XMODBENCH: BENCHMARKING CROSS-MODAL CAPABILITIES AND CONSISTENCY IN OMNI-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (OLLMs) aim to unify audio, vision, and text understanding within a single framework. While existing benchmarks primarily evaluate general cross-modal question-answering ability, it remains unclear whether OLLMs achieve modality-invariant reasoning or exhibit modality-specific biases. We introduce **XModBench**, a large-scale tri-modal benchmark explicitly designed to measure cross-modal consistency. XModBench comprises **60,828** multiple-choice questions spanning **five task families** and systematically covers all **six modality compositions** in question–answer pairs, enabling fine-grained diagnosis of an OLLM's modality-invariant reasoning, modality disparity, and directional imbalance. Experiments show that even the strongest model, Gemini 2.5 Pro, (i) struggles with spatial and temporal reasoning, achieving less than 60% accuracy, (ii) reveals persistent modality disparities, with performance dropping substantially when the same semantic content is conveyed through audio rather than text, and (iii) shows systematic directional imbalance, exhibiting lower consistency when vision serves as context compared to text. These findings indicate that current OLLMs remain far from truly modality-invariant reasoning, and position **XModBench** as a fundamental diagnostic tool for evaluating and improving cross-modal competence.

1 INTRODUCTION

Omni-modal large language models (OLLMs) integrate text, vision, and audio into a unified reasoning framework (Comanici et al., 2025; Xu et al., 2025; Xing et al., 2025; Su et al., 2023; Fu et al., 2025b; Cheng et al., 2024; Zhong et al., 2025). However, despite impressive advancements and expanded modality coverage, a key question remains: do these models reason in a truly modality-invariant manner, or do they still exhibit systematic biases tied to specific input modalities? For humans, cross-modal integration is typically seamless, yet it remains unclear whether OLLMs demonstrate comparable consistency. When the same semantic content is presented in different forms—spoken audio, written text, or visual images—do models still converge on the same correct answer? We refer to this property as *cross-modal consistency*: the ability to maintain stable predictions regardless of input modality, thereby demonstrating reasoning over shared semantic representations rather than relying on modality-specific cues. Although directly diagnosing whether current OLLMs achieve this goal is non-trivial, we can evaluate them through carefully designed benchmarks that expose inconsistencies. For instance, by posing semantically identical questions under different modality settings, we can test whether predictions diverge across modalities — an indicator of reliance on surface-level patterns rather than genuine modality-invariant reasoning.

Recent benchmarks have taken promising steps toward evaluating OLLMs, particularly through audio-visual tasks that reveal baseline cross-modality performance. Datasets such as Music AVQA (Li et al., 2022), AV-Reasoner (Lu et al., 2025), and Pano-AVQA (Yun et al., 2021) primarily probe fine-grained audio–visual reasoning, while broader efforts like AVQA (Yang et al., 2022), WorldSense (Hong et al., 2025), AV-Odyssey Bench (Gong et al., 2024), and OmniBench (Li et al., 2024b) expand to general multimodal understanding across diverse contexts. However, these benchmarks largely overlook whether models remain consistent across modalities. While other works (Park et al., 2025; Zhang et al., 2024) attempt to assess modality consistency, they are restricted to the vision–text setting within vision–language models.

Table 1: Comparison of multimodal question-answering (QA) benchmarks by modality coverage, task domains, and modality consistency.

Benchmark	#Q	Context Modality			Candidate Modality			Task Domain					Mod. Consist.
		Text	Vision	Audio	Text	Vision	Audio	Percep.	Spatial	Temporal	Ling.	Ext. Know.	
MME Bench (Fu et al., 2024a)	2,194	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗
MMBench (Liu et al., 2024)	3,217	✗	✓	✗	✓	✗	✗	✓	✓	✗	✓	✓	✗
OcrBench v2 (Fu et al., 2024b)	10,000	✗	✓	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗
SEED-Bench-2 (Li et al., 2024a)	24,371	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✗
AudioBench Wang et al. (2024)	24,371	✗	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
Audiopedia (Li et al., 2022)	45,867	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗
MMAU (Sakshi et al., 2024)	10,000	✗	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
AVQA (Yang et al., 2022)	57,335	✗	✓	✓	✓	✗	✗	✓	✓	✓	✗	✗	✗
Pano-AVQA (Yun et al., 2021)	51,700	✗	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
Music-AVQA (Li et al., 2022)	45,867	✗	✓	✓	✓	✗	✗	✓	✓	✓	✗	✗	✗
SAVE Bench (Sun et al., 2024)	4,350	✗	✓	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
Video-MME (Fu et al., 2025a)	2,700	✗	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗
WorldSense (Hong et al., 2025)	3,172	✗	✓	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗
AV-Reasoner (Lu et al., 2025)	1,027	✗	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗
AV-Odyssey Bench (Gong et al., 2024)	1,142	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
OmniBench (Li et al., 2024b)	4,555	✗	✓	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
XModBench (Ours)	60,828	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

To address this gap, we introduce **XModBench**, a benchmark specifically designed to evaluate cross-modal consistency in omni-modal large language models. We formulate all questions in a multiple-choice format, where each question naturally contains two components: (i) a *context* describing an object or event, and (ii) a set of *candidates* from which the model must select the correct one. Unlike prior benchmarks that typically fix either the context or the choices to a single modality (Yang et al., 2022; Li et al., 2024b), XModBench systematically covers all six cross-modal directions among audio, vision, and text (see Tab. 1). To ensure broad coverage and rigorous evaluation, XMODBENCH spans five domains—perception, spatial reasoning, temporal reasoning, linguistic understanding, and external knowledge. We curate data across these domains through re-annotation, synthetic construction, and targeted web collection, ensuring both diversity and balance across modalities. The resulting benchmark comprises **60,828** multiple-choice question–answer pairs (10,138 unique instances), each instantiated in six modality configurations that preserve identical semantics across audio, visual, and textual forms. This enables both large-scale evaluation and fine-grained diagnosis of cross-modal consistency. An overview of the benchmark design is illustrated in Fig. 1.

We systematically evaluate models on XMODBENCH, going beyond overall accuracy to provide fine-grained diagnosis of cross-modal reasoning. Specifically, we analyze three complementary dimensions: (1) **Task competence**—by averaging over all six modality directions, we assess model performance across perception, spatial, temporal, linguistic, and knowledge tasks, yielding task-centric comparisons of multimodal competence; (2) **Modality disparity**—we measure consistency when the same question is posed in different modalities, where high variability signals reliance on modality-specific cues rather than shared semantic representations; and (3) **Directional imbalance**—we compare accuracy when context and candidate modalities are swapped, revealing asymmetries in cross-modal grounding.

Our experiments show that current OLLMs fall short along all three axes. They perform strongly on perception and linguistic tasks (best models reach around 70%), but degrade by 15–25 points on spatial and temporal reasoning. Performance also drops sharply whenever audio is involved, underscoring that auditory representations remain the weakest link. Finally, accuracy is consistently higher when text serves as the candidate modality, highlighting incomplete bidirectional alignment across modalities. Together, these findings demonstrate that today's OLLMs remain far from achieving modality-invariant reasoning, underscoring the diagnostic value of XMODBENCH.

In summary, XMODBENCH makes the following key contributions:

1. **Cross-modal consistency benchmark.** We present XMODBENCH, the first tri-modal multiple-choice question-answering benchmark explicitly designed to evaluate cross-modal consistency, covering all six modality mappings among audio, vision, and text.
2. **Comprehensive coverage.** The benchmark spans five task families with 17 subtasks and 60,828 question–answer pairs, ensuring broad domain coverage and fine-grained diagnostics, while its balanced design enables fair assessment of modality-invariant reasoning.
3. **Diagnostic metrics.** We introduce *modality disparity* and *directional imbalance* to directly measure robustness and bidirectional alignment across modalities. Our experiments reveal

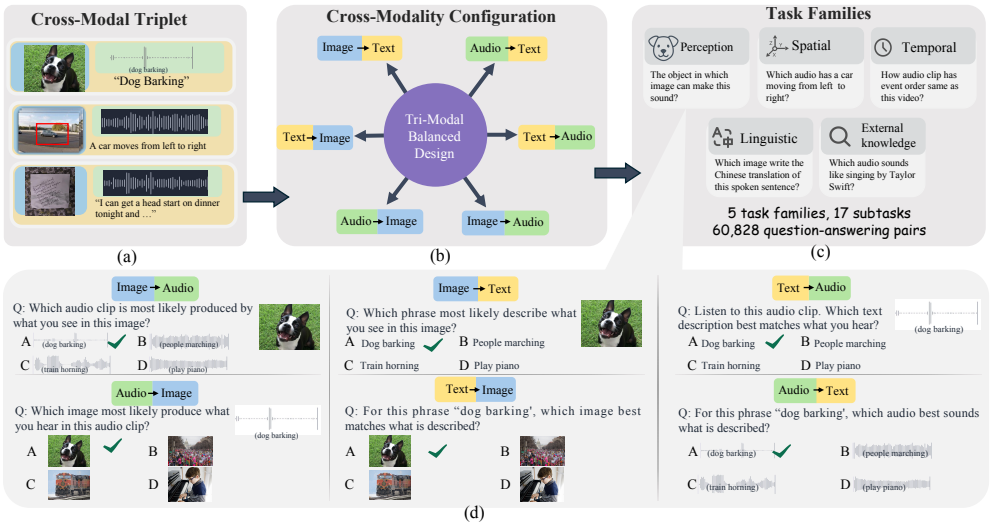

Figure 1: Overview of **XModBench**. (a) Instances are built from aligned text–image–audio triplets; (b) instantiated into six modality configurations by permuting context and candidate modalities; (c) spanning five domains with 17 subtasks and 60,828 question–answer pairs; and (d) illustrated with example multiple-choice questions under balanced modality settings.

systematic weaknesses in current OLLMs, providing actionable insights for developing more modality-invariant architectures and training strategies.

2 Related Work

Multimodal Question Answering (QA) Benchmarks. A number of benchmarks have been developed to evaluate multimodal large language models (MLLMs). Grouped by modality composition, Yin et al. (2024), Liu et al. (2024), and Li et al. (2024a) focus on the vision–text setting (covering both images and videos). For audio–text evaluation, representative efforts include Wang et al. (2024) and Sakshi et al. (2024). When combining audio and vision with text, a variety of benchmarks have emerged, including Yang et al. (2022), Li et al. (2022), Yun et al. (2021), Sun et al. (2024), Hong et al. (2025), Lu et al. (2025), Gong et al. (2024), Li et al. (2024b), and Zhou et al. (2025). Other recent works, such as Yang et al. (2025), further extend evaluation to diverse multimodal combinations. Despite their breadth, these benchmarks primarily emphasize coverage across tasks and modalities, while less attention has been paid to evaluating *cross-modal consistency*—whether models produce stable answers when the same semantic content is expressed in different modality forms. Our work fills this gap by explicitly designing a benchmark centered on modality-invariant reasoning.

Cross-Modality Consistency. Recent work has begun to investigate whether multimodal models behave consistently across modalities. Park et al. (2025) introduced the Modality Importance Score to quantify modality bias, which measures how much each modality contributes to answering questions in VideoQA. Zhang et al. (2024) further proposed the notion of cross-modal consistency between text and image, defining a consistent model as one that applies the same internal reasoning to semantically identical inputs across modalities, thereby yielding consistent outcomes. In contrast, other studies, such as Sung-Bin et al. (2024) and Choong et al. (2024), report instances of inconsistent audio-video reasoning, where models hallucinate non-existent sounds or visual signals, thereby exposing modality bias and cross-modal inconsistency. While these efforts provide pioneering insights into cross-modal consistency, they are typically confined to specific modality pairs. Our work not only expands the scope to cover a broader range of modality combinations for state-of-the-art OLLMs, but also conducts a deeper analysis of their cross-modality reasoning behavior on a comprehensive task suite.

