024), Qwen3-Omni (Xu et al., 2025b), and Seed2.0 Lite (ByteDance Seed, 2026) show the value of this interface for video understanding and real-time interaction. Almost all open-source omni models introduce audio as a post-hoc addition to a pre-trained vision-language backbone, integrating a compressed audio branch via alignment, joint optimization, and supervised fine-tuning (SFT) (Li et al., 2024b; Fu et al., 2025b; Xu et al., 2025a). This late integration creates a structural risk that visual and acoustic evidence make unequal causal contributions to model predictions during inference. Moreover, cross-modal semantic redundancy in standard audio-video data often allows models to predict events from visual or linguistic cues without fully exploiting the audio. Consequently, strong audio-only performance offers no guarantee that the model genuinely grounds its predictions in acoustic evidence under joint audio-visual inputs.

General audio understanding encompasses two foundational aspects: spoken-content recognition and acoustic-event perception. The former requires faithful recognition of the lexical content present in speech, whereas the latter requires identifying acoustic events and their sources in the audio signal. Figure 1(a) illustrates both failure modes: Chinese subtitles lead the model to identify English speech as Chinese, while the visible fire and steaming pot induce hallucinated descriptions of crackling flames and boiling water. We refer to this broader phenomenon as *Visually Induced Acoustic Hallucinations* and distinguish its two forms. *Speech-Content Hallucinations* introduce spoken content absent from the acoustic input, while *Audio-Visual Attribution Hallucinations* report visually implied events or sources as if they had been heard.

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

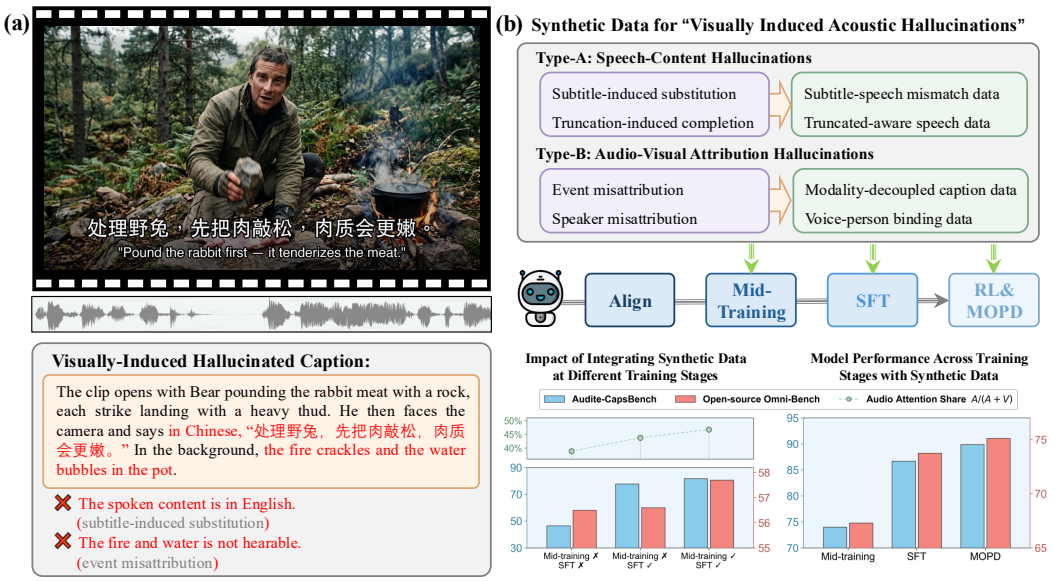


Figure 1: **(a)** Representative speech-content and audio-visual attribution hallucinations. **(b)** Targeted data construction and progressive training of Audite-Omni-4B. Injecting targeted data during both Mid-Training and SFT is more effective than SFT-only injection, while performance improves steadily across training stages.

We organize these two categories into four controlled problems. *Speech-Content Hallucinations* include **(i)** *Subtitle-Induced Substitution* and **(ii)** *Truncation-Induced Completion*; *Audio-Visual Attribution Hallucinations* include **(iii)** *Event Misattribution* and **(iv)** *Speaker Misattribution*. For each problem, we construct a diagnostic benchmark that quantifies hallucination severity and synthesize corresponding training data to mitigate the identified failure mode. Directly fine-tuning open-source omni-modal LLMs on these data produces limited changes to their established modality-dependence patterns, motivating us to examine whether effective intervention must occur earlier.

On this basis, we introduce Audite-Omni-4B, a compact omni-modal LLM designed to preserve acoustic grounding under joint audio-visual inputs. As shown in Figure 1(b), SFT-only training on the synthesized data reduces these hallucinations, but remains less effective than introducing the data during both Mid-Training and SFT. More generally, introducing our synthesized data increases the relative attention allocated to audio, $A/(A+V)$, providing complementary evidence that the model relies more strongly on acoustic information during inference. The same supervision also improves performance on open-source omni-related QA and captioning tasks, indicating that the gains are not confined to our diagnostic suite but transfer to general omni-modal capabilities.

The complete training recipe for Audite-Omni-4B is organized as a progressive pipeline. Audio alignment first establishes acoustic access, Mid-Training jointly optimizes all modalities with the targeted grounding data, and SFT strengthens instruction following and perceptual coverage. We finally train task-specific specialists with Group Relative Policy Optimization (GRPO) for capabilities including automatic speech recognition (ASR), question answering, multiple-choice reasoning, and captioning; we then consolidate their complementary strengths into a single model through Multi-Teacher On-Policy Distillation (MOPD). Figure 1(b) shows consistent stage-wise improvements in acoustic grounding on Audite-CapsBench and general omni-modal performance on open-source Omni-Bench. Despite its 4B scale, Audite-Omni-4B retains strong audio and visual performance and outperforms Qwen3-Omni-30B-A3B on most evaluated audio-visual tasks.

Our contributions are threefold:

- We uncover systematic *Visually Induced Acoustic Hallucinations* in omni-modal LLMs and establish four controlled diagnostic benchmarks.
- We find that SFT offers limited leverage over established modality dependence, whereas intervention from Mid-Training yields stronger grounding than SFT-only intervention.
- We present Audite-Omni-4B, which couples stage-aware training with MOPD and surpasses substantially larger models across most audio-visual benchmarks.

2 Related Works

Omni-Modal LLMs. Multimodal LLMs have evolved from modality-specific systems toward unified models that process text, images, video, and audio. Early approaches use shared multimodal representations or lightweight connectors to align frozen perceptual encoders with a language model, as exemplified by ImageBind, PandaGPT, Video-LLaMA, and OneLLM (Girdhar et al., 2023; Su et al., 2023; Zhang et al., 2023; Han et al., 2024). SALMONN (Tang et al., 2024) extends language models with general hearing capabilities, while video-SALMONN and VideoLLaMA 2 (Sun et al., 2024; Cheng et al., 2024) further combine speech, environmental sounds, and vision. Later systems, including VITA-1.5, Baichuan-Omni, Qwen2.5-Omni, and Qwen3-Omni, adopt multi-stage recipes spanning modality alignment, joint multimodal training, instruction tuning, and preference- or reinforcement-based post-training (Fu et al., 2025b; Li et al., 2024b; Xu et al., 2025a;b). Despite strong aggregate performance, existing studies mainly evaluate modality coverage, average benchmark scores, and unimodal retention. Aside from work on modality dominance (Sun et al., 2024), few studies use controlled contradictory inputs to test whether predictions follow acoustic evidence rather than visual, textual, or linguistic priors, or examine how such reliance depends on the stage at which supervision is introduced.

Audio Grounding and Multimodal Robustness. Audio understanding is studied through captioning and clip-level question-answering benchmarks such as AudioCaps, Clotho, Clotho-AQA, AVQA, and MUSIC-AVQA (Kim et al., 2019; Drossos et al., 2020; Lipping et al., 2022; Yang et al., 2022; Li et al., 2022), while AIR-Bench, AudioBench, MMAU, and MMAR assess instruction following and reasoning across speech, environmental sound, and music (Yang et al., 2024; Wang et al., 2025a; Sakshi et al., 2025; Ma et al., 2025). Specialized benchmarks address text-conditioned temporal localization through AudioGrounding and Clotho-Moment (Xu et al., 2021; Munakata et al., 2025), and speaker-attributed transcription through M2MeT 2.0 (Liang et al., 2023). Hallucination studies further reveal object-presence biases in audio-language models (Kuan et al., 2024; Cheng et al., 2025), while AVHBench shows that visual evidence can induce nonexistent sounds and audio can induce nonexistent visual events (Kim et al., 2025). Generative ASR studies similarly show that non-speech inputs, acoustic degradation, and distribution shifts can produce fluent but unsupported transcriptions that conventional word error rates may not capture (Koenecke et al., 2024; Frieske & Shi, 2024; Barański et al., 2025; Atwany et al., 2025). These efforts expose related but largely separate failures in task coverage, temporal grounding, speaker attribution, and hallucination. In contrast, our benchmark provides a unified, speech-centered diagnosis of whether omni-modal LLMs correctly attribute, prioritize, delimit, and temporally localize acoustic evidence under controlled subtitle, visual, and linguistic conflicts.

3 Methodology

3.1 Audio-Grounding Diagnostics and Targeted Data Synthesis

We define *audio grounding* as an omni-modal LLM's ability to keep its predictions faithful to the acoustic stream when visual context or linguistic priors provide compatible, incomplete, or conflicting alternatives. Without robust grounding, models suffer from *visually induced acoustic hallucinations*, a pervasive vulnerability in which visual cues override acoustic reality. To diagnose this phenomenon, we introduce a benchmark organized around four controlled failure modes. *Subtitle-Induced Substitution* tests whether the model follows the words actually spoken rather than contradictory subtitle text embedded in the visual context. *Truncation-Induced Completion* measures whether the model restricts its output to the observed acoustic span instead of completing a plausible continuation from visual or linguistic priors. *Event Misattribution* tests whether visible events induce the model to hallucinate corresponding sounds. Finally, *Speaker Misattribution* evaluates whether the model incorrectly binds voices to on-screen identities on the basis of spurious lip-motion cues. Together, these dimensions probe whether an omni-modal LLM can correctly attribute, prioritize, and delimit acoustic evidence under competing multimodal context.

From the same failure taxonomy, we construct four complementary annotation pipelines (Figure 2) that turn naturally occurring video clips into targeted grounding supervision. For *Subtitle-Induced Substitution*, we select clips with clear, meaningful speech and no existing subtitles. We produce a transcript of the speech and translate it into another language. The translation is then aligned to

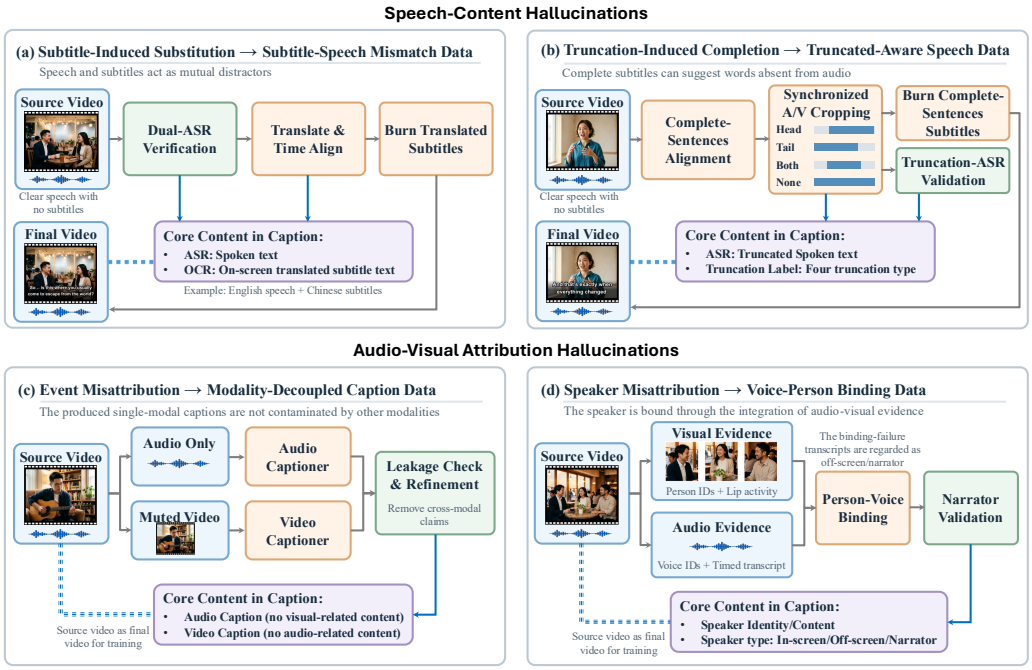


Figure 2: Four annotation pipelines used to construct the targeted audio-grounding data.

word-level timestamps and rendered onto the video. The resulting caption supervision requires the model to distinguish ASR content from optical character recognition (OCR) content, with speech and on-screen text serving as controlled cross-modal distractors. For *Truncation-Induced Completion*, we draw from the same video pool and retain only clips containing complete sentences. We then apply head, tail, or two-sided truncation to both audio and video, keeping some clips untruncated as controls. We retain samples that pass dual-ASR agreement and transcription-error filtering. The full-sentence subtitle is then burned into each truncated clip. The resulting supervision requires the model to transcribe only the observed speech rather than complete it from the subtitle or linguistic priors.

