OmniBench: Towards The Future of Universal Omni-Language Models

Yizhi Li[*, 1, 2] **Ge Zhang**[*, 1, 3] **Yinghao Ma**[*, 1, 4] **Ruibin Yuan**[1, 5] **Kang Zhu**[1, 3] **Hangyu Guo**[1] **Yiming Liang**[1] **Jiaheng Liu**[1, 6] **Zekun Wang**[1, 3] **Jian Yang**[1] **Siwei Wu**[1, 2] **Xingwei Qu**[1, 2] **Jinjie Shi**[4] **Xinyue Zhang**[1] **Zhenzhu Yang**[1] **Xiangzhou Wang**[1] **Zhaoxiang Zhang**[6] **Zachary Liu**[7] **Emmanouil Benetos**[4] **Wenhao Huang**[1, 3] **Chenghua Lin**[1, 2]

Abstract

Recent advancements in multimodal large language models (MLLMs) have focused on integrating multiple modalities, yet their ability to simultaneously process and reason across different inputs remains underexplored. We introduce **OmniBench**, a novel benchmark designed to evaluate models' ability to recognize, interpret, and reason across **visual**, **acoustic**, and **textual** inputs simultaneously. We define language models capable of such tri-modal processing as omni-language models (OLMs). OmniBench features high-quality human annotations that require integrated understanding across all modalities. Our evaluation reveals that: *i)* open-source OLMs show significant limitations in instruction-following and reasoning in tri-modal contexts; and *ii)* most baseline models perform poorly (around 50% accuracy) even with textual alternatives to image/audio inputs. To address these limitations, we develop **OmniInstruct**, an 96K-sample instruction tuning dataset for training OLMs. We advocate for developing more robust tri-modal integration techniques and training strategies to enhance OLM performance. Codes and data could be found at our repo.

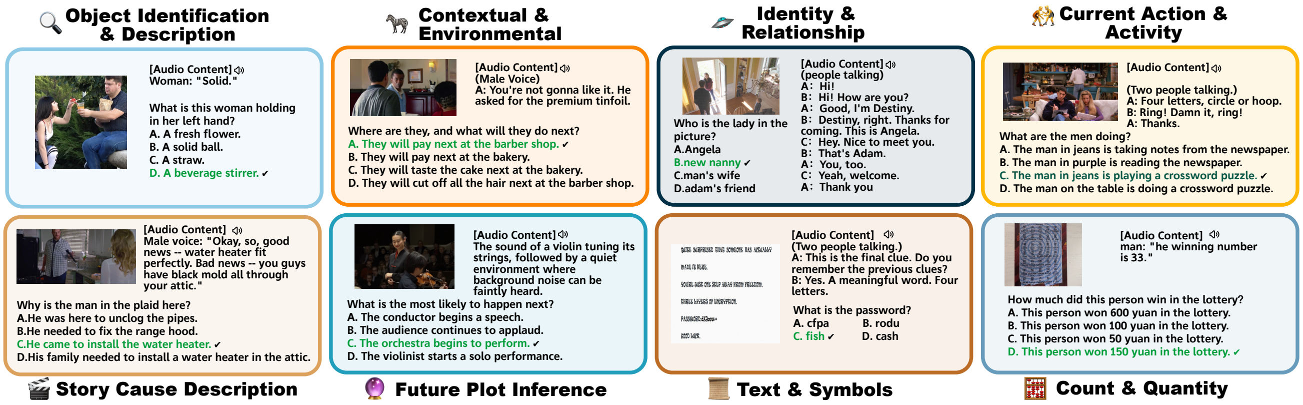
1. Introduction

The rapid advancement of artificial intelligence has ushered in a new era of multimodal large language models (MLLMs), capable of processing and interpreting diverse data types mainly involving images, audio, and text (Li & Lu, 2024). These models aim to emulate human-like understanding of the world by integrating information across multiple sensory modalities and learning a comprehensive context from the environment. While significant strides have been made in developing MLLMs handling two of the modalities, the ability to concurrently process and reason about the three modalities remains a frontier yet to be fully explored.