3 XModBench: Comprehensive Cross-Modal Balanced Benchmark

We introduce **XModBench**, a comprehensive multiple-choice question-answering (QA) benchmark designed to evaluate the cross-modal capabilities and consistency of OLLMs across audio, vision,

and text. A key feature of **XModBench** is its modality-balanced design, which creates six cross-modal variants of semantically identical questions to enable a controlled and fair evaluation of cross-modal capabilities and consistency (Sec. 3.1). The benchmark offers extensive domain coverage through five task families and seventeen subtasks (Sec. 3.2), all built upon meticulously curated, high-quality, and diverse tri-modal data (Sec. 3.3).

3.1 Benchmark Design

The central objective of XModBench is to evaluate whether models preserve *cross-modal consistency* when the same semantic content appears in different modalities. Each item is a four-choice multiple-choice question consisting of a `<context>` (question stem) and four `<candidates>` (answer options). By systematically permuting text (T), vision (V), and audio (A) across the `<context>` and `<candidates>`, we generate six modality configurations of the same question (see Fig. 1 (b) and (d)). This balanced design ensures that no single modality is privileged and enables consistent evaluation across all directions, which supports three diagnostic properties aligned with the goals of our benchmark:

(1) Task competence. Since each task is instantiated uniformly across all modality pairs, we measure competence by averaging accuracy across all context–candidate configurations. This yields a fair estimate of a model's overall capability for each task, independent of modality-specific biases.

(2) Modality disparity. By presenting semantically identical questions under different modality configurations, we keep the content fixed while varying only the modality. For example, to compare audio and vision, we examine cases where text provides the context with audio candidates (T↦A) versus text with visual candidates (T↦V), and similarly compare A↦T against V↦T settings. Differences in accuracy under these controlled comparisons reveal modality disparities, indicating the relative competence across different modalities.

(3) Directional imbalance. We examine inverse settings by swapping the modalities of context and candidates. For example, a model may perform well when vision serves as the context and text provides the options (V↦T), but perform worse when the same semantic content is presented as a text context with visual candidates (T↦V). Such differences indicate asymmetric grounding between the two modalities, and comparable asymmetries are also observed in the audio–text and audio–vision pair.

3.2 Task Taxonomy

XModBench covers five task families with seventeen subtasks, spanning perception, spatial reasoning, temporal reasoning, linguistic understanding, and external knowledge (see Fig. 2). Each task is formulated in the multiple-choice format and follows the modality-balanced configuration described in Section 3.1: a `<context>` is drawn from one modality and four `<candidates>` from another. In this section, we detail the design of these subtasks and specify how each instance is instantiated across modalities within every task.

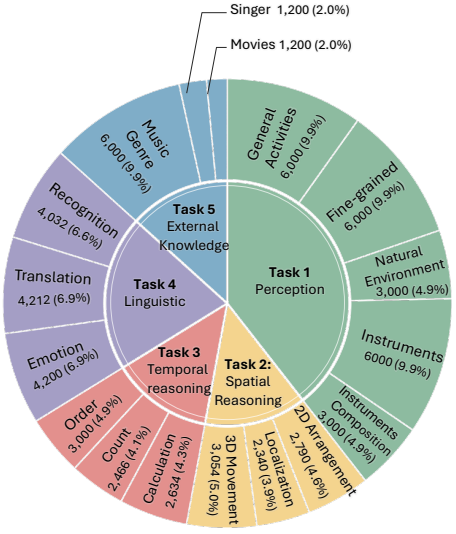

Figure 2: Distribution of XModBench's questions across five task families with specific subtasks.

Task 1. Perception. This task evaluates whether models can recognize the same object, activity, or scene across modalities. For example, a barking dog may appear as an image, as its sound, or as the text description "dog barking." Here, visual inputs are images, audio inputs are recordings of corresponding sounds, and text inputs are short labels or phrases. The data are drawn from diverse domains, including human activities, animal behaviors, musical instruments, and natural environments.

We divide perception into several subtasks. **General activity** recognition mixes candidates from diverse domains to test broad semantic alignment, while **fine-grained activity** recognition restricts candidates to a single domain (e.g., animal species or instrument types), thereby increasing difficulty and requiring precise discrimination. We further design domain-specific subtasks to capture

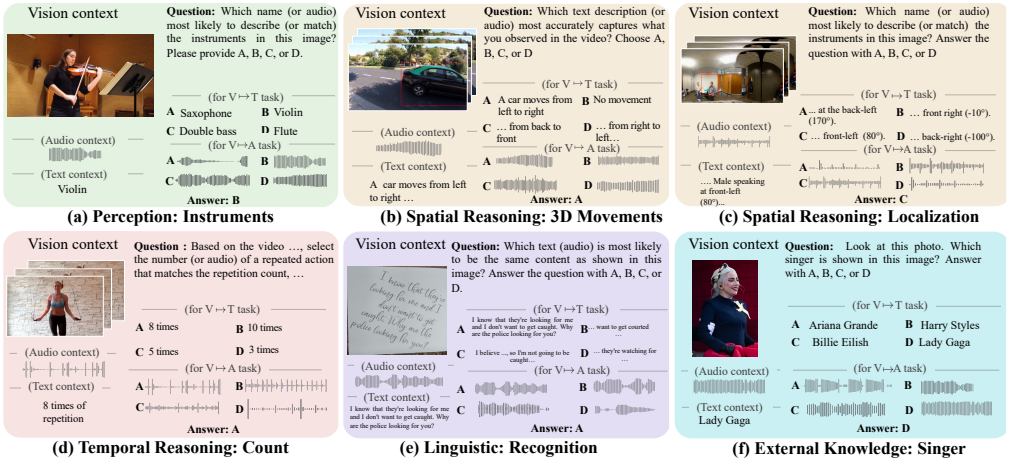

Figure 3: XModBench task examples. We show sample questions from six subtasks in the benchmark. Each question includes possible contexts from different modalities, and for the vision-context example, the candidates are given in either text or audio.

unique challenges: recognizing **natural environments** (e.g., rainfall, wind, fire), distinguishing **instruments** (e.g., violin, bass, cello), and identifying **instrument compositions** where multiple instruments are played together (e.g., violin and bass, or cello and flute). Illustrative examples are shown in Fig. 1(d) and Fig. 3(a).

Task 2. Spatial reasoning. This task evaluates whether models can interpret object positions and motion in 2D and 3D space, which is an important factor in vision–language models (Chen et al., 2024). We extend this ability to the omni-modal setting and design three subtasks. The first is **2D arrangement**, where the model determines the left–right order of objects such as musical instruments. Visual inputs are images of ordered layouts, audio inputs are stereo recordings with left–right cues, and text inputs describe the relative arrangement; distractors are generated by swapping or permuting positions. The second subtask, **3D localization**, using panoramic videos from Shimada et al. (2023), requires identifying the orientation of events in video frames, spatialized audio, and short textual descriptions (e.g., "a man speaking from the front-left"); distractors are produced by shifting the same scene to nearby but incorrect directions through camera or audio rotation. The third subtask, **3D movement understanding**, focuses on motion directions such as left–right or front–back, instantiated with street-view or action videos, spatialized audio of approaching or receding sounds, and textual trajectory descriptions (Fuentes et al., 2022); distractors are clips with incorrect motion patterns or mismatched vehicle types. Examples for the 3D movement and localization tasks are shown in Fig. 3(b) and (c), respectively.

Task 3. Temporal reasoning. This task evaluates whether models can understand **event order** and **frequency** across time in video and audio. We design three subtasks. The first is **temporal order**, where models infer the correct sequence of events from muted video segments, audio clips, or textual descriptions and align them across modalities. The second, **temporal counting**, requires recognizing the number of repeated actions such as tennis hits, jumps, or drum beats, with distractors differing in count. For example, a video may show a tennis player hitting the ball three times, and the model must select the audio clip with exactly three hits or the text "3 times." The third, **temporal calculation**, extends counting by applying simple arithmetic to the repetition number. For instance, if a video shows a person jumping three times and the query applies $2 \times$ count, the correct answer should correspond to six repetitions, given either as an audio clip with six jumps or as the text "6 times." An example of the temporal counting task is in Fig. 3(d).

Task 4. Linguistic understanding. This task covers recognition of linguistic content and interpretation of affective meaning. While prior work separates OCR for vision and ASR for audio (Fu et al., 2024b; Wang et al., 2024), XModBench unifies them in a cross-modal setting. We design three subtasks. The first, **linguistic recognition**, focuses on transcribing text from images, audio, or phrases; correct candidates require word-level precision, while distractors differ by only one or two words (see Fig. 3(e)). The second, **translation**, evaluates English–Chinese translation across modalities, with distractors introducing subtle shifts such as antonyms, degree modifiers (e.g., "very" $\rightarrow$

"a little"), or small changes in numbers and entities. The third, **emotion classification**, targets affective understanding in dialogue: audio inputs are spoken conversations, visual inputs are muted video clips with transcripts, and candidates represent emotions such as joy, sadness, or anger, with distractors drawn from closely related categories.

Task 5. External knowledge. Beyond perceptual and reasoning skills, some tasks require linking multimodal content with world knowledge. We design three subtasks. The first, **movie recognition**, presents audio clips from trailers, visual posters, or short text descriptions of the plot, with candidates drawn from films of similar genres or storylines. The second, **music genre classification**, uses album covers, short audio clips, or textual genre labels, with distractors from closely related genres (e.g., "jazz" vs. "blues"). The third, **singer identification**, provides names, portrait images, or audio clips of songs, with distractors sampled from artists of similar musical styles (see Fig. 3(f)).

3.3 Data Curation

The construction of XModBench follows a three-stage pipeline. We begin by collecting large-scale text–vision–audio triplets across all task domains, then generate task-specific multiple-choice questions, and finally apply both automated filtering and human verification to ensure quality and consistency.

Cross-modal data collection. We curate a large corpus aligned across vision, audio, and text by combining three sources: (i) re-annotated and extended data from existing multimodal datasets, such as adapting VGG-Sound for perception tasks or STARSS23 (Lee et al., 2022; Shimada et al., 2023) for spatial reasoning; (ii) synthetic or model-generated content to cover missing modalities, for example generating speech audio with FireRedTTS (Guo et al., 2025) or producing rendered text images for translation tasks; and (iii) web-collected samples for domains not well represented in public resources, such as singer portraits and songs for the *Singer Identification* task or trailers and posters for *Movie Recognition* from public YouTube videos. This design ensures both coverage and balance across all five task families. Detailed sources and processing procedures are described in Appendix F.

Question candidate generation. To ensure the correctness of both the generated questions and answers, we first construct task-specific multiple-choice templates using our annotated tri-modality data. The question descriptions are then refined by GPT-5 (OpenAI, 2025) solely to improve language fluency and stylistic diversity. Importantly, this refinement does not introduce any new information or alter the underlying semantics. Each question is instantiated with a context and four candidates under the modality-balanced configuration, ensuring consistent evaluation across all modality directions. Distractors are created to be semantically challenging but unambiguous, while templates are diversified with both human-written prompts and LLM-assisted variations.

LLM filtering and human-in-the-loop verification. To guarantee data quality at scale, we first adopt foundation models (OpenAI, 2025; Comanici et al., 2025) to filter out low-quality or ambiguous samples. Human annotators then double-check the filtered results to ensure accuracy. After questions are constructed, an internal round of testing is conducted by annotators, who resolve ambiguities and validate correctness. Feedback from this process is used to regenerate and retest questions until high-quality items are obtained.

Overall, this pipeline yields a high-quality benchmark with diverse and well-aligned multimodal content. More detailed descriptions of dataset sources, generation strategies, and signal-processing techniques are provided in Appendix F.

4 Experiments

4.1 Baselines

We evaluate XModBench on a diverse set of recent omni-modal large language models. The **Gemini series** (Comanici et al., 2025; Team et al., 2024) represents state-of-the-art closed-source omni-modal models, and we include multiple variants ranging from Gemini 1.5 Pro to Gemini 2.5 Pro. Note that OpenAI APIs do not currently support processing audio and visual modalities jointly within a single query; therefore, we omit the GPT series from our evaluation. For open-source systems, we include the latest **Qwen2.5-Omni** (Xu et al., 2025), **Baichuan Omni 1.5** (Li et al., 2025), and **EchoInk-R1** (Xing et al., 2025). Additional open-source omni-modal baselines include **VideoLLaMA 2** (Cheng et al., 2024), **VITA** (Fu et al., 2025b), the **Unified-IO 2** series (Large, XL, and

Table 2: Results on **XModBench**. We report (a) the performance under different input modalities across the full benchmark, and (b) the summary of average accuracy for each of the 5 task families. The highest scores are **bolded**, and the second highest are underlined.