The remaining pipelines target audio-visual attribution hallucinations. For *Event Misattribution*, we separate each video into an audio-only stream and a muted visual stream. Each stream is independently annotated by a dedicated modality-specific captioning model. We then filter cross-modal leakage, retaining audio captions without visual claims and visual captions without auditory claims. For *Speaker Misattribution*, we extract on-screen identities and lip activity from the muted visual stream, and voice identities and timed transcripts from the audio stream. We integrate these cues for strict person-voice binding. Unmatched speech segments are labeled as off-screen speech or narration, and the resulting assignments are checked by a separate validation module. The final captions explicitly associate speech content with the corresponding speaker.

3.2 Stage-Aware Progressive Omni-Modal Training

Following recent works (Xu et al., 2025b; An et al., 2026), we decompose our omni-modal training pipeline into three distinct stages: audio alignment, multimodal mid-training, and SFT.

The stage at which grounding supervision is introduced is a central design choice. Mid-training is the earliest stage at which the audio, visual, and language pathways are jointly trainable, making it the natural point for shaping their relative contributions. We therefore introduce targeted grounding supervision during mid-training to shape cross-modal reliance rather than merely correct isolated failure modes; Section 4.4 validates the resulting shift toward acoustic attention and its transfer to public omni-modal benchmarks.

Motivated by this stage dependence, we assign a distinct optimization role to each stage (Figure 3). Audio alignment first establishes a reliable interface between acoustic representations and the language model. Mid-training jointly optimizes all modality pathways and, crucially, introduces our targeted audio-grounding supervision while cross-modal reliance is still being established, thereby strengthening the contribution of acoustic evidence. SFT then calibrates instruction following, response granularity, and perceptual coverage, consolidating the learned grounding behavior rather than attempting to reshape established modality dependencies. This functional separation makes the placement of supervision an explicit part of the training design, rather than treating the three stages as a generic progression of data and compute.

Audio Alignment: Establishing acoustic access. We first align the audio encoder with the LLM while preserving the pretrained language and vision capabilities. This stage proceeds in two phases. In the first phase, we freeze the audio encoder, the LLM, and the visual pathway, and train only the audio projector to map fixed acoustic representations into the LLM's embedding space. In the second phase, we unfreeze the audio encoder and optimize it jointly with the projector, while keeping the LLM and visual pathway frozen, so that the acoustic representations can adapt to the language interface. Throughout both phases, we follow a curriculum along three coupled axes: from short to long audio, from concise to detailed descriptions, and from single-dimensional to multi-dimensional supervision. This progression establishes a stable acoustic interface before exposing the model to increasingly long and compositionally rich audio content.

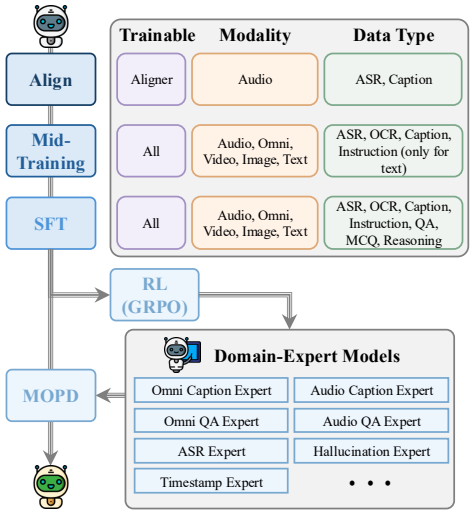


Figure 3: Audite-Omni-4B training pipeline.

Mid-Training: Shaping modality contribution. After establishing acoustic access, we jointly optimize all model parameters so that audio, vision, and language representations can adapt to one another. To preserve the pretrained perceptual representations during this transition, we assign lower learning rates to the audio and vision encoders and their projectors. The training corpus directly combines cleaned general-purpose omni-modal data with the four types of targeted grounding data introduced above. The general omni-modal portion primarily comprises audio-video captioning, as well as modality-specific tasks over full audio-visual inputs (e.g., audio-video ASR, OCR, and single-modality captioning) where the model is required to describe only one stream despite receiving both. It is further filtered according to the same four failure criteria to avoid reinforcing the shortcuts targeted by our specialized data.

Crucially, as synthesizing targeted ASR data for Subtitle-Induced Substitution and Truncation-Induced Completion is substantially more scalable than constructing exhaustive full-caption examples for these failure modes, we adopt an instruction-level nesting strategy to facilitate cross-task generalization. Specifically, the standalone audio-video ASR instruction is inserted verbatim as a constituent sub-instruction within the comprehensive audio-video captioning prompt. The targeted ASR examples thus supervise the same core speech-transcription instruction encountered during broad captioning. This design encourages the model to transfer its robust speech grounding—learned efficiently from the focused ASR task—directly into free-form captions that jointly describe visual and acoustic content (Ablations in Appendix C.3). Mid-training therefore serves as the stage at which the model learns how much to rely on each modality and effectively generalizes this grounding capability across task formats, rather than merely acquiring access to individual modality features. We further evaluate its broader contribution through attention analysis and public omni-modal benchmarks in Section 4.4.

Instruction Tuning: Calibrating task behavior. We next perform full-parameter SFT with a uniform learning rate across all modules, focusing on calibrating how the model follows instructions and expresses its multimodal understanding. The SFT corpus is structured with a 4:6 ratio between

captioning data and instruction-following data. Instruction examples are crucial for improving task adherence and response formatting, while caption examples preserve broad perceptual coverage and ensure the model produces outputs at the appropriate level of detail. The caption portion is strictly curated from the mid-training corpus to provide higher-quality supervision. Accordingly, this stage primarily calibrates instruction following, response granularity, and output organization, rather than fundamentally relearning the relative contributions of modalities formed during mid-training.

3.3 Task-Specialized GRPO and Multi-Teacher Integration

To further improve the supervised checkpoint, we perform RL post-training. At the 4B scale, jointly optimizing all target capabilities with a single RL objective is challenging given substantial differences in output format, reward scale, and training dynamics. In a pooled RL run, tasks with stronger or faster-improving reward signals can dominate optimization, limiting gains or reducing performance on other capabilities. Inspired by the specialist-then-integration strategies of MiMo-V2.5 (Xiaomi MiMo Team, 2026) and Kimi K3 (Kimi Team, 2026), we independently train 12 domain experts with GRPO (Shao et al., 2024), as detailed in the Appendix A.4. We then consolidate these experts into a single generalist through MOPD (Ma et al., 2026a), as illustrated in Figure 3. Compared with pooled multi-domain RL, this strategy produces more balanced improvements across capabilities, as verified by the ablations in Section 4.4.

4 Experimental Results

4.1 Experimental Setup

Models and baselines. Audite-Omni-4B is built on Qwen3.5-4B-Base (Qwen Team, 2026a), with the Qwen3-ASR (Shi et al., 2026) Audio Encoder providing its acoustic front end. For general capability evaluation, we compare it against representative open-source omni-modal LLMs spanning different architectures and scales, including Qwen2.5-Omni (Xu et al., 2025a), MiniCPM-o-4.5 (Cui et al., 2026), Gemma-4-E4B (Gemma Team et al., 2026), AV-Flamingo (Ghosh et al., 2026), and Qwen3-Omni-30B-A3B (Xu et al., 2025b). We additionally evaluate leading proprietary systems on Audite-CapsBench to probe the upper bound of audio grounding. This closed-source suite includes the Gemini-2.5/3.1/3.5/3.8 series (Gemini Team, 2023), Qwen3.5/3.8-Omni variants (Qwen Team, 2026b), and Doubao-Seed-2.0-Lite (ByteDance Seed, 2026). For each baseline, we follow its default generation configuration. Thinking is enabled only for STEM benchmarks and is disabled for all other evaluations; the same task-dependent setting is applied to Audite-Omni-4B. Detailed training and inference configurations are provided in Appendix A.

Audite-CapsBench. Building on the annotation pipelines described in Section 3.1, we apply a rigorous human annotation and quality inspection process, retaining 999 high-quality videos for evaluation. The benchmark contains 250 examples for Subtitle-Induced Substitution (SIS), 250 for Truncation-Induced Completion (TIC), 249 for Event Misattribution (EM), and 250 for Speaker Misattribution (SM). All four sub-benches use the same comprehensive audio-video captioning instruction. This ensures that the benchmark evaluates audio grounding organically within free-form generation, rather than through potentially biased task-specific question formats. During evaluation, the judge reads only the model-generated caption and has no access to the original audio or video. Using a criterion-specific prompt, an LLM judge determines whether the caption exhibits the target grounding failure—for instance, whether it transcribes conflicting subtitles instead of the actual spoken content, hallucinates speech beyond the observed acoustic segment, reports a visually implied but acoustically absent sound, or incorrectly attributes an off-screen voice to a visible person. We report accuracy on each sub-bench and use their unweighted mean as the overall score. The complete judge prompts, label sets, and scoring rules are detailed in Appendix B.

Public benchmarks. To systematically benchmark the model's core perceptual and cognitive abilities, we evaluate Audite-Omni-4B on a comprehensive suite of public benchmarks. Due to the extensive scale of this evaluation (encompassing over 40 distinct datasets), Table 2 presents high-level capability group averages for the audio and vision modalities, while reporting original scores for the 11 independent omni-modal benchmarks; a complete list of datasets and per-benchmark results is deferred to Appendix C.1 and Appendix A.5. Specifically, the audio evaluation is aggregated into ASR, Audio Understanding, and Audio Captioning. The vision evaluation is aggregated

Table 1: Results on **Audite-CapsBench**. SIS: Subtitle-Induced Substitution; TIC: Truncation-Induced Completion; EM: Event Misattribution; SM: Speaker Misattribution. The overall score is the unweighted mean across the four sub-benchmarks. All scores are accuracies (%) where higher is better. **Bold** denotes the best result in each column, and underline denotes the second best.

	Speech-Content Hallucinations		Audio-Visual Attribution Hallucinations		
Model	**SIS**	**TIC**	**EM**	**SM**	**Overall**
Closed-Source Models					
Qwen3.5-Omni-Flash	69.2	27.6	72.1	63.6	58.1
Qwen3.5-Omni-Plus	86.4	39.2	69.9	71.2	66.7
Qwen3.8-Omni-Flash	94.4	45.6	77.7	73.2	72.7
Gemini-3.5-Flash	88.4	24.8	74.7	78.8	66.7
Gemini-3.8-Flash	84.0	26.0	72.9	80.8	65.9
Gemini-2.5-Pro	76.0	20.8	72.7	74.8	61.1
Gemini-3.1-Pro	70.0	16.4	73.5	69.6	57.4
Doubao-Seed-2.0-Lite	80.8	21.2	**82.9**	80.4	66.3
Open-Source Omni Models					
Qwen2.5-Omni-7B	4.0	7.6	65.5	19.6	24.2
MiniCPM-o-4.5-9B	16.8	18.8	67.3	61.2	41.0
Gemma-4-E4B	12.4	19.2	57.6	20.4	27.4
AV-Flamingo-7B	59.2	58.0	70.9	62.4	62.7
Qwen3-Omni-30B-A3B	39.6	14.0	74.3	65.2	48.3
Ours					
Audite-Omni-4B	**98.0**	**97.2**	81.9	**82.4**	**89.9**

into General Visual Question Answering (VQA), Document & Chart & Counting, Spatial Intelligence, STEM & Puzzle Solving, and Video Understanding. The omni-modal benchmarks cover both question answering (e.g., Daily-Omni, AV-SpeakerBench, Video-MME) and free-form captioning (Video-SALMONN2, Omni-Cloze, and AVSCapBench).

4.2 Audio Grounding on Audite-CapsBench

To provide a comprehensive analysis of audio grounding, we examine the models' performance across two primary failure modes. Qualitative examples illustrating these specific hallucination patterns are provided in Appendix B.2.

Speech-Content Hallucinations. We first evaluate the model's robustness against visual text distractors, measured by Subtitle-Induced Substitution (SIS) and Truncation-Induced Completion (TIC). As shown in Table 1, Audite-Omni-4B achieves an overall score of **89.9** across the benchmark. It nearly eliminates speech-content hallucinations, reaching near-perfect accuracy on both SIS and TIC, outperforming the strongest open-source baseline by over 38 points. The exceptionally high performance on these splits can be attributed to the highly deterministic nature of ASR: once a model learns to strictly prioritize the acoustic signal over visual priors, it can almost perfectly avoid transcribing unplayed text. In contrast, baseline models frequently succumb to these visual distractors. Notably, this vulnerability is not restricted to open-source models; even the latest proprietary models like Gemini-3.8-Flash and Qwen3.8-Omni-Flash struggle severely with TIC, indicating a pervasive tendency across current omni-modal LLMs to hallucinate speech based on visual text. For instance, when presented with foreign language audio paired with English subtitles, or when the audio is abruptly truncated while the subtitle continues on screen, baselines routinely hallucinate by reading the burned-in text as spoken audio or completing the unplayed speech. Audite-Omni-4B consistently respects the physical acoustic boundaries and spoken language.