The social impact of these MLLMs is far-reaching, providing transformative capabilities for a variety of domains. In healthcare, VLMs and ALMs have contributed to diagnosing (Liu et al., 2023a; Hemdan et al., 2023), and potentially combining *three modalities* (Meskó, 2023). The integration of all vision, audio and text modalities is expected to significantly improve diagnostic accuracy and patient interaction, making healthcare more accessible and efficient. In urban environments, ALM can contribute to improving safety and traffic management by incorporating urban sound event detection during autonomous driving, such as recognizing audio of emergency vehicles and recognize their types or location with supplementary visual modality (Sun et al., 2021). In addition, audio contributes to biodiversity monitoring (Terenzi et al., 2021; Liang et al., 2024a) and can be greatly enhanced by MLLM's ability to analyse both audio and video from a variety of sensors. Finally, it may help robotics or LLM agents to provide better human-computer/robotic interaction (HCI/HRI) service in day-to-day life (Liang et al., 2024b; Su et al., 2023).

The challenge in advancing MLLMs lies not only in their development but also in our capacity to evaluate their performance comprehensively. Current benchmarks often solely focus on image or audios, or limited image-text (Yue et al., 2024; Zhang et al., 2024) or audio-text combinations (Yang et al., 2024) for the dual-modality vision-language models (VLMs) (Laurençon et al., 2024) or audio-language models (ALMs) (Chu et al., 2023a; Deng et al., 2023). This gap in evaluation tools has hindered the community to assess and improve the holistic capabilities of models right before the dawn of general-purpose MLLMs.

[*] Equal contribution [1] M-A-P [2] University of Manchester [3] 01.ai [4] Queen Mary University of London [5] Hong Kong University of Science and Technology [6] Nanjing University [7] Dartmouth College. Correspondence to: Ge Zhang <ge.zhang@01.ai>, Chenghua <c.lin@manchester.ac.uk>.


Object Identification & Description
Contextual & Environmental
Identity & Relationship
Current Action & Activity
QUITE SURPRISED THAT SOMEONE HAS ACTUALLY
MADE IT HERE.
YOU'RE JUST ONE STEP AWAY FROM FREEDOM.
THREE LAYERS OF ENCRYPTION.
PASSWORD:dXBocw==
GOOD LUCK.
Story Cause Description
Future Plot Inference
Text & Symbols
Count & Quantity

[Audio Content]
Woman: "Solid."

What is this woman holding in her left hand?
A. A fresh flower.
B. A solid ball.
C. A straw.
D. A beverage stirrer. ✓

[Audio Content]
(Male Voice)
A: You're not gonna like it. He asked for the premium tinfoil.

Where are they, and what will they do next?
A. They will pay next at the barber shop. ✓
B. They will pay next at the bakery.
C. They will taste the cake next at the bakery.
D. They will cut off all the hair next at the barber shop.

[Audio Content]
(people talking)
A: Hi!
B: Hi! How are you?
A: Good, I'm Destiny.
B: Destiny, right. Thanks for coming. This is Angela.
C: Hey. Nice to meet you.
B: That's Adam.
A: You, too.
C: Yeah, welcome.
A: Thank you

Who is the lady in the picture?
A.Angela
B.new nanny ✓
C.man's wife
D.adam's friend

[Audio Content]

(Two people talking.)
A: Four letters, circle or hoop.
B: Ring! Damn it, ring!
A: Thanks.

What are the men doing?
A. The man in jeans is taking notes from the newspaper.
B. The man in purple is reading the newspaper.
C. The man in jeans is playing a crossword puzzle. ✓
D. The man on the table is doing a crossword puzzle.

[Audio Content]
Male voice: "Okay, so, good news -- water heater fit perfectly. Bad news -- you guys have black mold all through your attic."

Why is the man in the plaid here?
A.He was here to unclog the pipes.
B.He needed to fix the range hood.
C.He came to install the water heater. ✓
D.His family needed to install a water heater in the attic.

[Audio Content]
The sound of a violin tuning its strings, followed by a quiet environment where background noise can be faintly heard.

What is the most likely to happen next?
A. The conductor begins a speech.
B. The audience continues to applaud.
C. The orchestra begins to perform. ✓
D. The violinist starts a solo performance.

[Audio Content]
(Two people talking.)
A: This is the final clue. Do you remember the previous clues?
B: Yes. A meaningful word. Four letters.

What is the password?
A. cfpa B. rodu
C. fish ✓ D. cash

[Audio Content]
man: "he winning number is 33."

How much did this person win in the lottery?
A. This person won 600 yuan in the lottery.
B. This person won 100 yuan in the lottery.
C. This person won 50 yuan in the lottery.
D. This person won 150 yuan in the lottery. ✓

Figure 1. Example Data from Different Categories. The main contextual information is provided by the corresponding image and audio, while the question and options are expressed with text. Playable audio demos are available at the demo page.