Model	Accuracy on 5 Task Families					Modality Configuration							Avg.
	Perc.	Spat.	Temp.	Ling.	Knwl.	A ↦ T	A ↦ V	T ↦ A	T ↦ V	V ↦ A	V ↦ T	*Std.*	
no context	25.5	24.8	24.9	24.7	25.5	25.1	24.3	25.4	24.8	25.3	25.7	0.4	25.1
Qwen2.5-VL	91.3	51.4	40.9	84.1	77.2	-	-	-	60.1	-	74.7	-	67.4
Intern3.5-VL	87.2	42.7	41.4	75.0	68.7	-	-	-	49.7	-	73.7	-	61.7
PandaGPT	24.6	25.7	24.4	25.5	23.1	24.5	25.0	23.8	25.2	24.5	25.1	0.5	24.7
Unified-IO 2	36.1	23.6	23.8	30.4	26.8	28.9	24.0	25.4	32.0	25.7	32.7	3.7	28.1
Unified-IO 2 XL	42.2	25.0	26.1	30.8	29.5	33.3	27.0	27.1	32.9	26.5	37.4	4.5	30.7
Unified-IO 2 XXL	43.7	28.3	27.7	31.2	34.0	37.4	25.0	31.2	37.8	26.7	39.9	6.3	33.0
VideoLLaMA 2	45.7	33.9	29.2	36.7	36.8	48.6	26.0	25.7	26.5	25.2	66.8	17.4	36.5
VITA	34.8	34.0	29.4	46.1	32.6	40.2	26.0	29.8	26.8	29.9	59.3	12.8	35.4
Baichuan Omni 1.5	58.9	34.9	30.0	62.8	56.7	47.8	35.8	40.5	56.2	38.6	73.0	14.0	48.7
EchoInk-R1	75.8	36.6	37.1	73.3	73.3	64.6	45.9	56.4	60.9	49.9	77.6	11.3	59.2
Qwen2.5-Omni	75.5	38.4	32.3	74.1	72.8	62.0	48.0	55.4	59.6	50.5	76.3	10.1	58.6
Gemini 1.5 Pro	56.2	40.1	37.1	72.6	69.4	52.4	38.2	48.6	70.4	40.7	79.9	16.7	55.0
Gemini 2.0 Flash	66.2	48.4	44.8	70.2	78.1	63.7	49.0	52.2	71.5	47.6	85.2	15.2	61.2
Gemini 2.5 Flash	66.1	48.0	48.6	73.1	82.8	62.6	51.2	55.1	75.7	51.9	86.0	14.2	63.7
Gemini 2.5 Pro	**75.9**	**50.1**	**60.8**	**76.8**	**89.3**	**71.0**	**58.9**	**64.4**	**79.8**	**60.8**	**88.6**	11.7	**70.6**
Human	91.0	89.7	88.9	93.9	93.9	92.4	91.5	91.1	91.8	86.4	95.6	3.0	91.5

XXL variants) (Lu et al., 2024), and **PandaGPT** (Su et al., 2023). Together, these models represent a broad spectrum of both closed- and open-source OLLMs.

4.2 Model Performances

Table 5 reports results across five task families and six cross-modal directions (Audio ↦ Text, Audio ↦ Vision, Text ↦ Audio, Text ↦ Vision, Vision ↦ Audio, Vision ↦ Text). The first subtable summarizes the average accuracy across all tasks for each modality configuration, while the remaining subtables present detailed performance within each task family. The highest scores are **bolded**, and the second highest are underlined. For each model, we also report the overall average accuracy (*Avg.*) and standard deviation (*Std.*) across the six configurations to quantify robustness to modality shifts. Details of the human evaluation are provided in Appendix E.

Performance by task families. Overall, the Gemini 2.0 and 2.5 series outperform all open-source systems. Among open models, Qwen2.5-Omni and EchoInk-R1 are the strongest baselines, surpassing Gemini 1.5 Pro by 3.6 and 4.2 points, respectively. Across the five task families, spatial and temporal reasoning remain the most challenging (Gemini 2.5 Pro achieves 50.1 and 60.8), whereas perception and linguistic tasks reach higher accuracy (75.9 and 76.8). The performance gap between open- and closed-source systems extends beyond spatial and temporal reasoning to external knowledge: while Qwen2.5-Omni and EchoInk-R1 perform comparably to Gemini 2.5 Pro on perception, the latter attains 89.3 on external knowledge. These results highlight persistent bottlenecks in open-source models, as closed-source systems likely benefit from broader web-scale pretraining and stronger spatial–temporal reasoning capabilities.

Performance by modality configurations. We also analyze performance consistency across modality configurations on the same tasks and observe clear divergences. Vision–text settings consistently outperform audio–text ones, confirming that visual representations are more strongly grounded than audio. In perception tasks, accuracy exceeds 90% with vision–text inputs but drops by over 20 points with audio–text. Audio–vision combinations without textual anchors yield the lowest scores, highlighting the difficulty of aligning heterogeneous signals. Among SOTA systems, Gemini 2.5 Pro (Avg. 70.6, Std. 11.7) shows the best balance of accuracy and stability, while Qwen2.5-Omni (Std. 10.1) and EchoInk-R1 (Std. 11.3) are the most consistent open models. By contrast, Gemini 1.5 Pro and Baichuan Omni 1.5 have standard deviations exceeding 14, reflecting weaker robustness to modality variation.

4.3 Modality Disparity Analysis

A key challenge for OLLMs is whether they handle audio, vision, and text equally rather than favoring one modality. XModBench enables this by instantiating identical semantics across modality settings. $\Delta_{T \text{ vs. } V} = (Acc_{A \mapsto \mathbf{V}} - Acc_{A \mapsto \mathbf{T}}) + (Acc_{\mathbf{V} \mapsto A} - Acc_{\mathbf{T} \mapsto A})$,

We quantify disparity via paired subtraction, e.g., which compares configurations that differ only by substituting **text** with **vision**, thereby isolating the effect of modality substitution on accuracy. Results in Fig. 4 show that $\Delta_{T \text{ vs. } A}$ exhibits the strongest disparity (-49 for Gemini 2.5 Pro), $\Delta_{V \text{ vs. } A}$ is moderate (-33), and $\Delta_{T \text{ vs. } V}$ remains smallest (-15). *These findings highlight audio as the most challenging modality, with vision showing moderate gaps and text remaining the most robust.*

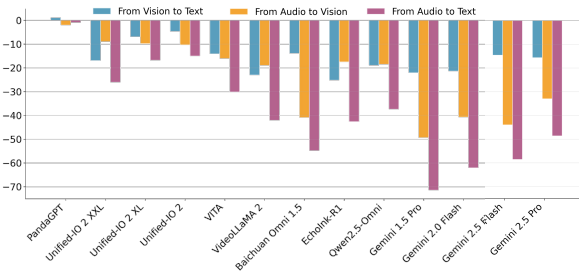

Figure 4: Modality disparity across different configurations. Negative scores indicate performance gaps, with the largest disparities observed between audio and text.

4.4 Directional Imbalance

We test whether models behave consistently when swapping the roles of context and candidates. We define *directional imbalance* as $\Delta_{X \leftrightarrow Y} = Acc(X \mapsto Y) - Acc(Y \mapsto X)$, the accuracy gap between inverse configurations for $(X, Y) \in \{(A, T), (V, T), (V, A)\}$. As shown in Fig. 5, vision–text and audio–text pairs exhibit notable asymmetries: Gemini 2.5 Pro drops by 8.8 points from T↦V to V↦T, and Qwen2.5-Omni shows a 16.6-point gap, while audio–text differences remain around 6–8 points. By contrast, **audio–vision** pairs are nearly symmetric but achieve much lower overall accuracy. *These findings suggest that directional imbalance mainly arises in text–vision and audio–text pairs, likely reflecting training data biases toward text as the dominant output modality.*

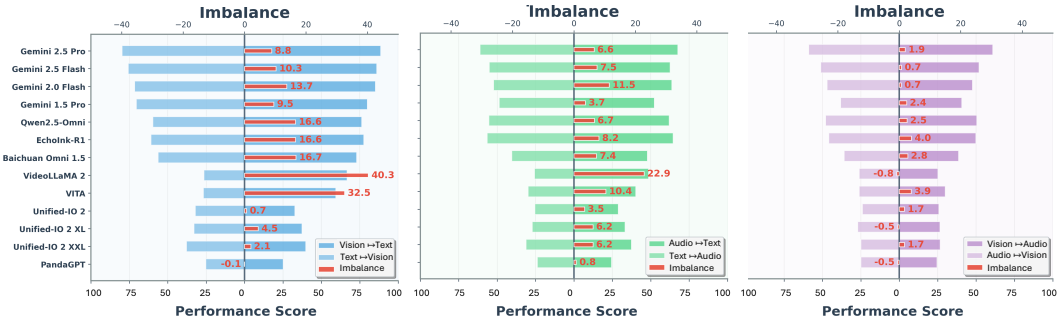

Figure 5: Directional imbalance: accuracy gaps between paired inverse settings among audio, vision and text. Models show clear asymmetries, especially in vision–text and audio–text pairs.

4.5 Failure Case Analysis

To better understand model errors, we prompt systems like Gemini 2.5 Pro and Qwen2.5-Omni to generate reasoning alongside their answers. As shown in Fig. 6, we observe common failure cases that reflect modality performance gaps and alignment issues. Example (a) shows a mismatch between audio-to-text and audio-to-image reasoning: while the model correctly identifies a didgeridoo by text, it fails to select the matching image, revealing inconsistent grounding. In example (b), Qwen-2.5 Omni misinterprets spatial audio motion when switching from audio-to-text to text-to-audio, reversing the vehicle's direction. These errors highlight persistent asymmetries in cross-modal reasoning that only emerge when the same task is posed across different modality combinations.

4.6 Triple-Domain Question Answering

Real-world omni-modal scenarios often present information jointly across modalities rather than in isolation. To approximate this setting, we extend XModBench tasks to an audio–visual context, where both sound and vision are provided in the question stem, while the candidates remain in text.

We evaluate this dual-context configuration using the Gemini series. Compared with single-modality baselines, the results (see Appendix D) show modest but consistent gains, indicating that models can

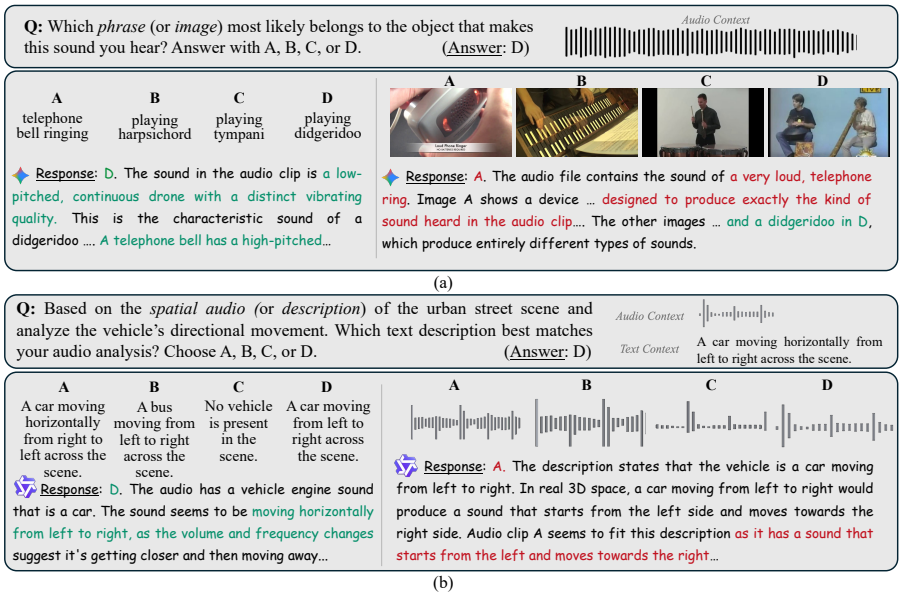

Figure 6: Failure cases. (a) Gemini 2.5 pro correctly identifies a didgeridoo in text but fails to match it with the corresponding image candidates. (b) shows Qwen2.5-Omni misinterprets spatial motion when switching candidates from text to audio. This cases illustrate asymmetries in cross-modal reasoning.

benefit from simultaneous multimodal cues. However, the improvements are not always additive, suggesting that current systems do not yet fully exploit complementary signals across modalities.