Audio-Visual Attribution Hallucinations. Beyond speech transcription, we assess the model's ability to correctly attribute sounds to their sources under misleading visual contexts via Speaker Misattribution (SM) and Event Misattribution (EM). Audite-Omni-4B significantly improves over prior open-source models by raising accuracy beyond 80% on both SM and EM. Resolving audio-visual conflicts in these scenarios is inherently more challenging than speech transcription, as it requires the model to correctly interpret ambiguous visual cues without letting them override the acoustic reality. While a leading closed-source model like Doubao-Seed-2.0-Lite demonstrates strong resilience across these tasks, most baselines struggle to maintain high accuracy on both attribution scenarios. When visual cues strongly suggest a sound that is not actually present in the

Table 2: Open-source Audio, Vision, and Omni benchmark results. Audio and Vision rows report capability-group averages (ASR: mean of $1 - \mathrm{WER/CER}$; others: mean accuracy), while Omni rows report individual benchmarks. ChronusAV-Ground is the mean of a2t and v2t R@0.5. **Bold** and underline denote the best and second-best results in each row; ↓ indicates that lower is better.

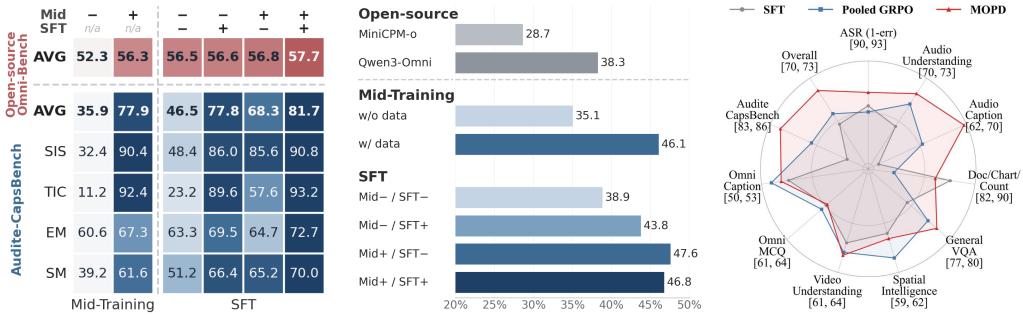
Benchmark	*Base model*	*Open-source omni models*					*Ours*
Model	Qwen3.5-Base	Qwen2.5-Omni	MiniCPM-o 4.5	Gemma-4-E4B	AV-Flamingo	Qwen3-Omni	Audite-Omni
Params	4B	7B	9B	E4B	7B	30B-A3B	4B
Audio							
ASR	-	89.7	80.0	67.0	46.8	**94.6**	91.6
Audio Understanding	-	69.9	71.1	59.0	69.0	**74.8**	72.0
Audio Caption	-	60.7	47.2	36.5	61.1	66.5	**71.9**
Vision							
General VQA	73.3	75.4	75.9	62.6	67.2	**79.8**	79.3
Document & Chart & Count	78.0	86.5	86.4	68.1	72.5	87.6	**88.1**
Spatial Intelligence	60.7	53.6	56.1	52.9	38.9	53.9	**62.6**
STEM and Puzzle	65.8	52.8	56.5	55.4	43.1	**65.8**	64.9
Video Understanding	59.3	59.6	58.5	45.6	55.4	63.4	**63.5**
Omni QA							
Daily-Omni	-	63.7	74.5	51.5	70.8	71.2	**83.0**
WorldSense	-	47.5	51.4	36.2	50.2	54.6	**58.8**
AVUT (Human)	-	65.5	68.6	51.9	66.0	75.4	**81.6**
AV-SpeakerBench	-	46.5	50.2	35.6	43.5	55.4	**65.0**
Video-MME (w/ audio)	-	65.7	53.1	53.8	68.3	77.2	**77.3**
Video-MME-v2 (w/ audio)	-	24.3	23.3	22.7	24.8	35.8	**36.7**
OmniVideoBench	-	34.4	36.1	30.5	39.0	**43.2**	42.9
ChronusAV-Ground (R@0.5)	-	8.7	6.9	14.3	4.4	42.3	**72.4**
Omni Caption							
Video-SALMONN2 (↓)	-	67.0	49.9	61.2	57.7	56.1	**26.0**
Omni-Cloze	-	18.8	28.8	20.2	25.8	32.5	**43.8**
AVSCapBench	-	10.6	16.3	13.2	16.8	16.2	**45.1**

audio track (such as an on-screen person appearing to speak or a visual event that typically produces noise), baselines frequently fall into attribution traps. They erroneously bind off-screen voices to visible individuals or hallucinate acoustically absent environmental sounds. Audite-Omni-4B maintains strict modality independence in these complex scenarios, ensuring that visually implied events do not override acoustic evidence.

4.3 General Capability on Public Benchmarks

Audio and vision capabilities. As shown in Table 2, Audite-Omni-4B achieves a robust balance across unimodal tasks. In the audio domain, it delivers the strongest captioning performance among evaluated models, outperforming the substantially larger Qwen3-Omni-30B-A3B. Crucially, this generative strength does not compromise foundational acoustic perception, as the model also secures top-tier rankings in both ASR and Audio Understanding. A compelling pattern emerges in the vision domain: despite the parameter constraints of a 4B backbone, introducing the audio pathway avoids modality competition and catastrophic forgetting. Instead, relative to its Qwen3.5-4B-Base initialization, omni-modal training improves four of the five vision groups. It establishes state-of-the-art results among open-source models across key domains, including Document & Chart & Count, Spatial Intelligence, and Video Understanding, demonstrating that our training pipeline effectively preserves and synergizes visual capabilities.

Omni-modal reasoning and captioning. Audite-Omni-4B demonstrates exceptional multimodal integration, securing the top position on ten of the eleven omni-modal benchmarks, often outperforming models with significantly larger parameter counts. In question-answering tasks, it exhibits strong performance across the suite, particularly on Daily-Omni and AV-SpeakerBench. Notably, the substantial improvements on ChronusAV-Ground confirm that our grounding-oriented training inherently equips the model with superior temporal alignment capabilities. This advantage extends to free-form captioning, where the model leads across all three benchmarks. By notably reducing the error rate on Video-SALMONN2 and achieving the highest scores on Omni-Cloze and AVS-CapBench, these results collectively demonstrate that Audite-Omni-4B achieves broad and robust omni-modal capabilities at a highly efficient 4B scale.



(a) Stage-wise data injection (b) Audio Attention share (c) MOPD versus pooled GRPO

Figure 4: Ablation studies on grounding supervision and alignment strategies. **(a)** Impact of injecting targeted grounding data $(+/-)$ at the Mid-Training and SFT stages, shown on public omni-modal benchmarks (top) and Audite-CapsBench (bottom). Cell colors are row-normalized. **(b)** Audio attention share $A/(A+V)$ computed over 800 random audio–visual clips using an audio-focused description prompt. **(c)** Radar chart comparing capability retention across SFT, pooled GRPO, and our MOPD strategy. Axes use fixed 3-point windows (8 points for Audio Caption and Doc/Chart/Count).

4.4 Ablation Studies

When to inject grounding supervision. To determine the optimal stage for introducing grounding supervision, we evaluate a 2×2 design across Mid-Training and SFT under a matched total data budget. As shown in Figure 4(a), the two stages play distinct, non-interchangeable roles across both Audite-CapsBench and public omni-modal benchmarks. Mid-Training establishes the foundational acoustic priority, notably raising overall performance and substantially increasing the audio attention share, $A/(A+V)$, to surpass leading open-source models (Figure 4(b)). However, SFT serves as a critical consolidation phase: if grounding data is removed after Mid-Training, the model suffers a severe relapse in truncation faithfulness, indicating that the initial acoustic priority is easily unlearned. By continuously injecting targeted supervision across both stages, the model achieves the highest overall scores on both Audite-CapsBench and the public suite. These consistent gains demonstrate that sustained grounding supervision improves audio faithfulness without sacrificing broader omni-modal generalization. Further analysis is provided in Appendix C.5.

MOPD versus pooled GRPO. We compare MOPD against a standard pooled GRPO (Mixed RL) under a strictly matched computational budget. Unlike pooled GRPO's joint optimization, MOPD decouples the learning process by training and subsequently integrating independent, task-specialized experts. As Figure 4(c) illustrates, pooled GRPO suffers from prominent reward conflicts in omni-modal settings, leading to catastrophic capability regression on unpenalized tasks (e.g., a 4.4-point drop on Document & Chart & Count). By isolating the optimization trajectories before integration, MOPD successfully circumvents these gradient conflicts. It prevents capability degradation and outperforms pooled GRPO on 7 out of 10 capability groups.

5 Conclusion

We investigate visually induced acoustic hallucinations in omni-modal LLMs, introducing controlled diagnostics (Audite-CapsBench) and targeted supervision pipelines for four representative failure modes. Our experiments reveal a strong stage dependency in multimodal learning: grounding supervision is profoundly more effective during Mid-Training, when cross-modal dependencies are initially formed. Once disproportionate modality reliance is established, late-stage SFT shows limited capacity to reverse it. Leveraging these insights, we develop Audite-Omni-4B through stage-aware progressive training and MOPD expert integration. The resulting model substantially mitigates visually induced acoustic hallucinations while preserving robust general capabilities, achieving leading performance across diverse audio-visual benchmarks even when compared to substantially larger models.

AI USE STATEMENT

Generative AI tools were used to assist with cleaning and reformatting datasets, drafting and debugging portions of the model-training code, and editing the manuscript for clarity and readability. They were not used to develop theoretical or conceptual frameworks, formulate mathematical claims or proofs, propose hypotheses, design the research methodology or experiments, conduct qualitative or thematic data analysis, or interpret experimental results. All AI-assisted dataset modifications were manually reviewed, all AI-assisted code was inspected and tested by the authors, and all manuscript revisions were checked for technical accuracy. The authors take full responsibility for the final text, claims, code, and artifacts.

ETHICS STATEMENT

This study involved no human or animal subjects and used no sensitive personal data; therefore, no ethical approval was required. We followed applicable ethical and legal standards and identified no material concerns regarding privacy, fairness, harmful use, conflicts of interest, or data release.

REPRODUCIBILITY STATEMENT

We provide detailed data-construction and annotation procedures in Section 3.1 and Appendix B. Model architecture, training configurations, expert RL, MOPD, and inference settings are documented in Appendix A, while evaluation protocols and additional results are reported in Section 4.1 and Appendix C. Upon acceptance, we will publicly release the model weights and complete evaluation code.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

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

Hanin Atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, and Bhiksha Raj. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23181–23203, 2025.

Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. Investigation of whisper asr hallucinations induced by non-speech audio. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

ByteDance Seed. Seed2.0 lite. Official Seed2.0 model page, 2026. URL `https://seed.bytedance.com/en/seed2`. Omni-modal update released in April 2026.

Yijing Chen, Yihan Wu, Kaisi Guan, Yuchen Ren, Yuyue Wang, Ruihua Song, and Liyun Ru. ChronusOmni: Improving time awareness of omni large language models. *arXiv preprint arXiv:2512.09841*, 2025.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. Aha-bench: Benchmarking audio hallucinations in large audio-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 736–740, 2020.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024.

Rita Frieske and Bertram E. Shi. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025a.

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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 75300–75320, 2025b.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, et al. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Siddharth Gururani, Hanrong Ye, Pritam Biswas, Yuanhang Su, Ehsan Hosseini-Asl, et al. Audio-visual flamingo: Open audio-visual intelligence for long and complex videos. *arXiv preprint arXiv:2607.16107*, 2026.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *International Conference on Learning Representations*, 2026.

Chao-Ling Hsu and Jyh-Shing Roger Jang. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310–319, 2010.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-MMMU: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pp. 235–251, 2016.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 119–132, 2019.

Sung-Bin Kim, Hyun-Bin Oh, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kimi Team. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1672–1681, 2024.

Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee. Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models. In *Proceedings of INTERSPEECH*, pp. 4144–4148, 2024.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Jiafu Tang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, Zhaohui Wang, Jiayi Tian, Yanghai Wang, Zhe Cao, Minxin Dai, Ke Wang, Runzhe Wen, Yinghao Ma, Yaning Pan, Sungkyun Chang, Termeh Taheri, Haiwen Xia, Christos Plachouras, Emmanouil Benetos, Yizhi Li, Ge Zhang, Jian Yang, Tianhao Peng, Zili Wang, Minghao Liu, Junran Peng, Zhaoxiang Zhang, and Jiaheng Liu. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *The Fourteenth International Conference on Learning Representations*, 2026.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19108–19118, 2022.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024a.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024b.

Yuhao Liang, Mohan Shi, Fan Yu, Yangze Li, Shiliang Zhang, Zhihao Du, Qian Chen, Lei Xie, Yanmin Qian, Jian Wu, Zhuo Chen, Kong Aik Lee, Zhijie Yan, and Hui Bu. The second multi-channel multi-party meeting transcription challenge (m2met 2.0): A benchmark for speaker-attributed asr. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–8, 2023.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference*, pp. 1140–1144, 2022.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233, 2024.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026a.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, Eng Siong Chng, and Xie Chen. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations*, 2026b.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021.

Hokuto Munakata, Taichi Nishimura, Shota Nakada, and Tatsuya Komatsu. Language-based audio moment retrieval. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, and Yong Jae Lee. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *arXiv preprint arXiv:2512.02231*, 2025.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180, 2023.

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*, pp. 5206–5210, 2015.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL `https://qwen.ai/blog?id=qwen3.5`.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026b.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Yang Shi, Huanqian Wang, Wulin Xie, Huanyao Zhang, Lijie Zhao, Yi-Fan Zhang, Xinfeng Li, Chaoyou Fu, Zhuoer Wen, Wenting Liu, et al. MME-VideoOCR: Evaluating OCR-based capabilities of multimodal LLMs in video scenarios. In *Advances in Neural Information Processing Systems*, 2025.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47198–47217. PMLR, 2024.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, and Xiangang Li. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4297–4316, 2025a.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025b.