To address this critical need, we introduce **OmniBench**, a pioneering universal multimodal benchmark designed to rigorously evaluate MLLMs' capability to recognize, interpret, and reason across visual, acoustic, and textual inputs *simultaneously*, which we define as the *omni-understanding and reasoning* ability of the omni-language models (**OLM**s) (Sun et al., 2024; Zhan et al., 2024; Lu et al., 2024b). For instance, one can only derive the correct answer of the sample question in Figure 1 by: *1*) recognizing elements from the given image and audio to reconstruct the context; *2*) interpreting the semantics and relationships among the multimodal objects according to the textual instruction formed as question and options; *3*) reasoning and then answering with the complementary information from all the modalities. We distinguishes OmniBench by enforcing a unique constraint: accurate responses necessitate an integrated understanding and reasoning of *all multimodal contexts*. This approach ensures a more realistic and challenging assessment of multimodal large language models, mirroring the complex, interconnected nature of human cognition. To ensure the evaluation reliability, the development of OmniBench relies on high-quality human annotations. Furthermore, OmniBench additionally includes the answer rationales provided by the annotators to enhance the validity and ensure the benchmark aligned with human-level understanding.

Our initial findings using OmniBench reveal critical limitations in the omni-understanding capabilities of existing MLLMs:

- Although the existing open-source omni-language models have been trained with data in the three modalities, most of them can surpass the performance of random guess accuracy but sometimes hard to follow the instruction when provided with image and audio together in certain cases.
- In contrast, the proprietary models perform better overall but suffer from more accuracy drops when ablating the image or audio input.
- Compared to models, the results of human evaluator show not only better overall performances but a distinct distribution (*e.g.*, much better on ”Sound Event” audios and ”Abstract Concept” tasks).
- In the context of using text as an alternative source of information to corresponding audio and images, the open-source VLMs and ALMs show relatively better results but remain in a preliminary level of capability to understand the given tri-modality context.

In the following sections, we 1) detail the data collection protocol of OmniBench; 2) present our evaluation results on current state-of-the-art MLLMs; 3) introduce the **Omni-Instruct** dataset for omni-language model supervised fine-tuning; and 4) discuss the implications of our findings for the future of research and development.

2. Related Work

Multimodal Large Language Models. Recent advances in multimodal large language models have produced specialized encoders for audio processing (Radford et al., 2022; Chen et al., 2022; Li et al., 2023b; Wu et al., 2023b), which have been integrated into more comprehensive systems (Tang et al., 2023; Gong et al., 2023). Notable progress in audio-focused dialogues has demonstrated promising capabilities in instruction-following and perception (Tang et al., 2023; Wang et al., 2023a; Wu et al., 2023a; Chu et al., 2023b). In the visual domain, significant strides have been made through models that leverage pre-trained image encoders (Dosovitskiy, 2020; Touvron et al., 2020; Liu et al., 2021; Radford et al., 2021; Zhai et al., 2023). These models have achieved success through visual-textual alignment, GPT-4 generated instruction data, and extensive pre-training (Li et al., 2023a; Liu et al., 2024b;a; Bai et al., 2023; Wang

Table 1. The Statistics of OmniBench Across Task Types. The word lengths of four options for each question are first averaged, and then the averages are calculated in group.

Statistic	Causal Inference			(Temporal-)Spatial Entity		Abstract Concept			
Sub-class of QA	**Current Action & Activity**	**Story Description**	**Plot Inference**	**Object Identification & Description**	**Contextual & Environmental**	**Identity & Relationship**	**Text & Symbols**	**Count & Quantity**	**Overall**
					Quantity				
Total	251	230	237	211	141	32	25	15	1142
Speech	78	182	179	162	104	31	22	13	771
Sound Event	147	27	37	28	25	1	-	-	265
Music	26	21	21	21	12	-	3	2	106
					Word Length				
Question	4.68	5.75	7.47	7.00	6.85	6.22	7.32	8.72	6.25
Option	8.85	7.77	8.92	6.47	5.68	10.38	11.22	6.60	8.81
Img. Rationale	18.27	19.62	24.40	24.94	18.34	22.69	24.80	29.16	21.19
Audio Rationale	23.11	20.50	24.40	20.97	18.27	24.92	23.10	53.84	22.90
Audio Content	13.21	17.91	29.87	28.03	14.41	19.01	23.31	35.16	18.37
					Multimodal Info.				
Img. Width	1283.75	1291.60	2394.93	1430.03	1141.39	1395.53	1338.51	1787.36	1322.36
Img. Height	842.32	776.11	2089.93	799.47	728.06	840.15	761.58	1168.04	818.64
Audio Len. (s)	7.35	9.82	11.22	11.43	8.03	8.63	11.43	15.63	9.22