Table 3: Overall performance of Gemini models under the dual-context setting (audio+visual context $\mapsto$ text). We compare with pairwise baselines (A $\mapsto$ T and V $\mapsto$ T), and report the stronger unimodal baseline $\max(\text{A} \mapsto \text{T}, \text{V} \mapsto \text{T})$.

Setting	Gemini 1.5 Pro	Gemini 2.0 Flash	Gemini 2.5 Pro
A $\mapsto$ T	52.76	63.71	70.99
V $\mapsto$ T	79.92	85.20	88.60
A+V $\mapsto$ T	82.53 (**+2.61**)	79.84	89.76

5 Discussion

Our benchmark results serve as a diagnostic tool, revealing how underlying data composition and training methodologies shape model behaviors. By correlating performance patterns with known model architectures and training reports, we derive three critical insights regarding interleaved data, domain coverage, and post-training dynamics.

5.1 Interleaved Data Correlates with Directional Symmetry

A key observation from our imbalance analysis is the link between interleaved training data and modality-swap robustness.

- **Balanced performance in interleaved models:** Public official reports indicate that models such as *Qwen-Omni* and Google's *Gemini* series incorporate massive-scale interleaved multimodal corpora (e.g., narrated videos, mixed audio–vision documents). Our benchmark corroborates this: these models exhibit relatively small performance gaps between Audio$\rightarrow$Vision and Vision$\rightarrow$Audio tasks. This suggests that seeing modalities appear interchangeably in context allows the model to build symmetric cross-modal bridges.
- **Asymmetry in lightly interleaved models:** Conversely, models trained primarily on disjoint datasets exhibit significant directional asymmetry. For instance, despite strong backbones, models relying on open-source data with limited interleaved audio–vision instruction pairs show a distinct bias. They often perform well in one direction (anchored by their dominant modality) but fail to generalize when the source-target modalities are swapped, indicating that insufficient interleaved supervision hinders directional robustness.

5.2 Domain Coverage Gaps and Encoder Bias

Performance inconsistencies across specific sub-tasks reveal "blind spots" in the training data distribution coverange, particularly regarding to the audio data.

- **Spoken vs. Non-Spoken Bias:** Many models utilizing speech-centric encoders (e.g., Whisper) show a sharp performance drop on non-verbal acoustic tasks, such as environmental sound classification and spatial reasoning. This implies a data-domain imbalance where the model is over-fitted to spoken language features at the expense of general acoustic awareness.
- **Specific Task Domain:** Distinct gaps in specific categories act as fingerprints for missing training data. For example, despite its high overall capacity, *Gemini 1.5* demonstrates limited capability in musical reasoning, suggesting an absence of music-theory-oriented data in its training corpus. Similarly, *EchoInk-R1* struggles with spatial-vision tasks relative to related families, pointing to a lack of spatial-centric visual content.

5.3 The Data Effect for Post-Training

A comparative analysis between *EchoInk-R1* and the *Qwen* series highlights how post-training strategies can alter—and sometimes degrade—multimodal alignment when the training data is limited, even though reinforcement learning is commonly assumed to improve generalization. While *EchoInk-R1* utilizes the *OmniInstruct* corpus (focused on spoken instruction following), *Qwen* incorporates interleaved multimodal conversations during post-training. This divergence leads to notable behavioral differences:

- **Alignment Erosion:** EchoInk-R1 shows decreased performance in cross-modal (AV/VA) tasks compared to Qwen. This suggests that aggressive fine-tuning on spoken-only instructions may cause "catastrophic forgetting" of the fine-grained cross-modal grounding acquired during pre-training.
- **Inefficacy for Spatial Domains:** The lack of improvement in spatial audio tasks for EchoInk, despite heavy instruction tuning, reinforces that modality balance is strictly required during post-training. Unimodal or text-centric fine-tuning cannot compensate for, and may actively harm, the model's ability to process complex multimodal signals.

We believe that the findings discussed above highlight the value of XModBench not only as an evaluation tool but also as a source of insight for model development. The broader challenge remains the limited transparency of training data in many state-of-the-art multimodal systems. XModBench provides a practical way to study the impact of such opacity by enabling controlled comparisons across models with different training paradigms. For future model developers, this allows clearer understanding of how data choices influence multimodal alignment. For existing model builders, greater openness about data sources would further support the refinement of data pipelines and help reduce modality inconsistencies.

Current benchmarks lack the multimodal-invariant structure and modality-swap design required to expose these effects, underscoring the role of XModBench in advancing both analysis and informed model development.

6 Conclusion

We introduced **XModBench**, a benchmark for diagnosing cross-modal consistency in omni-language models. By systematically interleaving audio, vision, and text across diverse tasks, XModBench enables fine-grained evaluation of modality disparities, directional imbalances, and modality invariant capability. Our results show that audio remains the most challenging modality, that models often behave asymmetrically in inverse settings such as text-vision and audio-text, and that combining audio and vision yields only modest gains. Overall, while current systems are strong in perception and language, they still lack stable and consistent reasoning across modalities, leaving ample room for progress toward truly modality-agnostic intelligence.

ETHICS STATEMENT

Our study does not involve private or sensitive personal data. All audiovisual samples are obtained from publicly available official sources, including previously published research datasets, content hosted on established open source platforms such as Hugging Face and Kaggle. For all newly generated labels and annotations, we perform manual verification to ensure correctness and to remove any potentially inappropriate content. All web-curated data are from publicly accessible and previously published sources without requiring special authentication. All materials are used solely for non-commercial academic research. We do not redistribute copyrighted video or audio; only derived features, annotations, and evaluation results are released.

REFERENCES

V.R. Algazi, R.O. Duda, D.M. Thompson, and C. Avendano. The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pp. 99–102, 2001. doi: 10.1109/ASPAA.2001.969552.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16867–16876, 2021.

Ssu-Yen Chen, Chao-Chun Hsu, Chuan-Chun Kuo, and Lun-Wei Ku. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*, 2018.

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

Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. Vidhal: Benchmarking temporal hallucinations in vision llms. *arXiv preprint arXiv:2411.16771*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024a. URL https://arxiv.org/abs/2306.13394.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025b.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024b.

Magdalena Fuentes, Bea Steers, Pablo Zinemanas, Martin Rocamora, Luca Bondi, Julia Wilkins, Qianyi Shi, Yao Hou, Samarjit Das, Xavier Serra, et al. Urban sound & sight: Dataset and benchmark for audio-visual urban scene understanding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 141–145. IEEE, 2022.

GbotHQ. Ocr dataset rendering. GitHub: https://github.com/GbotHQ/ocr-dataset-rendering/, 2024. MIT License, Accessed: YYYY-MM-DD.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

Hao-Han Guo, Yao Hu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, and Kun Xie. Fireredtts-1s: An upgraded streamable foundation text-to-speech system. *arXiv preprint arXiv:2503.20499*, 2025.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

Dongjin Kim, Sung Jin Um, Sangmin Lee, and Jung Uk Kim. Learning to visually localize sound sources from mixtures without prior source knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26467–26476, 2024.

Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *European Conference on Computer Vision*, pp. 34–50. Springer, 2022.

Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535, 2018.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024a.

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.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024b.

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. Springer, 2024.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024.

Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms. *arXiv preprint arXiv:2506.05328*, 2025.

Piotr Majdak, Péter Balazs, and Bernhard Laback. Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

Juan F. Montesinos, Olga Slizovskaia, and Gloria Haro. Solos: A dataset for audio-visual music analysis. *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2020. URL https://api.semanticscholar.org/CorpusID:219687731.

Andrada Olteanu. Gtzan dataset: Music genre classification. Kaggle Dataset: https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification/, 2024. Accessed: YYYY-MM-DD.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, August 2025.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19821–19829, 2025.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Rada Mihalcea, and Erik Cambria. MELD: A multimodal multi-party dataset for emotion recognition in conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 527–536, 2019.

Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society*, 45(6):456–466, 1997.

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. *arXiv preprint arXiv:2410.19168*, 2024.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, et al. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in neural information processing systems*, 36:72931–72957, 2023.

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. *arXiv preprint arXiv:2406.15704*, 2024.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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. *arXiv preprint arXiv:2406.16020*, 2024.

Christoph Wendler. Renderedtext. Hugging Face Dataset: https://huggingface.co/datasets/wendlerc/RenderedText, 2024. Accessed: YYYY-MM-DD.

Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

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.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Acvubench: Audio-centric video understanding benchmark. *arXiv preprint arXiv:2503.19951*, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2031–2041, 2021.

Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. Cross-modal consistency in multimodal large language models. *arXiv preprint arXiv:2411.09273*, 2024.

Yunhua Zhang, Ling Shao, and Cees GM Snoek. Repetitive activity counting by sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14070–14079, 2021.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *arXiv preprint arXiv:2505.20256*, 2025.

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

APPENDIX

A MINI BENCHMARK RESULT

We will release a standardized 6k-sample XModBench-Lite benchmark, consisting of 5 task families $\times$ 6 modality–configuration settings, with 200 examples per setting. The dataset is balanced across both task families and modality directions. The overall performance (see Tab. 4) trends and error patterns closely mirror those reported in Tab.2 of the main paper.

Table 4: Performance on 6k version of XModBench

Model	Accuracy on 5 Task Families					Modality Configuration						Avg.
	Perc.	Spat.	Temp.	Ling.	Knwl.	A $\rightarrow$ T	A $\rightarrow$ V	T $\rightarrow$ A	T $\rightarrow$ V	V $\rightarrow$ A	V $\rightarrow$ T	
w/o context	25.3	25.1	24.8	24.4	25.2	26.5	24.8	24.2	24.1	25.5	25.1	25.0
Qwen2.5-VL	91.5	51.9	40.5	84.3	76.5	-	-	-	68.2	-	72.8	60.5
Intern3.5-VL	88.2	41.8	48.5	75.8	62.4	-	-	-	46.5	-	73.1	69.8
PandaGPT	24.9	25.3	23.8	24.7	21.3	25.2	25.5	22.8	24.9	24.8	23.1	24.4
Unified-IO 2	36.5	24.8	24.1	31.2	27.5	29.8	24.2	25.5	32.1	25.9	33.5	28.4
Unified-IO 2 XL	42.5	26.3	26.0	32.5	30.6	33.8	26.8	27.5	34.2	27.1	38.8	31.2
Unified-IO 2 XXL	44.1	29.0	27.3	32.9	34.7	38.0	26.3	31.8	38.5	27.3	40.5	33.6
VideoLLaMA 2	46.1	34.2	29.0	37.5	37.4	49.1	26.2	26.0	27.3	25.5	67.8	36.9
VITA	35.6	31.8	29.2	46.1	30.5	45.2	26.5	26.8	26.1	24.5	52.5	33.6
Baichuan Omni 1.5	59.5	35.5	30.8	63.9	57.2	48.5	36.2	41.1	57.2	39.1	74.5	49.5
EchoInk-R1	73.1	35.8	36.4	73.3	72.4	66.5	42.1	56.0	65.5	46.8	72.3	53.2
Qwen2.5-Omni	**78.6**	37.1	31.2	74.2	77.8	69.5	45.2	54.5	58.1	56.8	74.5	51.4
Gemini 1.5 Pro	56.8	40.8	38.0	71.0	69.9	53.1	38.5	49.2	71.2	41.4	80.4	55.7
Gemini 2.0 Flash	65.4	48.9	41.5	72.2	71.2	68.5	44.2	50.1	74.3	47.5	**88.9**	67.2
Gemini 2.5 Flash	66.5	47.1	45.3	74.4	81.2	64.8	49.4	57.5	**77.2**	51.0	81.6	68.6
Gemini 2.5 Pro	74.8	**59.3**	**60.2**	**75.8**	**89.1**	**76.8**	**50.5**	**63.2**	75.1	**61.2**	82.5	**71.8**

B MODALITY CONFIGURATION SCORE UNDER FIVE TASK

Table 5 reports the detailed results for all six modality–configuration settings (A↦T, A↦V, T↦A, T↦V, V↦A, V↦T) across the five task families in XModBench (Perception, Spatial, Temporal, Linguistic, and Knowledge), as well as the overall average score on the full benchmark.