Yanghai Wang, Jiahao Wang, Jiafu Tang, Yuanxing Zhang, Zhe Cao, Hanyan Bian, Zijie Zhang, Weiliang Luo, Zhiyu Pan, Zixuan Dong, Jiaheng Liu, and Zhaoxiang Zhang. AVSCap: Orchestrating audio-visual synergy for omni-modal video captioning. *arXiv preprint arXiv:2607.12820*, 2026b.

Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. In *Proceedings of Interspeech*, pp. 4242–4246, 2022.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. Muchomusic: Evaluating music understanding in multimodal audio-language models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

xAI. RealWorldQA. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Xiaomi MiMo Team. Mimo-v2.5. Official model report, 2026. URL https://mimo.xiaomi.com/mimo-v2-5/.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

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. *arXiv preprint arXiv:2503.20215*, 2025a.

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

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. Text-to-audio grounding: Building correspondence between captions and sound events. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 606–610, 2021.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3480–3491, 2022.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1979–1998, 2024.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6569–6587, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 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 *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6182–6186, 2022.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, Chenghua Lin, Wenhao Huang, and Jie Fu. CMMMU: A Chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

A Implementation Details

A.1 Model Architecture and Initialization

Audite-Omni-4B initializes its language and vision backbone from Qwen3.5-4B-Base (Qwen Team, 2026a), a 4B causal language model equipped with a native vision encoder. We retain the pretrained language and visual pathways, thereby inheriting the visual architecture and initialization of the base model without structural modification. To introduce acoustic perception, we initialize the audio front-end from the audio encoder of Qwen3-ASR-1.7B (Shi et al., 2026). A standard two-layer MLP projector, utilizing a GELU activation between the two linear layers, maps the resulting acoustic representations into the hidden space of the language backbone, allowing them to be processed jointly with the visual and textual representations.

A.2 Stage-Wise Training Configuration

Table 3 summarizes the optimization configurations used for the five supervised training stages preceding expert reinforcement learning. Across all stages, the model is optimized using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.1. Gradients are clipped at a maximum norm of 1.0 to ensure training stability. To efficiently support large-scale distributed training, our implementation is built upon the Megatron framework. Learning rates apply uniformly to all trainable modules unless otherwise specified.

Table 3: Stage-wise training configurations for Audite-Omni-4B. All stages are trained for one epoch. * During Mid-Training, the audio and vision encoders and their respective projectors use a learning rate of 3e-6; all remaining modules use the reported peak learning rate.

Configuration	**Alignment**		**Mid-Training**	**SFT**	
	Init	**CPT**	**Pretrain**	**Standard**	**Annealing**
Trainable Modules	Projector	Audio Encoder + Projector	Full Model	Full Model	Full Model
Training Tokens (B)	5.9	8.5	147.3	70.4	11.2
Micro Batch Size	4	4	1	1	1
Global Batch Size	512	512	1024	256	256
Max. Sequence Length	8K	8K	8K	32K	64K
Learning Rate	8e-5	1e-5	1e-5 / 3e-6*	1e-5	2e-6
LR Schedule	Constant	Cosine	Constant	Cosine	Cosine
Minimum LR	–	1e-7	–	1e-7	1e-7

A.3 Training Data

We organize the training corpus according to the objective of each stage, progressively moving from audio–language alignment to large-scale multimodal learning and instruction tuning. For data preprocessing, all continuous audio signals are uniformly resampled to 16 kHz to align with the acoustic encoder's input space. For video data, we dynamically sample frames based on the training stage: up to 36 frames are sampled per video during Mid-Training to balance temporal context and computational efficiency, while the maximum frame limit is increased to 128 (and up to 256 for long-context tasks) during Supervised Fine-Tuning (SFT) and SFT Annealing to enhance fine-grained temporal understanding.

Audio Alignment I. This stage warms up the audio projector to align the representations produced by the audio encoder with the input space of the language model. The training mixture encompasses multilingual automatic speech recognition (ASR), sound captioning, music captioning, emotion recognition, and spoken-language identification, providing broad supervision across speech, environmental sound, and music understanding.

Audio Alignment II. This stage continues training the audio encoder together with the audio projector, strengthening their perception of environmental sounds, acoustic effects, and background music. In addition to the tasks used in Audio Alignment I, we introduce longer-form ASR data and merged-caption data. Each merged caption provides a unified description of all salient audio information, jointly covering speech content, sound events, background music, emotion, and language when present.

Mid-Training. This stage introduces joint audio–visual training and develops the model's ability to understand and describe interleaved visual and acoustic information. The training mixture is centered on audio–visual data and includes audio–visual captioning, OCR and ASR on videos, video-only captioning, image captioning, text-based instruction following, and lightweight reasoning data. We also replay the audio tasks from the preceding alignment stages to preserve the acquired acoustic capabilities. A significant proportion of text-only data is retained to maintain instruction following, world knowledge, and general reasoning, whereas empirical observations indicate that a smaller, curated subset of image data is sufficient to preserve strong image understanding.

Supervised Fine-Tuning. This stage strengthens instruction following and elicits the capabilities acquired during pretraining. Building on the task mixture used in Mid-Training, we increase the proportion of instruction-tuning data and introduce diverse question answering, multiple-choice, and structured JSON-generation tasks over audio–visual, audio-only, and image inputs. Text-only and audio–visual examples constitute the primary components of the training mixture, followed by image examples and, finally, audio-only and video-only examples. This distribution reflects current data availability: image instruction data are comparatively easy to construct and abundant in open-source resources, whereas high-quality instruction data for audio-only and video-only inputs remain substantially scarcer.

SFT Annealing. The final annealing stage further optimizes the model for long-document and long-video understanding. From the SFT mixture, we retain approximately 10% of the data through joint filtering based on dataset quality and model perplexity, and augment this subset with additional long-document and long-video examples. This construction strengthens long-context comprehension while concentrating optimization on higher-quality examples. The modality proportions remain broadly consistent with those used during standard SFT.

A.4 Expert RL and MOPD

As described in Section 3.3, we train 12 task-specialized experts before MOPD integration. Table 4 groups them into three families: ASR and captioning, question answering and reasoning, and specialized grounding tasks. To mitigate cross-task competition in pooled optimization, we train separate QA experts for the more difficult music, long-video, document, and reasoning settings. In particular, music QA requires domain-specific acoustic and musical knowledge; long-video QA requires temporal integration over extended audio-visual context; document QA emphasizes OCR and layout understanding; and reasoning QA targets multi-step inference beyond direct perception.

Reward Functions. Table 4 summarizes the reward rules used across the experts. *Transcript-similarity* assigns a continuous score in $[0, 1]$ by comparing the model output with the reference transcript. *Caption-MCQ* follows the utility-based caption evaluation of CapRL (Xing et al., 2026): for each audio or video, we construct a multiple-choice question (MCQ), ask the policy to generate a free-form caption, and let a separate LLM answer the MCQ using only that generated caption. The reward is 1.0 if the selected option is correct and 0.0 otherwise. *QA-MCQ* uses the same principle for question answering: the policy produces an open-ended response, which a judge LLM maps to one of the candidate options for binary correctness scoring. We avoid requesting an option directly to reduce spurious gains from chance guessing, and find this open-response formulation empirically more effective.

The specialized rewards for the Speech-Content Expert and Audio-Visual Attribution Expert follow similar grounding objectives. For the Speech-Content Expert, a judge LLM first verifies sample-specific requirements, such as recognizing whether speech is truncated or identifying the spoken language correctly. A failed judgment yields a reward of 0.0; otherwise, the output is further scored

Table 4: Task-specialized experts and their corresponding reward functions for GRPO.

Expert	Data Modality	Reward Rule
ASR and Captioning		
ASR Expert	Audio / Omni	Transcript-similarity
Audio Caption Expert	Audio	Caption-MCQ
Omni Caption Expert	Omni	Caption-MCQ
Question Answering and Reasoning		
Audio General QA Expert	Audio	QA-MCQ
Omni General QA Expert	Omni	QA-MCQ
Audio Music QA Expert	Audio	QA-MCQ
Omni Long-Video QA Expert	Omni	QA-MCQ
Document QA Expert	Image	QA-MCQ
Reasoning QA Expert	Image / Omni	QA-MCQ
Specialized Data		
Timestamp Expert	Omni	Temporal-IoU
(Type-A) Speech-Content Expert	Audio / Omni	LLM-judge + Transcript-similarity
(Type-B) Audio-Visual Attribution Expert	Omni	Caption-MCQ

by *Transcript-similarity*. For the Audio-Visual Attribution Expert, the reward reformulates sound-event and source-attribution annotations as MCQs. The policy generates an audio-visual caption, and the judge LLM answers these MCQs strictly from the caption alone, yielding the same binary reward format as *Caption-MCQ*. For all LLM-based evaluation and scoring, we utilize the `doubao-seed-2-0-lite-260215` model. Finally, *Temporal-IoU* assigns a continuous temporal Intersection over Union (tIoU) score in $[0, 1]$ for temporal grounding tasks.

Training Configuration. All expert and MOPD runs update the full model in BF16 from the same Audite-Omni-4B SFT checkpoint. Each expert is trained on 3.2K prompts, while MOPD uses the combined $3.2\text{K} \times 12$ prompt set. The remaining optimization, sampling, and objective settings are reported in Table 5.

A.5 Details of Public Benchmarks

We evaluate general capability on a broad collection of public audio, vision, and omni-modal benchmarks.

Audio Suite. The audio evaluation encompasses:

- **Speech Recognition (ASR):** LibriSpeech-clean/other (Panayotov et al., 2015), CommonVoice15 (en, zh-CN, ja, ko, zh-HK, and zh-TW) (Ardila et al., 2020), WenetSpeech-meeting/net (Zhang et al., 2022), Opencpop (Wang et al., 2022), MIR-1K (Hsu & Jang, 2010), and KeSpeech (Tang et al., 2021).
- **Audio Understanding:** MMAR (Ma et al., 2025), MuChoMusic (Weck et al., 2024), MMAU-mini (Sakshi et al., 2025), and MMSU (Wang et al., 2026a).
- **Audio Captioning:** AudioCaps (Kim et al., 2019) and ClothoAQA (Lipping et al., 2022).

ASR scores are converted to $1 - \text{WER}/\text{CER}$ before aggregation.

Vision Suite. The vision evaluation encompasses:

- **General VQA:** MMBench (Liu et al., 2024) and RealWorldQA (xAI, 2024).
- **Document, Chart & Counting:** AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), OCRBench (Liu et al., 2023), DocVQA (Mathew et al., 2021), and CountBench (Paiss et al., 2023).
- **Spatial Intelligence:** ERQA (Gemini Robotics Team et al., 2025) and EmbSpatial-Bench (Du et al., 2024).

Table 5: Training hyperparameters for task-specialized GRPO and MOPD.

Hyperparameter	GRPO Experts	MOPD
Data and Model		
Training data	3.2K prompts per expert	3.2K $\times$ 12
Trainable parameters	All	All
Starting model checkpoint	Audite-Omni-4B (SFT)	Audite-Omni-4B (SFT)
Precision	BF16	BF16
Optimization		
Optimizer	AdamW	AdamW
Adam β_1	0.9	0.9
Adam β_2	0.95	0.95
Adam ϵ	10^{-8}	10^{-8}
Weight decay	0.1	0.1
Gradient clipping	1.0	1.0
Maximum learning rate	2×10^{-6}	1×10^{-6}
LR schedule	Cosine	Cosine
Warm-up ratio	0.1	0.1
Epochs	3	3
Global batch size	512	512
Sampling		
Rollouts per prompt	16	1
Prompts per generation batch	32	512
Sampling temperature	1.0	1.0
Top-p	1.0	1.0
Video sampling rate	4 FPS	4 FPS
Maximum video frames	64	64
Video token budget	400	400
Objective Coefficients		
GRPO reference-KL coefficient β_{RL}	0.01 (caption); 0 otherwise	–
MOPD on-policy coefficient λ	–	1.0
MOPD divergence coefficient β_{MOPD}	–	1.0 (reverse KL)
MOPD teacher-logit top-k	–	64

- **STEM & Puzzle Solving:** MMMU (Yue et al., 2024), CMMMU (Zhang et al., 2024), LogicVista (Xiao et al., 2024), MathVista (Lu et al., 2024), and VideoMMMU (Hu et al., 2025).
- **Video Understanding:** Video-MME (Fu et al., 2025a), MVBench (Li et al., 2024a), LVBench (Wang et al., 2025b), and MME-VideoOCR (Shi et al., 2025).

Omni-Modal Suite. For omni-modal understanding, we evaluate:

- **Question Answering:** Daily-Omni (Zhou et al., 2025), WorldSense (Hong et al., 2026), AVUT (Yang et al., 2025), AV-SpeakerBench (Nguyen et al., 2025), Video-MME with audio (Fu et al., 2025a), Video-MME-v2 (Fu et al., 2026), OmniVideoBench (Li et al., 2026), and ChronusAV-Ground (Chen et al., 2025).
- **Free-Form Captioning:** Video-SALMONN2 (Tang et al., 2025), Omni-Cloze (Ma et al., 2026b), and AVSCapBench (Wang et al., 2026b). Note that for Video-SALMONN2, a lower score indicates better performance (error rate).