Table 5: Results on **XModBench** across 5 task families and 6 predefined cross-modal directions among **T**ext, **V**ision, and **A**udio. The first block reports the average accuracy across all tasks, followed by Task 1–5 (Perception, Spatial, Temporal, Linguistic, External knowledge). Scores are color-coded as < 30, 30–60, 60–90, ≥ 90, with the best in each column highlighted in **bold**.

Model	Overall Average								Task 1 - Perception							
	A↦T	A↦V	T↦A	T↦V	V↦A	V↦T	*Avg.*	*Std.*	A↦T	A↦V	T↦A	T↦V	V↦A	V↦T	*Avg.*	*Std.*
PandaGPT	24.5	25.0	23.8	25.2	24.5	25.1	24.7	0.5	24.5	24.7	24.8	24.5	24.6	24.7	24.6	0.1
Unified-IO 2	28.9	24.0	25.4	32.0	25.7	32.7	28.1	3.7	35.5	25.3	26.3	55.9	29.1	44.7	36.1	12.1
Unified-IO 2 XL	33.3	27.0	27.1	32.9	26.5	37.4	30.7	4.5	53.3	27.9	30.3	**59.1**	27.6	55.0	42.2	15.0
Unified-IO 2 XXL	37.4	25.0	31.2	37.8	26.7	39.9	33.0	6.3	55.0	26.9	39.0	64.2	26.7	50.2	43.7	15.4
VideoLLaMA 2	48.6	26.0	25.7	26.5	25.2	66.8	36.5	17.4	**74.7**	26.6	28.3	26.8	26.5	91.5	45.7	29.4
VITA	40.2	26.0	29.8	26.8	29.9	59.3	35.4	12.8	37.1	25.4	27.0	23.7	26.4	69.1	34.8	17.5
Baichuan Omni 1.5	47.8	35.8	40.5	56.2	38.6	73.0	48.7	14.0	42.7	36.3	45.6	87.8	50.3	90.7	58.9	24.0
EchoInk-R1	64.6	45.9	56.4	60.9	49.9	77.6	59.2	11.3	74.1	58.5	**69.3**	91.6	67.7	93.4	75.8	13.9
Qwen2.5-Omni	62.0	48.0	55.4	59.6	50.5	76.3	58.6	10.1	72.9	59.1	69.2	91.2	68.5	92.0	75.5	13.3
Gemini 1.5 Pro	52.4	38.2	48.6	70.4	40.7	79.9	55.0	16.7	41.0	27.9	45.0	95.8	32.1	95.3	56.2	31.1
Gemini 2.0 Flash	63.7	49.0	52.2	71.5	47.6	85.2	61.2	15.2	56.8	45.0	54.2	92.7	55.1	93.4	66.2	21.2
Gemini 2.5 Flash	62.6	51.2	55.1	75.7	51.9	86.0	63.7	14.2	52.6	44.3	53.4	95.4	56.0	95.0	66.1	22.8
Gemini 2.5 Pro	**71.0**	**58.9**	**64.4**	**79.8**	**60.8**	**88.6**	**70.6**	11.7	62.3	57.4	68.5	**97.1**	**72.6**	**97.6**	**75.9**	17.4
Human	92.4	91.5	91.1	91.8	86.4	95.6	91.5	3.0	92.9	94.2	91.3	89.2	85.4	92.9	91.0	3.2

Model	Task 2 - Spatial								Task 3 - Temporal							
	A↦T	A↦V	T↦A	T↦V	V↦A	V↦T	*Avg.*	*Std.*	A↦T	A↦V	T↦A	T↦V	V↦A	V↦T	*Avg.*	*Std.*
PandaGPT	25.5	26.6	26.0	27.2	25.8	23.1	25.7	1.4	21.9	25.3	24.8	26.0	24.5	23.9	24.4	1.4
Unified-IO 2	26.0	20.7	22.4	25.0	23.1	24.7	23.6	1.9	22.7	22.4	25.1	24.3	25.8	22.4	23.8	1.5
Unified-IO 2 XL	24.8	23.0	25.8	26.0	26.0	24.5	25.0	1.2	22.3	24.5	28.8	22.1	26.0	32.7	26.1	4.1
Unified-IO 2 XXL	29.6	23.6	30.9	25.5	29.5	30.7	28.3	3.0	24.3	27.4	25.3	29.6	25.2	34.4	27.7	3.8
VideoLLaMA 2	43.9	27.8	24.4	27.5	25.2	54.3	33.9	12.3	31.0	25.0	27.7	25.9	25.8	39.8	29.2	5.6
VITA	42.3	28.9	24.6	30.9	25.1	52.2	34.0	11.0	31.1	25.1	26.1	24.6	27.6	41.7	29.4	6.5
Baichuan Omni 1.5	38.1	28.0	25.1	31.7	25.3	61.2	34.9	13.8	27.0	25.2	23.9	26.9	25.0	52.2	30.0	10.9
EchoInk-R1	41.3	27.2	26.8	34.0	28.0	62.2	36.6	13.7	38.2	26.2	38.6	31.1	26.9	61.6	37.1	13.1
Qwen2.5-Omni	41.8	31.2	26.7	34.4	28.6	67.8	38.4	15.3	26.9	28.7	36.6	25.6	25.3	50.8	32.3	10.0
Gemini 1.5 Pro	37.2	31.2	24.5	51.4	23.7	72.8	40.1	19.0	37.1	27.2	31.0	47.3	24.5	55.7	37.1	12.2
Gemini 2.0 Flash	45.2	**43.1**	29.2	56.4	33.5	83.0	48.4	20.4	51.8	30.8	38.6	48.0	27.4	72.0	44.8	16.3
Gemini 2.5 Flash	**45.6**	31.4	30.2	71.2	26.7	83.2	48.0	23.8	48.8	39.6	39.1	51.4	38.0	74.6	48.6	13.9
Gemini 2.5 Pro	41.0	32.9	**32.1**	**75.8**	**30.3**	**88.3**	**50.1**	25.4	**76.4**	**54.4**	**57.7**	**55.4**	**50.6**	**70.6**	**60.8**	10.3
Human	93.3	93.3	81.7	86.7	86.7	96.7	89.7	5.7	90.0	85.0	86.7	91.7	83.3	96.7	88.9	4.9

Model	Task 4 - Linguistic								Task 5 - External Knowledge							
	A↦T	A↦V	T↦A	T↦V	V↦A	V↦T	*Avg.*	*Std.*	A↦T	A↦V	T↦A	T↦V	V↦A	V↦T	*Avg.*	*Std.*
PandaGPT	28.0	24.3	20.7	24.7	24.3	31.3	25.5	3.6	22.8	23.9	22.6	23.6	23.3	22.6	23.1	0.5
Unified-IO 2	32.4	27.5	27.6	27.9	25.2	41.7	30.4	6.0	28.2	24.2	25.8	27.1	25.3	29.9	26.8	2.1
Unified-IO 2 XL	34.4	31.7	24.5	28.8	23.6	41.8	30.8	6.8	31.9	27.9	26.2	28.6	29.5	32.9	29.5	2.5
Unified-IO 2 XXL	39.9	23.0	25.5	30.1	22.3	46.6	31.2	9.9	38.4	23.9	35.3	39.5	29.7	37.4	34.0	6.0
VideoLLaMA 2	50.3	25.2	24.2	25.2	24.1	71.2	36.7	19.8	42.9	25.5	23.9	27.0	24.4	76.9	36.8	20.9
VITA	52.2	26.8	47.1	29.9	47.9	72.5	46.1	16.6	38.5	24.1	24.2	24.7	22.6	61.2	32.6	15.2
Baichuan Omni 1.5	77.0	45.7	65.8	51.8	58.7	77.6	62.8	13.1	54.3	43.9	41.9	82.9	33.7	83.2	56.7	21.5
EchoInk-R1	86.0	57.4	74.6	64.4	70.1	87.3	73.3	11.8	83.3	60.4	72.7	83.6	56.6	83.3	73.3	12.3
Qwen2.5-Omni	85.6	61.8	73.6	64.6	71.5	87.5	74.1	10.6	83.0	59.2	70.7	82.5	58.6	83.2	72.8	11.8
Gemini 1.5 Pro	86.2	52.4	72.3	68.7	70.7	85.5	72.6	12.0	62.3	52.5	70.2	88.8	52.3	90.3	69.4	17.0
Gemini 2.0 Flash	83.6	57.5	68.6	67.3	60.9	83.4	70.2	11.1	81.2	68.3	70.5	93.1	61.3	94.2	78.1	13.6
Gemini 2.5 Flash	84.1	**68.3**	70.9	66.8	64.4	84.4	73.1	8.9	82.0	72.2	81.7	93.9	74.5	92.7	82.8	9.0
Gemini 2.5 Pro	**84.9**	67.5	**75.5**	**76.1**	**65.8**	**91.4**	**76.8**	9.9	**90.3**	**82.5**	**88.2**	**94.6**	**84.8**	**95.1**	**89.3**	5.1
Human	89.2	96.7	97.5	93.3	91.7	95.0	93.9	2.8	96.7	88.3	98.3	98.3	85.0	96.7	93.9	5.8

C TASK SPECIFICED MODEL PERFORMANCE

C.1 TASK 1: PERCEPTUAL TASK

Table 6: T1 (Perception) Results

Model		Perception Task				
Model	**Task**	**General**	**General - Hard**	**Scene**	**Instruments**	**Instruments-multi**
Gemini 2.5 Pro	Audio ↦ Text	81.05	71.39	67.20	47.75	44.09
	Audio ↦ Vision	76.26	65.25	64.60	44.30	36.60
	Text ↦ Audio	79.95	79.22	75.05	59.05	49.30
	Text ↦ Vision	98.90	97.87	90.80	97.90	99.80
	Vision ↦ Audio	88.73	79.35	84.40	61.92	48.79
	Vision ↦ Text	98.37	97.50	95.00	97.19	99.80

Continued on next page

Table 6 – continued from previous page

Model		Perception Task				
Model	**Task**	**General**	**General - Hard**	**Scene**	**Instruments**	**Instruments-multi**
Gemini 2.5 Flash	Audio ↦ Text	81.00	50.00	51.01	45.82	35.27
	Audio ↦ Vision	62.63	50.39	47.60	30.99	29.92
	Text ↦ Audio	79.80	59.13	57.34	37.90	32.99
	Text ↦ Vision	98.96	91.45	90.20	96.50	99.74
	Vision ↦ Audio	82.10	60.59	67.54	39.80	29.92
	Vision ↦ Text	98.39	96.62	92.60	89.88	97.27
Gemini 2.0 Flash	Audio ↦ Text	81.10	62.07	54.00	47.05	39.80
	Audio ↦ Vision	67.45	51.68	49.80	31.50	24.80
	Text ↦ Audio	79.95	60.64	53.80	38.80	37.60
	Text ↦ Vision	98.95	91.45	80.40	96.90	95.60
	Vision ↦ Audio	82.50	66.45	53.00	37.90	35.40
	Vision ↦ Text	96.95	90.22	84.80	96.70	98.40
Gemini 1.5 Pro	Audio ↦ Text	80.90	36.38	29.20	30.93	27.40
	Audio ↦ Vision	34.35	30.00	28.40	23.60	23.00
	Text ↦ Audio	80.25	45.88	41.80	31.10	26.20
	Text ↦ Vision	98.75	95.88	89.40	98.10	97.00
	Vision ↦ Audio	41.85	34.38	31.80	27.70	25.00
	Vision ↦ Text	95.10	94.62	87.80	98.90	100.00
Qwen2.5 Omni	Audio ↦ Text	80.00	74.50	79.20	69.37	61.40
	Audio ↦ Vision	71.10	54.30	59.80	58.30	51.80
	Text ↦ Audio	81.20	69.90	78.80	67.40	48.80
	Text ↦ Vision	94.90	87.70	89.60	90.80	92.80
	Vision ↦ Audio	83.90	68.50	61.20	68.30	60.60
	Vision ↦ Text	97.50	87.00	88.00	91.80	95.80
EchoInk	Audio ↦ Text	87.55	74.80	77.10	68.20	63.00
	Audio ↦ Vision	74.60	58.40	49.00	58.20	52.10
	Text ↦ Audio	84.57	66.40	79.40	69.14	46.80
	Text ↦ Vision	95.00	91.80	88.40	89.78	92.80
	Vision ↦ Audio	82.80	68.80	60.40	69.14	57.52
	Vision ↦ Text	96.00	95.20	87.80	92.38	95.79
Baichuan Omni 1.5	Audio ↦ Text	55.85	44.05	46.40	36.44	31.00
	Audio ↦ Vision	44.45	37.43	43.20	29.60	26.60
	Text ↦ Audio	63.80	50.90	53.60	32.80	27.00
	Text ↦ Vision	97.35	88.10	81.20	84.50	88.00
	Vision ↦ Audio	68.25	53.75	58.80	38.40	32.20
	Vision ↦ Text	95.90	87.12	86.80	92.70	90.80
VideoLLaMA 2	Audio ↦ Text	86.84	76.85	77.26	75.86	56.89
	Audio ↦ Vision	26.82	26.82	24.45	29.26	25.82
	Text ↦ Audio	30.69	28.21	28.89	28.45	25.07
	Text ↦ Vision	25.49	27.49	26.03	29.25	25.89
	Vision ↦ Audio	29.28	27.47	25.30	29.26	21.04
	Vision ↦ Text	97.05	91.48	87.45	89.40	92.23
VITA	Audio ↦ Text	43.30	32.99	39.18	39.18	30.93
	Audio ↦ Vision	22.68	20.62	28.87	28.87	25.77
	Text ↦ Audio	28.96	24.24	24.92	28.28	28.62
	Text ↦ Vision	20.62	25.77	31.96	21.65	18.56
	Vision ↦ Audio	23.57	29.29	24.92	25.93	28.28
	Vision ↦ Text	64.95	73.20	58.76	74.23	74.23
Unified IO 2	Audio ↦ Text	49.05	45.26	32.04	26.46	24.44
	Audio ↦ Vision	27.00	26.84	30.65	19.40	22.45
	Text ↦ Audio	26.68	25.86	25.27	27.64	26.20
	Text ↦ Vision	73.28	56.44	72.67	27.26	49.89
	Vision ↦ Audio	27.89	24.09	43.22	24.05	26.26
	Vision ↦ Text	55.83	44.21	32.81	48.66	41.86
Unified IO 2 XL	Audio ↦ Text	76.64	71.68	57.82	33.68	26.87
	Audio ↦ Vision	28.04	25.21	34.89	22.29	29.22
	Text ↦ Audio	39.47	28.46	33.02	23.80	26.84
	Text ↦ Vision	81.89	68.28	69.87	22.09	53.29
	Vision ↦ Audio	26.46	24.43	35.05	24.80	27.03
	Vision ↦ Text	61.10	51.83	53.06	60.49	48.50
Unified IO 2 XXL	Audio ↦ Text	83.63	71.20	45.88	41.45	32.83
	Audio ↦ Vision	29.07	23.28	27.41	27.09	27.87
	Text ↦ Audio	59.87	44.10	35.40	27.68	28.09
	Text ↦ Vision	86.07	73.08	71.29	36.07	54.44
	Vision ↦ Audio	28.66	29.81	24.82	24.01	26.07
	Vision ↦ Text	53.46	48.64	40.49	61.85	46.68
PandaGPT	Audio ↦ Text	25.03	28.80	24.49	24.30	19.99
	Audio ↦ Vision	26.52	27.37	24.63	24.89	20.15

Continued on next page

Table 6 – continued from previous page

Model		Perception Task				
Model	**Task**	**General**	**General - Hard**	**Scene**	**Instruments**	**Instruments-multi**
PandaGPT	Text ↦ Audio	25.40	29.65	24.25	24.77	20.08
	Text ↦ Vision	25.07	28.68	24.22	24.52	20.19
	Vision ↦ Audio	25.26	28.81	24.50	24.52	19.85
	Vision ↦ Text	25.26	28.70	24.64	24.90	20.05

C.2 TASK 2: SPATIAL REASONING

Table 7: T2 (Spatial) Task Results

Model		Spatial Task		
Model	**Task**	**Arrangement**	**Moving Direction**	**Indoor**
Gemini 2.5 Pro	Audio ↦ Text	28.82	69.39	24.87
	Audio ↦ Vision	24.73	40.65	33.38
	Text ↦ Audio	30.09	39.02	27.09
	Text ↦ Vision	95.70	58.85	72.73
	Vision ↦ Audio	29.01	38.10	23.86
	Vision ↦ Text	95.21	85.23	84.56
Gemini 2.5 Flash	Audio ↦ Text	27.53	83.53	25.64
	Audio ↦ Vision	26.54	36.03	31.61
	Text ↦ Audio	25.81	35.34	29.37
	Text ↦ Vision	91.40	66.44	55.71
	Vision ↦ Audio	27.44	26.44	26.12
	Vision ↦ Text	91.40	84.05	74.25
Gemini 2.0 Flash	Audio ↦ Text	28.82	82.71	24.10
	Audio ↦ Vision	26.45	37.58	35.38
	Text ↦ Audio	27.31	39.41	21.01
	Text ↦ Vision	67.53	66.99	34.62
	Vision ↦ Audio	25.81	45.78	28.86
	Vision ↦ Text	89.25	99.02	60.76
Gemini 1.5 Pro	Audio ↦ Text	29.25	57.37	24.87
	Audio ↦ Vision	27.10	32.87	33.59
	Text ↦ Audio	19.25	34.26	20.00
	Text ↦ Vision	64.30	50.80	39.23
	Vision ↦ Audio	23.66	21.82	25.57
	Vision ↦ Text	95.48	80.00	43.04
Qwen2.5 Omni	Audio ↦ Text	21.29	75.28	28.89
	Audio ↦ Vision	28.60	35.83	29.23
	Text ↦ Audio	20.22	31.52	28.35
	Text ↦ Vision	45.38	26.98	30.77
	Vision ↦ Audio	23.87	34.69	27.34
	Vision ↦ Text	80.86	81.63	41.01
EchoInk	Audio ↦ Text	27.79	61.62	34.34
	Audio ↦ Vision	24.97	25.59	30.98
	Text ↦ Audio	26.60	28.96	24.92
	Text ↦ Vision	46.80	25.59	29.63
	Vision ↦ Audio	24.88	31.31	27.95
	Vision ↦ Text	80.13	61.62	44.78
Baichuan Omni 1.5	Audio ↦ Text	28.39	71.43	14.36
	Audio ↦ Vision	28.17	28.51	27.18
	Text ↦ Audio	22.37	27.21	25.57
	Text ↦ Vision	35.70	36.32	23.08
	Vision ↦ Audio	25.38	22.95	27.59
	Vision ↦ Text	71.40	82.95	29.37

Continued on next page

Table 7 – continued from previous page

Model		Spatial Task		
Model	**Task**	**Arrangement**	**Moving Direction**	**Indoor**
VideoLLaMA 2	Audio ↦ Text	31.40	62.44	37.76
	Audio ↦ Vision	27.40	27.75	28.22
	Text ↦ Audio	26.76	27.04	19.53
	Text ↦ Vision	27.36	27.01	28.25
	Vision ↦ Audio	25.63	29.28	20.77
	Vision ↦ Text	46.96	84.21	31.70
VITA	Audio ↦ Text	29.90	77.32	19.59
	Audio ↦ Vision	30.93	26.80	28.87
	Text ↦ Audio	23.23	25.59	25.00
	Text ↦ Vision	29.90	31.96	30.93
	Vision ↦ Audio	24.92	25.59	24.66
	Vision ↦ Text	57.73	55.67	43.30
Unified IO 2	Audio ↦ Text	23.03	20.47	34.40
	Audio ↦ Vision	21.98	17.20	22.89
	Text ↦ Audio	23.50	20.69	22.87
	Text ↦ Vision	25.63	24.03	25.22
	Vision ↦ Audio	24.09	17.32	27.86
	Vision ↦ Text	28.60	24.10	21.49
Unified IO 2 XL	Audio ↦ Text	23.09	28.42	22.88
	Audio ↦ Vision	22.20	20.09	26.75
	Text ↦ Audio	24.82	22.92	29.70
	Text ↦ Vision	24.18	24.25	29.56
	Vision ↦ Audio	24.12	21.78	32.17
	Vision ↦ Text	27.41	24.10	21.93
Unified IO 2 XXL	Audio ↦ Text	22.58	30.07	36.18
	Audio ↦ Vision	24.54	24.02	22.37
	Text ↦ Audio	25.85	38.11	28.71
	Text ↦ Vision	25.39	30.02	21.10
	Vision ↦ Audio	25.45	28.33	34.77
	Vision ↦ Text	30.36	30.91	30.80
PandaGPT	Audio ↦ Text	25.42	25.62	25.44
	Audio ↦ Vision	27.22	25.63	26.91
	Text ↦ Audio	27.06	25.58	25.27
	Text ↦ Vision	27.01	25.95	28.57
	Vision ↦ Audio	27.16	25.53	24.57
	Vision ↦ Text	21.19	25.72	22.34

C.3 Task 3: Temporal Reasoning

Table 8: T3 (Temporal) Task Results

Model		Temporal Task		
Model	**Task**	**Order**	**Counting**	**Calculation**
Gemini 2.5 Pro	Audio ↦ Text	96.18	57.36	75.78
	Audio ↦ Vision	95.38	37.88	29.87
	Text ↦ Audio	95.39	50.00	27.60
	Text ↦ Vision	99.80	35.85	30.63
	Vision ↦ Audio	96.35	34.70	20.65
	Vision ↦ Text	99.80	40.58	71.46
Gemini 2.5 Flash	Audio ↦ Text	41.40	49.60	55.40
	Audio ↦ Vision	58.99	33.07	26.88
	Text ↦ Audio	61.00	29.40	27.00
	Text ↦ Vision	99.15	29.45	25.51

Continued on next page

Table 8 – continued from previous page

Model		Temporal Task		
Model	**Task**	**Order**	**Counting**	**Calculation**
	Vision ↦ Audio	63.39	26.58	24.13
	Vision ↦ Text	99.20	53.37	71.22
Gemini 2.0 Flash	Audio ↦ Text	43.60	52.60	59.20
	Audio ↦ Vision	33.40	30.17	28.93
	Text ↦ Audio	61.40	28.80	25.60
	Text ↦ Vision	81.40	33.33	29.16
	Vision ↦ Audio	33.40	28.22	20.65
	Vision ↦ Text	99.20	57.87	58.90
Gemini 1.5 Pro	Audio ↦ Text	34.40	30.00	47.00
	Audio ↦ Vision	32.00	24.44	25.10
	Text ↦ Audio	38.60	30.20	24.20
	Text ↦ Vision	82.00	33.88	26.14
	Vision ↦ Audio	27.60	23.87	21.88
	Vision ↦ Text	98.40	25.26	43.56
Qwen2.5 Omni	Audio ↦ Text	28.20	25.80	26.60
	Audio ↦ Vision	34.80	22.22	28.96
	Text ↦ Audio	63.80	19.40	26.60
	Text ↦ Vision	24.80	22.90	28.96
	Vision ↦ Audio	26.40	23.57	25.93
	Vision ↦ Text	85.00	41.41	25.93
EchoInk	Audio ↦ Text	35.00	48.48	30.98
	Audio ↦ Vision	30.98	23.57	23.91
	Text ↦ Audio	68.69	21.89	25.25
	Text ↦ Vision	43.10	22.56	27.61
	Vision ↦ Audio	31.99	23.57	25.25
	Vision ↦ Text	93.60	46.80	44.44
Baichuan Omni 1.5	Audio ↦ Text	23.00	34.40	23.60
	Audio ↦ Vision	23.80	25.74	25.99
	Text ↦ Audio	23.40	23.00	25.40
	Text ↦ Vision	25.80	26.23	28.77
	Vision ↦ Audio	25.20	28.34	21.47
	Vision ↦ Text	70.20	53.18	33.13
VideoLLaMA 2	Audio ↦ Text	25.82	35.90	31.23
	Audio ↦ Vision	25.23	25.80	24.03
	Text ↦ Audio	34.29	22.09	26.70
	Text ↦ Vision	26.66	26.06	24.90
	Vision ↦ Audio	27.03	23.64	26.67
	Vision ↦ Text	50.40	32.44	36.67
VITA	Audio ↦ Text	26.26	38.14	28.87
	Audio ↦ Vision	16.49	31.17	27.52
	Text ↦ Audio	26.80	27.61	23.91
	Text ↦ Vision	22.68	25.62	25.58
	Vision ↦ Audio	28.62	26.71	27.59
	Vision ↦ Text	42.27	49.66	33.10
Unified IO 2	Audio ↦ Text	24.28	18.25	25.44
	Audio ↦ Vision	21.50	22.61	23.03
	Text ↦ Audio	30.02	23.46	21.89
	Text ↦ Vision	25.25	24.85	22.86
	Vision ↦ Audio	25.46	26.29	25.65
	Vision ↦ Text	27.68	16.25	23.37
Unified IO 2 XL	Audio ↦ Text	24.65	24.63	17.47
	Audio ↦ Vision	26.03	30.21	17.39
	Text ↦ Audio	27.52	28.83	30.02
	Text ↦ Vision	25.09	19.16	22.17
	Vision ↦ Audio	22.64	24.92	30.57