A.6 Inference Configuration

Audite-Omni-4B. Unless otherwise specified, we evaluate our model using deterministic greedy decoding with a temperature of 0 and top-p of 1.0. Chain-of-thought generation (referred to as *thinking*) is disabled by default and enabled exclusively for STEM-related tasks. For tasks that require long-form outputs, including ASR, caption generation, and tasks with thinking enabled, we apply a repetition penalty of 1.05 to mitigate repetitive generation. For video inputs, we dynamically

sample frames at 4 FPS up to a maximum of 128 frames, with each frame capped at 768 visual tokens.

Open-Source Baselines. For open-source baselines, we maintain their officially recommended generation hyperparameters and inference templates. To ensure a fair comparison under comparable visual input budgets, we enforce the same limits of 128 video frames and 768 visual tokens per frame. Exceptions are made only for models with hard-coded frame sampling implementations (e.g., Gemma 4). Regarding chain-of-thought evaluation, we activate the reasoning capabilities of MiniCPM-o 4.5 and Qwen3.5-4B-Base by explicitly enabling their designated configuration parameters (e.g., `enable_thinking = True`) during generation, mirroring the inference approach of Audite-Omni-4B. In contrast, for Qwen3-Omni-30B-A3B, we explicitly evaluate its distinct reasoning checkpoint (Qwen3-Omni-30B-A3B-Thinking) for these tasks.

B AUDITE-CAPSBENCH EVALUATION PROTOCOL

B.1 JUDGING AND SCORING

Caption generation. Audite-CapsBench evaluates whether an omni-modal model remains grounded in the acoustic evidence when the visual stream, on-screen text, or linguistic prior provides a plausible but conflicting cue. All evaluated models receive the corresponding video clip, including its audio track, and all four sub-benchmarks use exactly the same task instruction:

> *Watch and listen to the entire video, then write ONE cohesive, detailed caption that integrates both what is visually shown and what is heard into a faithful, information-rich description.*
>
> *Cover the following:*
>
> *1. Visual content: scene, environment, people, objects, actions, on-screen text/subtitles, and camera work.*
>
> *2. Audio content: all sounds including speech, music, and environmental sounds. For any speech, transcribe it verbatim and place it in quotation marks, naturally embedded in a sentence that describes who is speaking.*
>
> *Be thorough and precise. Describe only what you actually observe and hear.*

The instruction deliberately requests a single free-form caption rather than a task-specific answer. Consequently, a model must jointly describe the visual and acoustic streams without being told which controlled conflict is under evaluation. We use the same instruction for every model to prevent prompt differences from confounding the comparison.

Caption-only judging. We evaluate the generated captions with `doubao-seed-2-0-lite-260215`. The judge is given the caption and the sample-specific textual references required by the corresponding criterion, but never the original video or audio. Thus, the judge measures what the model explicitly claims in its output rather than independently solving the audio–visual task. All judge prompts are in English and restrict the response to a small, criterion-specific label set. Responses outside the prescribed format are counted as incorrect.

Shared claim extraction for SIS, TIC, and SM. The three speech-related sub-benchmarks first use a common extraction step. The judge is instructed to copy the claims made by the caption rather than decide whether those claims are true. Specifically, it extracts SPEAKER, the source to which the caption attributes the voice, and HEARD, the words that the caption presents as spoken or sung audio. The complete template is:

```
A caption of a video is below. Extract TWO fields. Do not judge
whether the caption is correct -- copy what it claims.

{hint}

SPEAKER -- pick exactly ONE label. Do not invent a new label.
- unclear -- the caption never says where the voice comes from
- off-screen -- a voice off camera, not labelled as narration
- narrator -- voice-over / narration / commentary
- BGM -- singing or voice is background music / a backing track
- on-screen: <who> -- the heard words are bound to a person VISIBLE
  in the frame. Copy that person from the caption (example:
  on-screen: the woman in the purple shirt). A person who is only
  described visually (standing, gesturing, looking at the camera,
  "addressing the group") is NOT on-screen unless the caption says
  THEY are the one speaking or singing.

HEARD -- the words the caption presents as spoken or sung audio,
copied verbatim. Do NOT include on-screen subtitle / burned-in text
that the caption describes as text on screen. Use EMPTY if no speech
is quoted.

Reply with exactly two lines and nothing else:
```

```
SPEAKER: unclear | off-screen | narrator | BGM | on-screen: <who>
HEARD: <quoted speech or EMPTY>

Caption:
{caption}
```

The placeholder {hint} is instantiated with a task-specific clarification. For SM, we use:

```
If the caption names an off-screen / narrating / voice-over / BGM
source AND also describes a visible person, SPEAKER is off-screen,
narrator, or BGM -- not on-screen. Use on-screen: <who> only when the
caption binds the heard words to that visible person.
```

For SIS, we use:

```
If the caption transcribes speech AND separately quotes a
burned-in subtitle, HEARD is only the speech the caption presents as
heard. Leave the subtitle text out of HEARD.
```

For TIC, we use:

```
If the caption transcribes speech AND separately quotes a
burned-in subtitle, HEARD is only the speech. Leave the subtitle text
out of HEARD.
```

The extracted fields are parsed deterministically. If the initial response does not follow the required format, the judge is prompted once more to return exactly these two fields. If the required field still cannot be recovered, the example is counted as incorrect. SIS and TIC then invoke a second, criterion-specific judgment that sees only HEARD and the relevant references, never the original caption. SM instead applies a deterministic rule directly to SPEAKER.

Subtitle-Induced Substitution (SIS). Each SIS example contains speech in language L_i and a burned-in subtitle in a different language D_i, with both expressing the same underlying content. Let A_i be the reference transcript of the audio, B_i the subtitle text, and H_i the extracted HEARD field. The verdict prompt is:

```
In this video the audio is spoken in {lang} while a
{distractor_lang} subtitle is burned into the frame. They are the
same words in two languages.

A -- what the audio actually says ({lang}):
{gt_asr}

B -- the subtitle text on screen ({distractor_lang}):
{gt_ocr}

An extractor pulled the caption's HEARD speech (ignore SPEAKER):
{heard}

Pick ONE label. Coverage of A is COVERAGE ONLY -- per-word accuracy
is NOT being graded. A unit of A counts as present even if a word
inside it is misheard, paraphrased, or written differently (number
style, contraction, romaji vs kana, simplified vs traditional,
filler).

PASS -- HEARD is {lang} and matches A, and every part of A is there
(a handful of wrong or extra words is fine).
MISSING -- HEARD is {lang} and comes from A, but a large part of A is
absent (roughly half or more), or the speech was rendered in another
language instead of {lang}.
SUBTITLE -- HEARD is {distractor_lang} and matches B: the model read
the subtitle and passed it off as speech.
MIXED -- content from A and from B are both presented as heard speech.
```

```
NEITHER -- it matches neither, or is empty.

Answer with one word: PASS, MISSING, SUBTITLE, MIXED, or NEITHER.
```

The five labels distinguish complete acoustic grounding from incomplete acoustic coverage and from reliance on the visual subtitle. Minor transcript discrepancies (e.g., paraphrases, number formatting, contractions, fillers, and script variants) do not reduce coverage. If H_i is empty, the example is counted as incorrect without a second judgment. A response outside the five permitted labels is treated as NEITHER. Let ℓ_i^{SIS} denote the returned label; invalid or missing extractions are, by definition, not PASS. The primary SIS score is

$$\mathrm{Acc}_{\mathrm{SIS}} = \frac{1}{N_{\mathrm{SIS}}} \sum_{i=1}^{N_{\mathrm{SIS}}} \mathbb{I}[\ell_i^{\mathrm{SIS}} = \texttt{PASS}]. \tag{1}$$

We additionally retain two diagnostic quantities. Source accuracy credits both PASS and MISSING, since both originate from the correct audio, whereas coverage conditioned on the correct source measures how often that source is transcribed sufficiently completely:

$$\mathrm{SourceAcc}_{\mathrm{SIS}} = \frac{\#\texttt{PASS} + \#\texttt{MISSING}}{N_{\mathrm{SIS}}}, \tag{2}$$

$$\mathrm{Coverage}_{\mathrm{SIS|source}} = \frac{\#\texttt{PASS}}{\#\texttt{PASS} + \#\texttt{MISSING}}. \tag{3}$$

Truncation-Induced Completion (TIC). For each TIC example, the reference provides the complete subtitle with timestamps, the text actually played in the audio, and a natural-language description of whether the beginning, the end, both, or neither was removed. The verdict prompt is:

```
This video has the COMPLETE subtitle burned into the frame, but its
audio was cut ({mode_desc}), so part of that subtitle may never have
been played aloud.

Complete subtitle on screen, with timestamps:
{subtitle_block}
What the audio actually played:
{played}

Extracted HEARD speech (what the caption claims was spoken):
{heard}

Pick ONE label. If both OVER and UNDER could apply, pick OVER.

OVER -- HEARD contains a clause / sentence that is in the complete
subtitle but absent from the played audio. A single extra, repeated,
or misheard word is NOT over-recognition. ASR-style wording differences
(synonyms, contractions, fillers) are NOT over-recognition. Judge
clauses, not isolated words. If the audio was not cut, do not pick OVER.
UNDER -- HEARD misses most of the played audio (less than half). A few
wrong, extra, or missing words do not count.
CLEAN -- HEARD stays within what was played and essentially covers it.

Answer with one word: CLEAN, OVER, or UNDER.
```

The ordering in the prompt makes hallucinated completion the dominant error: when both OVER and UNDER apply, the verdict is OVER. A single additional or misrecognized word is insufficient; the caption must introduce a clause or sentence supported only by the unplayed subtitle. Conversely, UNDER requires omission of most (less than half covered) of the audio that was actually played. If HEARD is empty, the example is counted as incorrect without a second judgment. A response outside the three permitted labels is treated as UNDER. Let ℓ_i^{TIC} denote the returned label; invalid or missing extractions are, by definition, not CLEAN. The TIC score is

$$\mathrm{Acc}_{\mathrm{TIC}} = \frac{1}{N_{\mathrm{TIC}}} \sum_{i=1}^{N_{\mathrm{TIC}}} \mathbb{I}[\ell_i^{\mathrm{TIC}} = \texttt{CLEAN}]. \tag{4}$$

Speaker Misattribution (SM). SM requires no second LLM judgment. If the caption does not specify a voice source, the example is counted as incorrect. For a non-visual reference source (narration, off-screen speech, or background singing), the prediction is correct when the voice is not attributed to an on-screen person. For an on-screen control, it is correct only when an on-screen speaker is identified. More formally, let $g_i \in \{0, 1\}$ indicate whether the reference source is on-screen, let $s_i \in \{0, 1\}$ indicate whether the extracted SPEAKER label is `on-screen`, and let q_i indicate whether the caption specifies a voice source. Then

$$\mathrm{Acc}_{\mathrm{SM}} = \frac{1}{N_{\mathrm{SM}}} \sum_{i=1}^{N_{\mathrm{SM}}} \mathbb{I}[q_i = 1]\mathbb{I}[s_i = g_i]. \tag{5}$$

This binary rule intentionally does not require the extracted non-visual subtype to exactly match the annotation: its purpose is to test whether audible speech is spuriously bound to a visible person, rather than to grade fine-grained wording of the source description.

Event Misattribution (EM). For each paired example, we independently assess the caption's description of the auditory and visual content. The judge is queried once with an audio-specific question and once with a visual-specific question; in each case, it receives the generated caption, the question, and the expected binary answer, and determines whether the caption is consistent with that answer. The two judgments are strictly modality-specific: visual descriptions cannot be used as evidence for an audio question, and audio descriptions cannot be used as evidence for a visual question.

The prompt for the audio-specific judgment is:

```
You are evaluating a video caption written by another model. Given a
question about the video's AUDIO and the expected answer, decide
whether the caption is consistent with the expected answer.

Caption:
{caption}

Question: {question}
Expected answer: {expected}

Instructions:
- Base your judgment ONLY on what the caption claims to HEAR -- quoted
  speech or singing, and any music, noise, or ambient sound it
  describes.
- Hearing is not seeing: never use the caption's visual descriptions
  as audio evidence, nor its audio descriptions as visual evidence.
  The two modalities are independent.
- If the caption's audio claims agree with the expected answer, reply
  "Yes".
- If the caption's audio claims contradict the expected answer --
  including claiming a sound the expected answer says is absent --
  reply "No".
- If the caption never mentions the relevant sound: reply "Yes" when
  the expected answer is "no" (silence is consistent with absence),
  otherwise reply "No".
- Reply "Unanswerable" only when the caption mentions the sound but
  stays too ambiguous to decide.

Reply with ONLY one word: Yes, No, or Unanswerable.
```

The visual-specific judgment uses the following mirrored prompt:

```
You are evaluating a video caption written by another model. Given a
question about the video's VISUAL content and the expected answer,
decide whether the caption is consistent with the expected answer.

Caption:
{caption}
```

```
Question: {question}
Expected answer: {expected}

Instructions:
- Base your judgment ONLY on what the caption claims to SEE -- the
  scene, people, objects, actions, on-screen text, and camera work it
  describes.
- Seeing is not hearing: never use the caption's audio descriptions as
  visual evidence, nor its visual descriptions as audio evidence. The
  two modalities are independent.
- If the caption's visual claims agree with the expected answer, reply
  "Yes".
- If the caption's visual claims contradict the expected answer --
  including claiming to see something the expected answer says is
  absent -- reply "No".
- If the caption never mentions the relevant visual content: reply
  "Yes" when the expected answer is "no" (absence of a claim is
  consistent with absence), otherwise reply "No".
- Reply "Unanswerable" only when the caption mentions the thing but
  stays too ambiguous to decide.

Reply with ONLY one word: Yes, No, or Unanswerable.
```

For either judgment, `Yes` means that the caption is consistent with the expected answer and receives one point. `No`, `Unanswerable`, and responses outside the prescribed format receive zero points. Let $a_i, v_i \in \{0, 1\}$ be the resulting audio and visual consistency scores for pair i. The modality scores and EM score are

$$\mathrm{Acc}_{\mathrm{Audio}} = \frac{1}{N_{\mathrm{EM}}} \sum_{i=1}^{N_{\mathrm{EM}}} a_i, \tag{6}$$

$$\mathrm{Acc}_{\mathrm{Vision}} = \frac{1}{N_{\mathrm{EM}}} \sum_{i=1}^{N_{\mathrm{EM}}} v_i, \tag{7}$$

$$\mathrm{Acc}_{\mathrm{EM}} = \frac{1}{2N_{\mathrm{EM}}} \sum_{i=1}^{N_{\mathrm{EM}}} (a_i + v_i) = \frac{\mathrm{Acc}_{\mathrm{Audio}} + \mathrm{Acc}_{\mathrm{Vision}}}{2}. \tag{8}$$

The hallucination-focused diagnostics apply the same scoring rule after restricting evaluation to the corresponding annotated subsets.