Continued on next page

Table 8 – continued from previous page

Model		Temporal Task		
Model	**Task**	**Order**	**Counting**	**Calculation**
	Vision ↦ Text	37.30	36.42	24.44
Unified IO 2 XXL	Audio ↦ Text	24.41	26.81	21.62
	Audio ↦ Vision	25.26	29.68	27.17
	Text ↦ Audio	28.83	22.43	24.61
	Text ↦ Vision	23.70	37.78	27.37
	Vision ↦ Audio	23.63	24.69	27.28
	Vision ↦ Text	41.69	38.50	22.95
Panda	Audio ↦ Text	25.85	16.77	23.17
	Audio ↦ Vision	26.06	22.60	27.31
	Text ↦ Audio	25.72	22.81	25.80
	Text ↦ Vision	26.31	22.77	29.02
	Vision ↦ Audio	26.10	22.77	24.59
	Vision ↦ Text	25.51	22.94	23.37

C.4 Task 4: Linguistic Task

Table 9: T4 Linguistic Task Results

Model		Linguistic Task		
Model	**Task**	**Recognition**	**Translation**	**Emotion**
Gemini 2.5 Pro	Audio ↦ Text	97.16	96.58	60.86
	Audio ↦ Vision	91.65	67.95	42.75
	Text ↦ Audio	80.35	81.62	64.51
	Text ↦ Vision	93.58	67.38	67.31
	Vision ↦ Audio	80.81	73.22	43.43
	Vision ↦ Text	99.54	100.00	74.54
Gemini 2.5 Flash	Audio ↦ Text	94.05	97.44	60.86
	Audio ↦ Vision	68.01	93.30	43.67
	Text ↦ Audio	76.92	81.34	54.43
	Text ↦ Vision	72.88	67.24	60.14
	Vision ↦ Audio	74.95	72.93	45.22
	Vision ↦ Text	99.40	96.72	57.14
Gemini 2.0 Flash	Audio ↦ Text	92.86	97.29	60.57
	Audio ↦ Vision	68.30	67.66	36.43
	Text ↦ Audio	69.79	81.20	54.86
	Text ↦ Vision	73.92	67.66	60.43
	Vision ↦ Audio	66.52	73.08	43.00
	Vision ↦ Text	96.43	97.15	56.71
Gemini 1.5 Pro	Audio ↦ Text	94.94	97.15	60.43
	Audio ↦ Vision	73.96	46.72	36.57
	Text ↦ Audio	83.33	80.91	52.57
	Text ↦ Vision	76.93	66.81	62.43
	Vision ↦ Audio	80.80	92.02	39.20
	Vision ↦ Text	96.73	96.44	63.29
Qwen2.5 Omni	Audio ↦ Text	94.64	96.72	65.29
	Audio ↦ Vision	62.95	73.36	48.94
	Text ↦ Audio	81.25	86.75	52.71
	Text ↦ Vision	65.03	69.09	59.79
	Vision ↦ Audio	82.44	88.60	43.57
	Vision ↦ Text	97.17	97.72	67.57
EchoInk	Audio ↦ Text	92.93	95.96	69.02
	Audio ↦ Vision	64.98	71.38	35.69
	Text ↦ Audio	80.47	81.48	61.95

Continued on next page

Table 9 – continued from previous page

Model		Linguistic Task		
Model	**Task**	**Recognition**	**Translation**	**Emotion**
	Text ↦ Vision	68.35	67.68	57.24
	Vision ↦ Audio	81.48	85.86	43.10
	Vision ↦ Text	96.63	97.31	68.01
Baichuan Omni 1.5	Audio ↦ Text	87.05	96.01	48.00
	Audio ↦ Vision	55.36	56.55	25.25
	Text ↦ Audio	64.29	84.94	48.29
	Text ↦ Vision	55.95	52.56	46.99
	Vision ↦ Audio	65.03	84.06	27.14
	Vision ↦ Text	92.56	96.72	43.43
VideoLLaMA 2	Audio ↦ Text	69.04	67.40	14.48
	Audio ↦ Vision	24.82	26.00	24.68
	Text ↦ Audio	22.82	22.02	27.68
	Text ↦ Vision	25.03	25.80	24.65
	Vision ↦ Audio	24.07	23.25	25.01
	Vision ↦ Text	83.86	86.80	43.00
VITA	Audio ↦ Text	39.18	73.20	44.33
	Audio ↦ Vision	24.74	24.74	30.93
	Text ↦ Audio	39.73	55.56	46.13
	Text ↦ Vision	30.93	25.77	32.99
	Vision ↦ Audio	53.87	61.95	27.95
	Vision ↦ Text	86.60	88.66	42.27
Unified IO 2	Audio ↦ Text	62.01	14.06	21.05
	Audio ↦ Vision	35.66	20.90	25.83
	Text ↦ Audio	26.60	26.36	29.85
	Text ↦ Vision	25.89	26.00	31.82
	Vision ↦ Audio	24.24	25.14	26.27
	Vision ↦ Text	66.06	18.90	40.01
Unified IO 2 XL	Audio ↦ Text	69.63	17.26	16.29
	Audio ↦ Vision	45.46	26.28	23.28
	Text ↦ Audio	27.82	23.75	21.90
	Text ↦ Vision	30.65	25.26	30.47
	Vision ↦ Audio	25.07	23.70	21.88
	Vision ↦ Text	75.27	23.23	27.02
Unified IO 2 XXL	Audio ↦ Text	72.67	17.63	29.46
	Audio ↦ Vision	18.23	27.43	23.24
	Text ↦ Audio	23.04	25.97	27.47
	Text ↦ Vision	31.09	27.84	31.42
	Vision ↦ Audio	19.43	23.31	24.06
	Vision ↦ Text	78.04	26.88	34.88
PandaGPT	Audio ↦ Text	27.12	28.83	28.03
	Audio ↦ Vision	22.38	22.23	28.23
	Text ↦ Audio	22.06	18.88	21.20
	Text ↦ Vision	22.10	24.96	27.04
	Vision ↦ Audio	22.55	22.40	27.99
	Vision ↦ Text	33.96	32.67	27.20

C.5 Task 5: External Knowledge

Table 10: T5 (External) Task Results

Model		External Task		
Model	**Task**	**Genre**	**Movie**	**Singer**
Gemini 2.5 Pro	Audio ↦ Text	83.28	93.00	94.67
	Audio ↦ Vision	74.80	89.90	82.67
	Text ↦ Audio	78.16	94.50	91.95
	Text ↦ Vision	85.76	97.99	100.00
	Vision ↦ Audio	72.42	92.00	90.00
	Vision ↦ Text	88.95	96.45	100.00
Gemini 2.5 Flash	Audio ↦ Text	83.78	93.00	69.13
	Audio ↦ Vision	63.36	82.41	70.92
	Text ↦ Audio	78.56	90.45	76.00
	Text ↦ Vision	85.00	97.99	98.67
	Vision ↦ Audio	63.96	88.32	71.33
	Vision ↦ Text	86.34	98.00	93.71
Gemini 2.0 Flash	Audio ↦ Text	83.50	88.00	72.00
	Audio ↦ Vision	62.40	86.50	56.00
	Text ↦ Audio	78.46	82.50	50.67
	Text ↦ Vision	84.50	98.00	96.67
	Vision ↦ Audio	66.43	79.50	38.00
	Vision ↦ Text	87.50	95.00	100.00
Gemini 1.5 Pro	Audio ↦ Text	61.70	78.00	47.33
	Audio ↦ Vision	42.90	74.50	40.00
	Text ↦ Audio	63.53	84.50	62.67
	Text ↦ Vision	82.10	95.50	88.67
	Vision ↦ Audio	45.59	74.00	37.33
	Vision ↦ Text	87.10	95.00	88.67
Qwen2.5 Omni	Audio ↦ Text	89.50	79.50	80.00
	Audio ↦ Vision	61.40	67.50	48.67
	Text ↦ Audio	85.65	70.50	56.00
	Text ↦ Vision	74.20	94.50	78.67
	Vision ↦ Audio	81.82	60.50	33.33
	Vision ↦ Text	79.00	92.50	78.00
EchoInk	Audio ↦ Text	87.54	82.50	80.00
	Audio ↦ Vision	61.95	68.00	51.33
	Text ↦ Audio	84.51	73.00	60.67
	Text ↦ Vision	77.78	93.00	80.00
	Vision ↦ Audio	62.63	64.50	42.67
	Vision ↦ Text	79.12	93.50	77.33
Baichuan Omni 1.5	Audio ↦ Text	65.60	56.00	41.33
	Audio ↦ Vision	45.30	54.50	32.00
	Text ↦ Audio	25.75	60.00	40.00
	Text ↦ Vision	77.00	94.50	77.33
	Vision ↦ Audio	27.15	46.50	27.33
	Vision ↦ Text	81.20	94.50	74.00
VideoLLaMA 2	Audio ↦ Text	62.60	26.59	39.38
	Audio ↦ Vision	26.23	23.56	26.72
	Text ↦ Audio	24.85	21.59	25.34
	Text ↦ Vision	26.40	28.55	26.09
	Vision ↦ Audio	25.67	23.56	24.10
	Vision ↦ Text	68.27	80.55	82.02
VITA	Audio ↦ Text	46.39	40.21	28.87
	Audio ↦ Vision	20.62	26.80	24.74
	Text ↦ Audio	21.89	25.50	25.33
	Text ↦ Vision	20.62	31.96	21.65
	Vision ↦ Audio	23.23	22.00	22.67
	Vision ↦ Text	47.42	81.44	54.64

Continued on next page

Table 10 – continued from previous page

Model		External Task		
Model	**Task**	**Genre**	**Movie**	**Singer**
Unified IO 2	Audio ↦ Text	31.83	22.53	30.09
	Audio ↦ Vision	22.30	29.03	21.40
	Text ↦ Audio	26.25	24.51	26.71
	Text ↦ Vision	34.46	26.03	20.71
	Vision ↦ Audio	25.45	20.57	30.01
	Vision ↦ Text	27.90	34.59	27.33
Unified IO 2 XL	Audio ↦ Text	36.80	27.52	31.40
	Audio ↦ Vision	29.23	29.09	25.40
	Text ↦ Audio	24.12	25.09	29.34
	Text ↦ Vision	34.41	24.57	26.68
	Vision ↦ Audio	26.51	32.55	29.41
	Vision ↦ Text	24.86	35.76	38.05
Unified IO 2 XXL	Audio ↦ Text	57.68	22.71	34.70
	Audio ↦ Vision	26.83	20.56	24.42
	Text ↦ Audio	47.92	26.52	31.43
	Text ↦ Vision	51.85	24.01	42.75
	Vision ↦ Audio	25.06	30.57	33.40
	Vision ↦ Text	28.20	36.55	47.36
Panda	Audio ↦ Text	25.77	21.32	21.24
	Audio ↦ Vision	25.74	24.49	21.42
	Text ↦ Audio	22.11	25.18	20.37
	Text ↦ Vision	24.63	24.64	21.40
	Vision ↦ Audio	23.93	24.58	21.29
	Vision ↦ Text	26.32	21.07	20.39

C.6 Evaluation Cost

We provide a detailed evaluation cost section as a reference of usage. We evaluate on the full version (60k sample) of XModBench, API-based models we test **Gemini 2.5 Pro**, we report the *token usage* for evaluating the overall benchmark and each task family . For open-source models we report **Qwen2.5-Omni**, we report the *evaluation runing time*, using with eight `A6000` GPUs and each GPU run one process.