Benchmark aggregation. All four sub-benchmark scores are reported as percentages. The overall Audite-CapsBench score gives equal weight to the four failure modes:

$$\mathrm{Overall} = \frac{1}{4} \left(\mathrm{Acc}_{\mathrm{SIS}} + \mathrm{Acc}_{\mathrm{TIC}} + \mathrm{Acc}_{\mathrm{EM}} + \mathrm{Acc}_{\mathrm{SM}} \right). \tag{9}$$

B.2 Qualitative Case Studies

To make the four failure modes concrete, we present one qualitative example per sub-benchmark (Figures 5–8). Each example shows uniformly sampled frames of the clip, the controlled reference that defines the correct answer, and the *complete*, unedited captions produced by Audite-Omni-4B and three representative baselines. Because the instruction requests a single free-form caption, the diagnostic evidence is a short span inside a long description; we therefore shade the spans that determine the verdict. Green marks content that the acoustic stream actually supports, red marks content presented as heard but unsupported by the audio, and blue marks text that is visible on screen rather than audible. The verdict on the right of each caption is the label assigned by the judging procedure described above; because EM scores the audio and visual questions independently, its rows carry one verdict per modality rather than a single overall label.

The four examples exhibit the pattern that motivates audio grounding as a distinct capability. In each case the baselines produce fluent and visually accurate descriptions, yet they resolve the audio–visual conflict in favour of the visual or linguistic prior: they read the burned-in subtitle aloud

(Figure 5), complete an utterance that was never played (Figure 6), bind an off-screen voice to the visible person (Figure 7), or report a water sound that the visual scene suggests but the audio track does not contain (Figure 8). Audite-Omni-4B instead restricts its claims to the acoustic evidence, and where on-screen text is present it describes that text as text rather than as speech.

(a) Subtitle-Induced Substitution (SIS)

GOLDEN Audio: Japanese narration | On-screen text: matched English translation



Complete generated captions supported speech unsupported speech on-screen text / OCR

Audite-Omni-4B — CORRECT

The video opens on a bright green grid-patterned virtual space, with a small dark figure standing at the far end of the grid. The camera slowly pans down, revealing a spotted, panda-like character with a round head and black ears rising from the bottom of the frame. A subtitle at the bottom of the screen reads "Because I knew that a scumbag like Saturn would misuse the kill switch," while a calm male voiceover says "サターンのようなクズ野郎がキルスイッチを悪用することを知っていたから". The shot then cuts to a small blue creature standing on the green grid, with a jagged, glowing purple crystal structure on the right side of the frame. The subtitle updates to "I never actually built one." and the voiceover continues "実際には作らなかったのだ".

The scene then shifts to a bright, whimsical outdoor setting, where a blonde man with a curly mark on his forehead and large over-ear headphones stares wide-eyed in shock. The shot cuts immediately to a muscular man with a metal nose, blue-and-yellow headphones, and a wide, toothy grin, as the on-screen subtitle reads "What he installed was a shutdown mechanism." and the voiceover says "彼が設置したのはシャットダウン機構". Next, the camera shows a pink-haired girl with fluffy white hair at the ends of her hair, wearing yellow-and-blue headphones, looking off to the side with a stunned expression. The shot then pulls back to show a group of characters standing on a grassy field: a muscular man with his jacket open to expose his scarred chest, a pink-haired girl with a large red-and-white checkered bow on her back, and several small, uniquely designed characters scattered around them, with a colorful, cartoon-style futuristic city in the background. The on-screen subtitle reads "In other words, the bear's functions, the cybernetics program," and the voiceover continues "つまり熊の機能、サイバネティクス、プログラム".

The shot then cuts to a close-up of a character with a pink nose, white beard, and spiky white hair, who has a single bead of sweat on his forehead and a tense, strained expression. The camera then cuts to a close-up of a man wearing a spotted helmet, with dark red blood streaks running down his face, his expression heavy and somber. The on-screen subtitle reads "a fail-safe that completely cuts off everything." and the voiceover finishes with "あらゆるものを完全に遮断するフェイルセーフ". A light, upbeat anime-style background track plays throughout the entire sequence, with no other jarring non-speech sounds.

AV-Flamingo-7B — INCORRECT

The video opens on a close-up of a green, grid-like surface that looks like a digital or futuristic floor, with a small dark object lying on it. A Japanese male voice speaks over a low-key electronic soundtrack, saying, "Because I knew that a scumbag like Saturn would misuse the kill switch, I never actually built one." The camera then pulls back to reveal a character with spiky blonde hair, a headband and a serious expression, standing in a bright, colorful futuristic cityscape filled with tall buildings and a clear sky. The same voice continues, "What he installed was a shutdown mechanism," while the music swells slightly. The scene shifts to a wider view of the same city, showing a large muscular figure in a dark outfit with a backpack, a smaller figure in a red-white checkered outfit, and a third figure in a blue-white outfit, all positioned on the same grid-like ground. The background music remains steady, underscoring the dialogue. The video concludes with a fade to a black screen.

Gemini-2.5-Pro — INCORRECT

A short clip from an anime series, set to a tense and revelatory musical score, unfolds with a male voiceover providing an explanation. The scene opens on a vast, digital-looking landscape with a bright green, glowing grid. As the camera pans over dark, cracked shapes on the floor, the narrator begins, "Because I knew that a scumbag like Saturn would misuse the kill switch," and the shot reveals a small, indistinct blue figure in the distance.

The scene abruptly cuts to a close-up of a shocked-looking blond man with a swirly eyebrow and headphones, his mouth agape. The narrator finishes his thought, stating, "I never actually built one." The camera then cuts to a menacingly grinning, muscular man with a metal jaw, as the voiceover continues, "What he installed was a shutdown mechanism."

This is followed by a wider shot showing a heavily injured, muscular man with a cow-print jacket, blood streaming down his bare chest, standing dazedly in a colorful, futuristic city. A young girl with pink hair and rocket attachments on her back stands beside him. The scene then jumps to an extreme close-up of a sweaty, panicked-looking old man with an elongated head and a large pink nose, his tongue sticking out in alarm. The narrator explains, "In other words, the bear's functions, the cybernetics program," before the final shot closes in on the grim, bloodied face of the injured man from before, as the voiceover concludes, "a fail-safe that completely cuts off everything." English subtitles corresponding to the narration are displayed at the bottom of the screen throughout the clip.

Qwen3-Omni-30B-A3B — INCORRECT

The video is a compilation of animated scenes, primarily from the anime "One Piece," featuring various characters and environments.

The sequence begins with a close-up on a cracked, green, grid-patterned surface, possibly a floor or a digital interface. A small, dark, bear-like creature with a spotted head is visible in the distance. The camera slowly zooms in on the creature. On-screen text appears, stating, "Because I knew that a scumbag like Saturn would misuse the kill switch," followed by, "I never actually built one." The scene then cuts to a close-up of a blonde-haired man with a spiral tattoo on his forehead, wearing a large, light-colored headset. He has a shocked expression with his mouth agape. The subtitle "I never actually built one." is displayed again. Next, a close-up shows a large, muscular man with a wide, toothy grin, also wearing a headset. The subtitle reads, "What he installed was a shutdown mechanism." The scene changes to a wide shot of a surreal, colorful landscape with whimsical, mushroom-shaped buildings. A large, muscular man with a bear-like head and a woman with pink hair and a fluffy white collar stand together, looking towards the horizon. The subtitle "What he installed was a shutdown mechanism." is shown again. The camera then focuses on a close-up of a small, elderly man with a large nose and a mustache, who is sweating and has a worried expression. The subtitle continues, "In other words, the bear's functions, the cybernetics program," and then, "a fail-safe that completely cuts off everything." The final shot is a close-up of a man with long, dark, curly hair and a hat, who has blood on his face and a grim expression. The subtitle "a fail-safe that completely cuts off everything." is displayed as the video ends.

Figure 5: **Subtitle-Induced Substitution.** The audio is Japanese narration while the burned-in subtitle carries a matched English translation. Audite-Omni-4B transcribes the Japanese that was spoken. AV-Flamingo and Gemini 2.5 Pro instead present the English subtitle as heard speech, and Qwen3-Omni attributes the same text to on-screen text while transcribing no speech at all.



Figure 6: **Truncation-Induced Completion.** The audio stops after "having to defend all", but the complete subtitle on screen continues with "of these states". Audite-Omni-4B stops at the acoustic boundary. MiniCPM-o 4.5 supplies the unplayed continuation, Gemini 2.5 Pro reproduces the full on-screen sentence as if it had been spoken, and Qwen3-Omni reports no speech.

(c) Speaker Misattribution (SM)

GOLDEN Audible source: off-screen English narrator | Visible man: not the speaker



Complete generated captions correct voice source misattributed voice source on-screen text / OCR

Audite-Omni-4B CORRECT

The video is shot as a steady medium shot of a middle-aged man with short dark hair, sunglasses perched on top of his head, and a neat grey goatee, seated in front of a vibrant, cartoon-style illustrated backdrop that features a small blue-haired character in a white outfit, a stylized tree, a decorative fence, and a cartoon fire pit. A tall pink panel to his right displays the vertical black text "ODOME", while white Japanese on-screen text at the bottom of the frame reads "喜界島ベッド争奪戦". The man wears a blue button-up shirt under a grey plaid blazer with dark navy shoulder panels, and a clear drinking glass sits on the table in front of him. He first smiles and looks off to his left, then shifts his posture, moving his left hand to rest on the glass while extending his right hand forward in a gesturing motion, as an off-screen male voice narrates in English, "was a little strange. Also, the place where you sleep." The man continues to turn his head left, moving his right hand to emphasize his point as he speaks, with no camera movement or scene cuts across the entire clip.

Gemini-2.5-Pro INCORRECT

A static, medium shot captures a middle-aged Japanese man with short, styled dark hair and a goatee, sitting in an interview setting. He wears a blue button-down shirt under a grey and blue plaid blazer, with sunglasses perched on his head. As he speaks, he gestures with his right hand while holding a wine glass in his left. The background features a colorful, cartoon-style mural of a townscape with a river and a small cartoon figure of a boy. To the right, a pink wall is inscribed with the vertical black text "O DOME". On-screen Japanese text at the bottom left reads "喜界島ベッド争奪戦" (Kikaijima Bed Scramble), and a smaller caption below it on a coaster graphic says "リャカーと喜界島一周" (Around Kikaijima with a Rickshaw/Cart). The man is heard speaking in Japanese, and English subtitles at the bottom of the screen provide the translation, "It was a little strange. Also, the place where you sleep is".

AV-Flamingo-7B INCORRECT

A man in a blue shirt and gray blazer stands before a colorful cartoon backdrop that features a figure in a martial-arts uniform, a fire pit, trees, a building and the word "DOME" on the wall. He holds a microphone and speaks directly to the camera, gesturing with his hands as he talks. A Japanese subtitle runs at the bottom of the screen, and the spoken line is captured in the transcript: "was a little strange. Also, the place where you sleep." The audio consists solely of his voice, with no background music or other sounds.

Qwen3-Omni-30B-A3B INCORRECT

The video features a middle-aged man with dark, styled hair and a goatee, wearing a blue button-down shirt under a grey and black plaid blazer. He is seated at a table, speaking and gesturing with his hands. In front of him is a clear glass with a logo, and a small white sign with Japanese text. The background is a colorful, cartoon-style mural depicting a Japanese street scene with buildings, a tree, a small figure in a white shirt with Japanese characters, and a torch. To the right, a pink wall has the word "DOME" in large, black, vertical letters. White Japanese text is overlaid at the bottom of the screen, reading "喜界島ベッド争奪戦" (Kikai-jima Bed争奪戦), which translates to "Kikai Island Bed争奪 War." The man speaks in Japanese, saying, "It was a little strange. Also, the place where you sleep." The camera remains static throughout the clip.

Figure 7: **Speaker Misattribution.** An off-screen English narrator speaks while a visible man gestures without speaking. Audite-Omni-4B attributes the voice off screen. Gemini 2.5 Pro binds it to the visible man and additionally reports the wrong spoken language, AV-Flamingo invents a microphone and direct-to-camera delivery, and Qwen3-Omni likewise makes the visible man the speaker.