Table 11: Evaluation cost estimation for models across the five task families and the full benchmark.

Model	**Perc.**	**Spat.**	**Temp.**	**Ling.**	**Knwl.**	**Total**
Gemini 2.5 Pro (*Token usage*)	26.0M	13.5M	25.1M	4.3M	14.0M	82.9M
Qwen2.5-Omni (*Hours*)	6.3	1.4	1.4	1.4	2.1	12.7

D Interleaving Visual Audio Input

In the preceding experiments, we showed that omni-language models exhibit varying performance in pairwise cross-modal reasoning, particularly between vision–text and audio–text tasks. Yet, real-world multimodal scenarios are more complex: information from multiple modalities often arrives simultaneously and must be processed in an integrated manner. To address this challenge, we extend all tasks in XModBench to an audio–visual context configuration, where the question stem provides both audio and visual cues, while the candidate space remains identical to the original text-based setting.

We evaluate this dual-context setup using the Gemini series of models, which represent some of the most advanced omni-language systems available. The results, presented in Tab. 12, enable a direct comparison with the pairwise baseline and reveal how models leverage—or fail to leverage—simultaneous multimodal evidence.

Table 12: Overall performance of Gemini models under the dual-context setting (audio+visual context $\mapsto$ text). We compare with pairwise baselines (A $\mapsto$ T and V $\mapsto$ T), and report the stronger unimodal baseline $\max(\text{A} \mapsto \text{T}, \text{V} \mapsto \text{T})$.

Setting	Gemini 1.5 Pro	Gemini 2.0 Flash	Gemini 2.5 Pro
A $\mapsto$ T	52.76	63.71	70.99
V $\mapsto$ T	79.92	85.20	88.60
A+V $\mapsto$ T	82.53 (**+2.61**)	79.84	89.76

E Human Survey

To evaluate human performance and establish reference baselines, we conducted a user study on a subset of **XModBench**. Participants answered multiple-choice questions under different modality configurations, with Figure 7 showing a screenshot of the interface and example questions. For each subtask, we collected responses from 10 valid participants per modality configuration.

F Techiniqal Details in Triplet Data collection and Processing.Data for Each Subtask

In this section, we provide detailed descriptions of the data sources are collected, and how each data in each modality are processed for each subtask in XModBench.

F.1 Perceptual Recognition

General Categories. We utilize the VGGSound Source (VGG-SS) dataset(Chen et al., 2021; Kim et al., 2024), a large-scale video benchmark designed for sound source localization, which provides video-level annotations across diverse sound activities. The dataset covers 200 categories with approximately 5,000 video clips, where sound sources are annotated with bounding boxes to ensure clear visibility in each clip. For our benchmark, we extract a 2-second segment corresponding to the loudest audio channel as the audio input, and randomly sample a single frame from the same clip as the visual input. The activity class name serves as the textual description. To construct multiple-choice questions, four additional activity labels are randomly sampled as distractors, resulting in four candidate answers per instance. We then use Gemini 2.5-flash lite to(Comanici et al., 2025) filter if each instance if the audio and video frame is clear to be hear and the image frame and audio are all match the category name.

Fine-grained Categories. This subtask uses the same pool of video clips as the General Categories setting. The difference lies in reorganizing the activity classes into eight fine-grained clusters: *Animal sounds*, *Musical instruments*, *Human activities*, *Transportation*, *Tools and utilities*, *Urban sounds*, *Human speech*, and *Natural sounds*. For each instance, we select the target activity along with four distractor activities sampled from the same fine-grained cluster. This ensures that all answer choices belong to the same semantic domain, making the recognition task more challenging and diagnostic within a coherent category group.

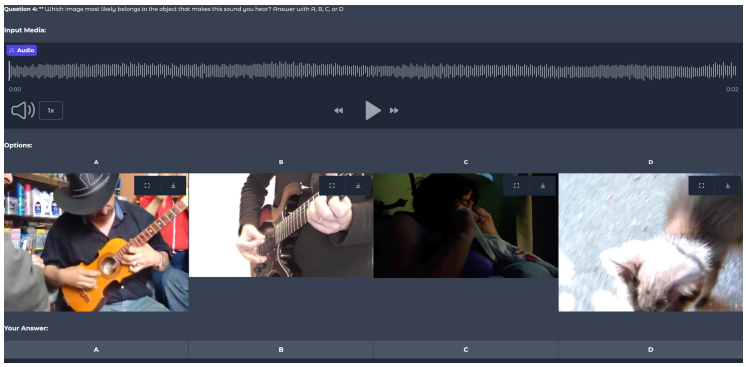

Figure 7: Sample question of human survey

Natural Environment. We draw data from the Landscapes dataset(Lee et al., 2022), which consists of ambient audio–video clips capturing natural outdoor scenes. Following the same selection protocol as in the General Categories task, we extract a 2-second segment from the dominant audio channel as the audio input, and randomly sample one frame from the corresponding video as the visual input. The dataset's categorical labels are used as the textual descriptions.

Instruments. Instrument data is collected from the Solos dataset(Montesinos et al., 2020), which contains recordings of 13 distinct instruments: violin, viola, cello, double bass, flute, oboe, clarinet, bassoon, saxophone, trumpet, horn, trombone, and tuba. We use the video frames as the visual modality, the isolated performance recordings as the audio modality, and the instrument names as textual labels.

Instrument Composition. We employ the URMP dataset(Li et al., 2018), a multimodal corpus designed for music performance analysis, which provides video and audio recordings of ensemble performances. For this subtask, we leverage clips containing multiple instruments playing together, using the mixture audio as input, sampled video frames as the visual modality, and instrument combination labels as text.

F.2 Spatial Reasoning

2D Horizontal Arrangement. This subtask is derived from the URMP dataset(Li et al., 2018), which contains multi-instrument ensemble recordings with annotated left-to-right spatial positions of each performer and independent audio channels per instrument. We construct multiple-choice questions by generating three distractor options through random shuffling of instrument order along the horizontal axis. For the visual modality, cropped player images are concatenated into a composite frame that preserves their spatial arrangement. For the audio modality, stereo spatialization is synthesized by assigning distinct azimuth values to each shuffled configuration and adjusting the relative channel balance using a panning algorithm (e.g., vector-base amplitude panning(Pulkki, 1997)). This design ensures that listeners can clearly perceive the relative horizontal positions of the instruments.

3D Localization. This subtask builds on the STARSS23 dataset(Shimada et al., 2023), which provides panoramic video with time-stamped annotations of sound source depth, azimuth, and activity. For the visual modality, we annotate sound sources with bounding boxes and generate alternative views by rotating the camera perspective to $+90^\circ$, 180°, and -90° (positive defined as left). The corresponding videos are created through spatial cropping of frames. For the audio modality, we utilize the four-channel microphone array (MIC) recordings and simulate azimuthal rotation by first encoding the array signals into first-order Ambisonics (FOA), applying a 2D rotation matrix to the X–Y components, and decoding back into microphone signals with loudness normalization. To further enhance perceptual realism, each spatial microphone signal is additionally processed with head-related transfer functions (HRTFs) in the SOFA format(Majdak et al., 2013; Algazi et al., 2001).

3D Movements. This subtask is based on the Urbansas dataset(Fuentes et al., 2022), which provides street-view traffic videos with detailed audio annotations indicating vehicle types and the presence of off-screen sounds. Each clip includes labels specifying the vehicle category, whether the sound source is visible in the video, and its temporal activity. We curate video segments from this dataset and highlight the target vehicle by overlaying a red bounding box to establish clear audio–visual correspondence.

F.3 Temporal Reasoning

Event Order. This subtask is derived from 2-second video clips in the VGGSound Source (VGG-SS) dataset(Chen et al., 2021), originally used in the Perceptual Recognition task where each clip is annotated with an activity class label. For temporal ordering, we randomly sample 3–5 clips from different classes and generate four candidate event sequences by shuffling their order. Each sequence is represented across three modalities: (i) a text description (e.g., "Event A $\rightarrow$ Event B $\rightarrow$ Event C"), (ii) a concatenated video sequence, and (iii) a concatenated audio sequence. Multiple-choice questions are formed by selecting one sequence as the correct answer and presenting the stem in one modality, while the four candidate sequences are given in another modality.

Repetition Count. Following the setup in(Zhang et al., 2021), this subtask focuses on counting repeated events. Visual data is generated from synthetic renderings of repeated object actions, while audio data consists of temporal patterns with clear repetitions (e.g., sequences of knocks or claps). Text prompts explicitly query the number of repetitions in either modality.

Repetition Calculation. Also inspired by(Zhang et al., 2021), this subtask extends beyond direct counting by requiring simple arithmetic over observed repetitions. Both audio and video are rendered with variable fre-

quencies of repeated events, while the text prompts encode arithmetic formulations that ask models to compute totals (e.g., "three knocks plus two knocks").

F.4 Linguistic Understanding

Linguistic Recognition. This subtask targets recognition of textual content across modalities. Images are collected from OCR-rendered text data(Wendler, 2024), each paired with its ground-truth transcript. Audio is generated from these transcripts using a TTS system(Guo et al., 2025), allowing for cross-modal recognition between text, vision, and speech.

Translation. This subtask examines cross-lingual translation. Input sequences consist of English text with multiple-choice options in Chinese. Text data is derived from OCR-rendered images(Wendler, 2024), while translations are generated using Gemini(Team et al., 2024). Visual inputs are rendered using the OCR dataset rendering toolkit(GbotHQ, 2024), and audio is synthesized from both languages with a TTS system(Guo et al., 2025).

Dialogue Emotion. This subtask focuses on multimodal emotion recognition in conversational settings. Visual data consists of face videos displaying emotional expressions extracted from multi-party dialogue clips(Chen et al., 2018; Poria et al., 2019). Each dialogue is paired with transcripts and annotated with categorical emotions (anger, disgust, fear, sadness, surprise, and joy). We filter clips to lengths between 5–30 seconds. The video data is stripped of original audio but accompanied by transcripts to enable inference of emotion from dialogue and facial expression. Audio inputs consist of the original speech tracks, and text inputs are provided as the emotion category names.

F.5 External Knowledge

Music Genre Classification. This subtask evaluates music genre recognition. We collect audio samples from the GTZAN dataset(Olteanu, 2024), covering multiple musical styles. To complement the audio, we also collect representative album cover images for each genre category.

Movie Matching. This subtask requires linking multimodal cues to movie identities. We collect a set of recent films from IMDb. For the visual modality, we use official posters. To prevent trivial text matching between posters and movie titles, we use written plot summaries from IMDb as the text modality. Audio is sampled as 30-second clips from publicly available trailers on YouTube.

Singer Identification. This subtask targets cross-modal recognition of popular singers. Images of singers are collected from the web, while audio consists of short clips (3–5 songs each) sampled from their publicly available music videos on YouTube. Text inputs include singer names and associated biographical metadata. We select a diverse set of internationally recognized artists, including American singers Ariana Grande, Bad Bunny, Billie Eilish, Bruno Mars, Chappell Roan, Harry Styles, and Chinese singers David Tao, Eason Chan, Faye Wong, G.E.M., and Jay Chou.

G LLM Usage

We used large language models (LLMs) to assist in the preparation of this paper. Their role was limited to language editing such as proofreading and rephrasing. All ideas, experiments, and analyses were conceived and conducted by the authors.