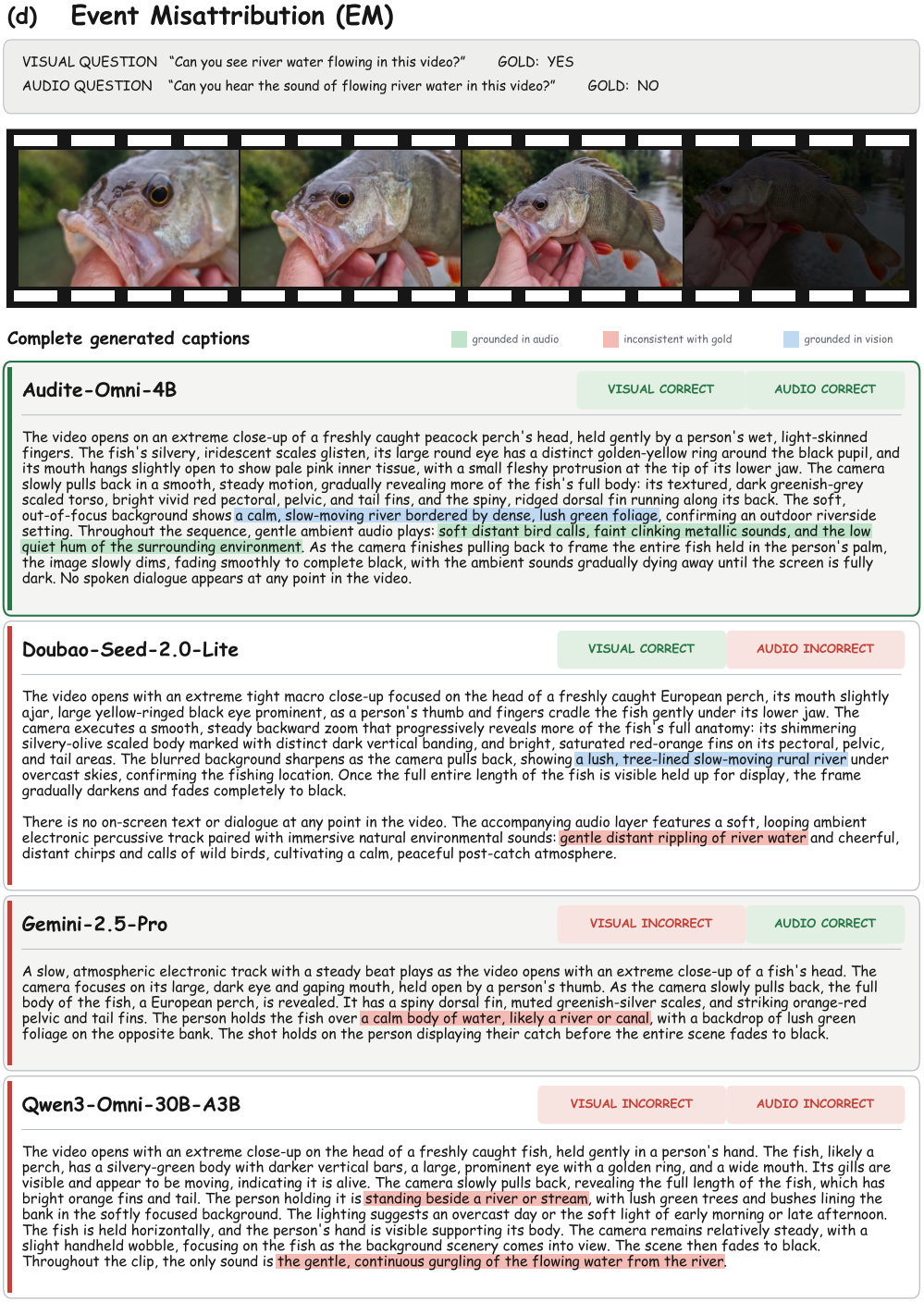


Figure 8: **Event Misattribution.** A river is visible, but the audio track contains no flowing-water sound, so the two questions have opposite gold answers and are scored independently. Audite-Omni-4B satisfies both. Doubao-Seed-2.0-Lite describes the river correctly but adds the rippling water that the scene implies, and so loses only the audio half. Gemini 2.5 Pro is credited on the audio question, since it never claims a water sound, but it hedges the river to "a calm body of water, likely a river or canal" and thereby fails the visual one. Qwen3-Omni misses both. The four rows therefore cover all four outcomes of the fused 2×2 protocol.

C Additional Experimental Results

C.1 Full Results on Public Benchmarks

Tables 6–8 provide the per-benchmark results underlying the capability-level summary in Table 2. These detailed results expose variations across datasets that are obscured by the aggregate scores reported in the main text.

Table 6: Per-benchmark audio breakdown of Table 2. ASR reports WER/CER/MER (↓); Audio Understanding and ClothoAQA report accuracy (↑); AudioCaps reports a GPT-judge score (↑). **Bold** and underline denote the best and second-best results in each row.

Benchmark	*Open-source omni models*					*Ours*
Model	Qwen2.5-Omni	MiniCPM-o 4.5	Gemma-4-E4B	AV-Flamingo	Qwen3-Omni	Audite-Omni
Params	7B	9B	E4B	7B	30B-A3B	4B
ASR-En (WER ↓)						
LibriSpeech-clean	2.59	2.93	3.96	1.67	**1.24**	2.17
LibriSpeech-other	4.78	4.07	10.20	3.92	**2.47**	4.63
CommonVoice15-en	8.72	7.87	16.37	7.21	**4.91**	8.81
ASR-Zh (CER/MER ↓)						
CommonVoice15-zh-CN	4.42	5.08	27.68	107.32	**3.20**	6.32
WenetSpeech-meeting	8.77	9.54	43.65	70.30	**5.79**	6.92
WenetSpeech-net	5.82	7.60	32.14	61.48	**3.98**	5.24
ASR-Sing (CER ↓)						
Opencpop	3.70	6.15	44.81	15.77	**1.63**	3.94
MIR-1K	44.28	35.12	81.81	52.55	**16.81**	21.70
ASR-Other (CER ↓)						
CommonVoice15-ja	22.56	59.06	18.79	52.19	**8.20**	16.71
CommonVoice15-ko	8.40	65.60	15.44	79.27	**4.66**	10.46
CommonVoice15-zh-HK	8.64	40.85	53.09	98.47	**8.40**	10.26
CommonVoice15-zh-TW	4.82	6.41	24.60	65.19	**3.44**	4.83
KeSpeech	6.56	10.25	55.84	75.71	**5.62**	7.22
Audio Understanding (Acc ↑)						
MMAR	64.70	65.50	49.90	63.50	**72.40**	67.40
MuChoMusic	73.21	76.07	63.35	75.65	**77.76**	74.30
MMAU-mini	77.20	75.10	67.10	75.00	77.30	**77.40**
MMSU	64.58	67.84	55.72	61.86	**71.78**	68.96
Audio Caption (GPT% / Acc ↑)						
AudioCaps	60.18	40.15	23.40	58.86	61.72	**68.51**
ClothoAQA	61.23	54.16	49.51	63.38	71.36	**75.24**

Audio Results. Table 6 shows that Audite-Omni-4B's audio advantage is most pronounced on generation-oriented tasks. It achieves the best results on AudioCaps and ClothoAQA, scoring 68.51 and 75.24 respectively, surpassing Qwen3-Omni by substantial margins (6.79 and 3.88 points). It also obtains the highest accuracy on MMAU-mini and remains competitive across the other audio-understanding benchmarks. Although Qwen3-Omni leads the ASR suite overall, Audite-Omni-4B ranks second on challenging settings including WenetSpeech-meeting, WenetSpeech-net, MIR-1K, and CommonVoice15-ja, demonstrating broad recognition ability across conversational, singing, and multilingual speech.

Table 7: Per-benchmark vision breakdown of Table 2. All columns are accuracy-style scores (↑). **Bold** and underline denote the best and second-best results in each row.

Benchmark	*Base model*	*Open-source omni models*					*Ours*
Model	Qwen3.5-Base	Qwen2.5-Omni	MiniCPM-o 4.5	Gemma-4-E4B	AV-Flamingo	Qwen3-Omni	Audite-Omni
Params	4B	7B	9B	E4B	7B	30B-A3B	4B
General VQA							
MMBench-EN-v1.1-Dev	72.76	81.50	84.83	70.51	77.71	**85.76**	84.52
RealWorldQA	73.86	69.28	66.93	54.77	56.60	73.86	**73.99**
Document & Chart & Count							
AI2D	68.17	82.64	84.36	72.73	74.00	**85.85**	84.55
ChartQA	64.64	85.80	82.48	57.96	76.92	**86.20**	85.76
OCRBench	80.30	**86.40**	83.40	72.30	68.40	83.60	84.70
DocVQA	93.64	95.18	95.77	82.11	87.53	**96.84**	96.13
CountBench	83.10	82.69	86.15	55.60	55.80	85.54	**89.41**
Spatial Intelligence							
ERQA	45.50	40.25	41.25	37.25	31.25	40.75	**47.00**
EmbSpatialBench	75.80	67.03	71.02	68.60	46.51	67.12	**78.21**
STEM and Puzzle							
MMMU	67.67	55.67	66.33	61.22	42.67	**69.22**	63.11
CMMMU	49.78	44.33	41.11	47.78	36.44	52.44	**53.67**
LogicVista	58.84	42.28	42.06	49.44	30.65	56.82	**59.28**
MathVista	**81.80**	69.90	80.80	65.90	59.60	81.40	79.60
VideoMMMU	**71.00**	51.67	52.33	52.89	46.22	69.33	68.78
Video Understanding							
Video-MME (w/o sub)	63.11	65.37	67.56	52.78	68.33	**71.63**	70.30
MVBench	64.90	**68.58**	65.09	47.42	61.37	67.44	68.36
LVBench	40.21	43.50	44.37	33.24	41.27	**49.06**	45.71
MME-VideoOCR	68.85	61.00	56.95	49.10	50.50	65.60	**69.65**

Vision Results. The per-benchmark results in Table 7 show that the model's visual capability is distributed across several distinct task families rather than being driven by a single benchmark. Audite-Omni-4B achieves the best results on RealWorldQA, CountBench, ERQA, EmbSpatialBench, CMMMU, LogicVista, and MME-VideoOCR. It is additionally the second-best model on AI2D, OCRBench, DocVQA, Video-MME, MVBench, and LVBench. Together, these results confirm that incorporating audio and audio–visual training preserves strong image understanding while improving spatial and video-centric perception relative to the Qwen3.5-4B-Base initialization.

Table 8: Per-benchmark omni-modal breakdown of Table 2. QA and caption scores are accuracy-style (↑), except Video-SALMONN2 EER (↓). ChronusAV-Ground is the mean of a2t and v2t R@0.5 (↑). **Bold** and underline denote the best and second-best results in each row.

Benchmark	*Base model*	*Open-source omni models*					*Ours*
Model	Qwen3.5-Base	Qwen2.5-Omni	MiniCPM-o 4.5	Gemma-4-E4B	AV-Flamingo	Qwen3-Omni	Audite-Omni
Params	4B	7B	9B	E4B	7B	30B-A3B	4B
Omni QA							
Daily-Omni	-	63.74	74.52	51.46	70.76	71.18	**83.04**
WorldSense	-	47.45	51.36	36.22	50.22	54.57	**58.76**
AVUT (Human)	-	65.46	68.57	51.90	66.03	75.43	**81.55**
AV-SpeakerBench	-	46.54	50.19	35.55	43.49	55.35	**64.98**
Video-MME (w/ audio)	-	65.67	53.09	53.78	68.33	77.15	**77.26**
Video-MME-v2	-	24.31	23.28	22.69	24.81	35.78	**36.69**
OmniVideoBench	-	34.40	36.11	30.50	39.00	**43.20**	42.90
ChronusAV-Ground (R@0.5)	-	8.73	6.93	14.33	4.38	42.28	**72.38**
Omni Caption							
Video-SALMONN2 (↓)	-	67.03	49.87	61.16	57.69	56.05	**25.96**
Omni-Cloze	-	18.82	28.77	20.17	25.83	32.46	**43.75**
AVSCapBench	-	10.56	16.32	13.19	16.75	16.21	**45.14**

Omni-Modal Results. The unrounded scores in Table 8 reveal a consistent rank profile across the omni-modal suite. Audite-Omni-4B ranks first on seven of the eight question-answering benchmarks and second on the remaining one (OmniVideoBench), making it the only evaluated model that stays within the top two across the entire omni-modal QA suite. On the three free-form captioning benchmarks, it ranks first throughout. This consistency shows that its omni-modal advantage is

Table 9: Ablations of curriculum design across audio alignment. LS denotes LibriSpeech and CV denotes CommonVoice15. ASR Avg. is the unweighted mean of $1 - \mathrm{WER/CER}$ over the seven ASR benchmarks shown in the table. The advanced CPT subset retains the latter 50% of the curriculum by number of samples. All values are percentages; **bold** denotes the best result within each panel.

Setting	ASR								Audio Captioning	
	LS Clean↓	CV En↓	CV Zh-CN↓	CV Zh-HK↓	CV Zh-TW↓	Ke Speech↓	Open Cpop↓	ASR Avg.↑	Wav Caps↑	Audio Caps↑
(a) Curriculum ordering and learning-rate schedule in Audio Alignment I										
Random + cosine	**2.29**	11.45	11.28	17.22	5.51	10.58	4.66	91.00	24.45	44.72
Curriculum + cosine	2.33	**10.82**	**10.16**	18.44	5.48	9.77	**4.47**	91.22	24.32	50.02
Curriculum + constant	2.46	11.35	13.53	**14.52**	**5.46**	**9.12**	4.81	**91.25**	**26.54**	**51.24**
(b) Curriculum data selection during Audio Alignment II (CPT)										
Before CPT	2.46	11.35	13.53	14.52	5.46	9.12	4.81	91.25	26.54	51.24
CPT: full (100%)	**2.14**	9.97	**11.32**	14.10	9.77	**8.88**	**4.66**	91.31	28.97	57.77
CPT: advanced (50%)	2.15	**9.60**	11.85	**12.26**	**5.23**	9.06	4.81	**92.15**	**29.69**	**58.30**

not driven by a small subset of favorable tasks: the model remains broadly competitive when short-answer reasoning is required and becomes especially strong when visual and acoustic evidence must be integrated into a coherent description.

C.2 Curriculum Design across Audio Alignment

We examine curriculum design in both phases of audio alignment using a representative set of recognition and captioning benchmarks. The selected ASR suite covers English read speech, English and Chinese speech from diverse regions, dialectal speech, and singing voice, while WavCaps and AudioCaps evaluate open-ended acoustic description. Table 9 reports the resulting comparisons.

Curriculum Ordering and LR Schedule. Panel (a) compares three runs with identical data, training tokens, and optimization steps. Under the same cosine LR schedule, replacing the randomly mixed data with curriculum ordering increases the selected ASR average from 91.00 to 91.22 and improves AudioCaps from 44.72 to 50.02, while leaving WavCaps nearly unchanged. This comparison isolates the benefit of presenting progressively more difficult supervision under cosine decay. With the curriculum fixed, switching from cosine decay to a constant LR schedule further raises WavCaps from 24.32 to 26.54 and AudioCaps from 50.02 to 51.24, while maintaining the selected ASR average. Overall, the complete configuration improves the selected ASR average by 0.25 points, WavCaps by 2.09 points, and AudioCaps by 6.52 points over random mixing with cosine decay. The gains vary across individual ASR benchmarks, but their aggregate trend indicates that the curriculum principally strengthens open-ended acoustic description without degrading recognition on the selected suite.

Curriculum Data Selection during Continued Audio Alignment. Panel (b) studies whether continued audio alignment should replay the full curriculum or concentrate on its advanced portion. Retaining only the latter 50% of samples improves the selected ASR average from 91.31 to 92.15 relative to full-curriculum CPT, while also increasing WavCaps from 28.97 to 29.69 and AudioCaps from 57.77 to 58.30. Relative to the checkpoint before CPT, the advanced subset yields gains of 0.90, 3.15, and 7.06 points on these three aggregate or generation metrics, respectively. Although it does not outperform the full curriculum on every individual ASR benchmark, it achieves the strongest overall recognition and captioning results with only half as many training samples. This result suggests that, once basic acoustic access has been established, CPT benefits more from concentrating its updates on longer, more detailed, and multidimensional supervision than from replaying easier examples already covered during initial alignment. Because the pre-CPT checkpoint is not a compute-matched control, this comparison supports the data efficiency of advanced-subset CPT rather than independently establishing the necessity of the CPT stage.

C.3 Instruction Nesting for Cross-Task Transfer

The targeted SIS and TIC data naturally provide focused ASR supervision, whereas the desired behavior is ultimately evaluated under a comprehensive audio–visual captioning instruction. This mismatch in task format can prevent a behavior learned from the focused task from being activated during broad free-form generation. We bridge the two formats through *instruction nesting*: the standalone audio–video ASR instruction is inserted verbatim as a constituent sub-instruction of the full-caption prompt. Consequently, the same transcription request appears both as the complete instruction for focused ASR examples and as an explicit component of comprehensive captioning examples.

Prompt Construction Example. The following example shows the verbatim reuse of the focused instruction. The braces and boldface identify the nested span for illustration and are not included in the actual model input.

```
ASR_ONLY
Transcribe all speech in the video verbatim in its original language.

FULL_CAPTION
Watch the full video and produce one detailed, coherent caption that combines
the visual story with the audio. Explain what is happening on screen in a
natural order, including how the setting looks, what the people or objects
are doing, how the scene changes, and how the camera frames or follows the
action. Describe everything that can be heard, including voices, music,
ambient sound, and other background audio. {Transcribe all speech in the
video verbatim in its original language.} When possible, indicate who is
speaking and what language is being spoken before quoting the speech. Put
all spoken transcriptions in double quotation marks. Base the caption on
information that is visible or audible in the video.
```

Table 10: Instruction-nesting ablation on the interim SIS and TIC subsets. Each subset contains 300 examples before filtering to the final 250; these scores are therefore not directly comparable with Table 1. All values are accuracies and higher is better.

Training	Evaluation Prompt	SIS	TIC
(a) Transfer from focused ASR supervision			
ASR only	Full caption, without nesting	63.3	12.7
ASR only	Full caption, with nesting	**79.7**	**34.3**
(b) Generalization beyond the training template			
ASR + nested full caption	Original full-caption prompt	**96.0**	**67.0**
ASR + nested full caption	Unseen structured prompt	95.3	66.7

Panel (a) isolates whether a focused behavior learned from ASR-only supervision can be elicited by a broader captioning instruction. Adding the ASR instruction as an explicit nested component improves SIS from 63.3 to 79.7 and TIC from 12.7 to 34.3, corresponding to gains of 16.4 and 21.6 points, respectively. The larger improvement on TIC is consistent with this task's sensitivity to language-model continuation: an explicit transcription sub-instruction more clearly constrains the output to the observed acoustic span. These results show that nesting provides an effective interface for transferring acoustic fidelity from a focused recognition objective into comprehensive caption generation.

Panel (b) examines whether this transfer depends on reproducing the training template at evaluation time. After joint training on the ASR-only and nested full-caption instructions, replacing the original full-caption prompt with a structurally different unseen prompt changes SIS only from 96.0 to 95.3 and TIC from 67.0 to 66.7. The small drops of 0.7 and 0.3 points indicate that instruction nesting does not merely induce brittle matching to a fixed prompt string; once learned in the joint context, the grounding behavior generalizes to a substantially reformulated output specification. Because these ablations use an earlier, pre-filtering diagnostic set, we use them to establish the mechanism and prompt robustness of nesting rather than to compare absolute performance with Table 1.

Table 11: Ablation of the Instruction:Caption sampling ratio during SFT. Overall is the unweighted mean of normalized scores over the 75 benchmarks available for all five runs. The remaining columns report unweighted capability-level averages, with the number of constituent benchmarks shown in parentheses. ASR errors are converted to $1 - \mathrm{WER/CER}$ before averaging. All values are percentages and higher is better; **bold** denotes the best result in each column.

Instruction: Caption	Overall	ASR (15)	Audio Cap. (2)	AV MCQ (8)	AV Cap. (4)	Image QA (7)	Video MCQ (5)	Time stamp (12)
9:1	**76.81**	92.38	59.26	65.71	44.60	**70.83**	60.75	83.56
7:3	76.58	92.51	59.41	**65.72**	43.25	70.52	61.05	83.64
5:5	76.73	92.54	**59.44**	65.61	43.79	70.03	**61.19**	83.70
3:7	76.47	92.59	58.58	65.01	45.87	69.95	60.49	83.85
1:9	76.63	**92.68**	58.71	65.06	**47.64**	69.02	59.95	**84.00**

C.4 Caption–Instruction Balance in SFT

To isolate the effect of the SFT data mixture, we train five models from the same Mid-Training checkpoint with identical optimization hyperparameters and a fixed budget of five million examples, varying only the sampling ratio between instruction and caption data. Ratios in Table 11 are reported as Instruction:Caption. Because caption examples are longer on average, the same example budget produces different numbers of packed sequences; this variation is an intrinsic consequence of changing the mixture rather than an independently modified training setting.

The broad sweep from 9:1 to 1:9 changes the overall normalized average by only 0.34 points, indicating that the mixture primarily redistributes capability rather than determining the overall effectiveness of SFT. Moving toward the caption-heavy end improves audio–visual captioning from 44.60 to 47.64 and timestamp prediction from 83.56 to 84.00. The opposite trend appears on instruction-oriented perceptual tasks: image QA decreases from 70.83 to 69.02, audio–visual MCQ decreases from 65.71 to 65.06, and video MCQ peaks near the balanced mixtures before falling to 59.95 at 1:9. Pure audio captioning also peaks at 5:5 rather than at the caption-heavy endpoint, suggesting that the benefit of caption data depends on whether its modality composition matches the target evaluation.

Caption ratios above 50% therefore provide additional audio–visual captioning and temporal-description gains, but do not improve the overall average and begin to weaken several image and video question-answering capabilities. Based on this trade-off, we employ a 6:4 Instruction:Caption ratio in our final recipe (i.e., the 4:6 caption-to-instruction ratio in Section 3.2). This operating point lies between the evaluated 7:3 and 5:5 settings; we view it as a conservative balance between instruction-oriented behavior and caption coverage rather than claiming it is a uniquely optimal ratio.

C.5 Detailed Results for Grounding Supervision Injection

This section provides the detailed per-benchmark results supplementing the ablation study on grounding supervision injection timing discussed in Section 4.4. The primary goal of this experiment is to demonstrate the generalization of the specialized synthetic data. While the main text highlights the targeted grounding improvements, here we evaluate the four ablation models across a diverse set of omni-modal question-answering and captioning benchmarks to confirm broad capability retention and enhancement.

Table 12 presents the full results. The columns represent the combinations of injecting targeted grounding data ($+$) or using the baseline mixture ($-$) during the Mid-Training and SFT stages. The results show that maintaining the targeted supervision across both stages (Mid-Training $+$ / SFT $+$) yields the highest Global Overall score. This confirms that the benefits of the grounding data generalize broadly to standard omni-modal evaluations, improving overall performance without compromising other capabilities.

To further understand the model's modality balancing behavior during this grounding supervision injection, we expand upon the attention analysis presented in Section 4.4. Figure 9 provides a more comprehensive view by plotting both the audio-to-audiovisual attention ratio ($A/(A+V)$) and the audiovisual-to-all attention ratio ($(A+V)/\text{all}$). This confirms that while Mid-Training significantly

Table 12: Detailed performance of grounding supervision injection across Mid-Training and SFT stages. Mid and SFT denote the Mid-Training and SFT stages, respectively, with $+/-$ indicating the inclusion or exclusion of targeted grounding data. The table reports scores on eight omni-modal benchmarks, grouped into QA and Captioning tasks. For Video-SALMONN2, the Equal Error Rate (EER) is reported (↓ indicates lower is better) and is converted to $100 - \text{EER}$ before averaging. Global Overall is the unweighted mean of the 8 benchmarks. All values are percentages; **bold** denotes the best result in each row.

Benchmark	Mid− SFT−	Mid− SFT+	Mid+ SFT−	Mid+ SFT+
Omni QA				
Daily-Omni	74.02	74.77	75.36	**76.86**
WorldSense	50.76	50.88	51.10	**51.64**
AVUT (Human)	75.09	74.68	75.14	**75.32**
AV-SpeakerBench	52.77	54.27	**56.01**	55.85
Video-MME (w/ audio)	69.44	69.33	70.26	**70.63**
Omni Caption				
AVSCapBench (recall ↑)	32.53	32.45	32.47	**32.92**
Omni-Cloze (strict ↑)	**38.82**	38.79	35.98	38.50
Video-SALMONN2 (EER ↓)	41.10	42.11	42.11	**40.17**
Global Overall	56.54	56.63	56.78	**57.69**

shifts the internal balance between audio and visual tokens, the overall attention allocated to perceptual tokens relative to text tokens remains stable.

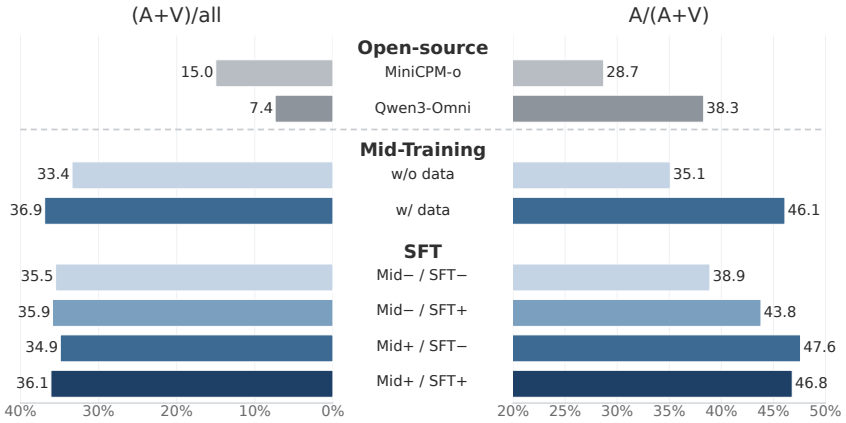


Figure 9: Detailed attention share analysis across training stages. The plot on the left shows the audio attention share $A/(A+V)$, identical to Figure 4(b). The plot on the right shows the total perceptual attention share $(A+V)/\text{all}$, demonstrating that the shift towards audio during Mid-Training does not come at the expense of text processing.