et al., 2023b; Young et al., 2024). While most existing MLLMs focus on single-modality input processing, open-source models capable of processing multiple modalities generally show less competitive capabilities compared to closed-source alternatives. In this context, we define omni-language models (OLMs) as those capable of processing at least three different modalities simultaneously[1].

Multimodal Understanding Benchmark. Current vision-language benchmarks evaluate various capabilities including OCR, spatial awareness, multimodal information retrieval, and reasoning skills (Wu et al., 2024a; Yu et al., 2023; Liu et al., 2023c; Chen et al., 2024a; Yue et al., 2024; Zhang et al., 2024; Wu et al., 2024b). These benchmarks assess models through multiple-choice tasks, complex vision-language problems, and multi-image relational understanding. In the audio domain, several benchmarks focus on speech recognition, audio QA tasks, and sound classification (Bu et al., 2017; Du et al., 2018; Panayotov et al., 2015; Drossos et al., 2020; Gong et al., 2022). However, there remains a significant gap in comprehensive benchmarks that can assess models' ability to simultaneously process and integrate information from textual, audio, and visual inputs.

Audio-Visual Understanding Datasets. Existing audio-visual question answering datasets have primarily focused on object and sound identification (Yun et al., 2021; Li et al., 2022; Liu et al., 2024c; Yang et al., 2022). While these datasets cover various aspects like temporal understanding and counting, they often lack comprehensive evaluation of causal inference and abstract reasoning. Some datasets don't require true multimodal integration for answering questions, as responses can often be deduced from a single modality. Recent work has attempted to address these limitations through human annotation and improved alignment between modalities (Chen et al., 2023; Gemmeke et al., 2017), but still lacks instruction-following evaluation capability. These limitations highlight the need for more comprehensive datasets that can effectively assess models' multimodal integration and reasoning abilities.

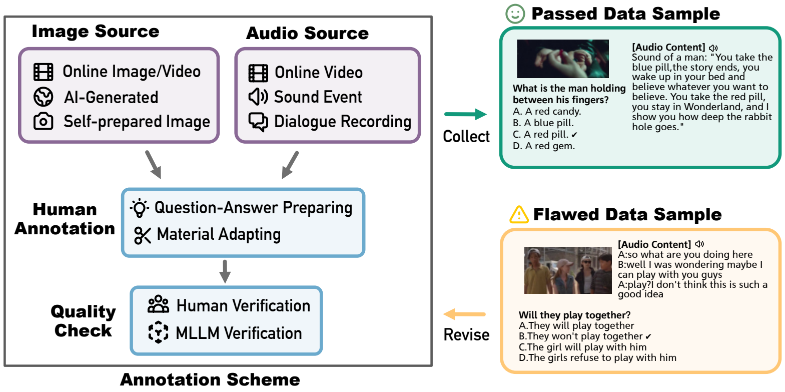
3. OmniBench

The OmniBench aims to create a comprehensive benchmark for evaluating multimodal large language models that support simultaneous image, audio, and text inputs. While OmniBench is designed to evaluate the understanding capability of MLLMs on cross-modality complementary information, the models are required to interpret the multimodal input and provide accurate text answer. The problem could be formulated as following: given a tuple of (image, audio, text), the model is required to recognize the objects, re-build the contexts, and conduct reasoning based on the given information. The design logic and statistics of the dataset and the annotation protocols are introduced in this section.

3.1. Benchmark Design

Building on the foundation of existing multimodal benchmarks, our OmniBench introduces a refined taxonomy for task categorization that effectively captures a wide range of cognitive and reasoning abilities. Our framework organizes tasks into three primary categories: (1) *(temporal)-spatial entity*, which includes *Object Identification* for recognizing distinct entities and *Contextual & Environmental* for discerning the setting or backdrop of the events; (2) *causal inference*, comprised of *Story Cause Description* to infer

[1] We target the models able to concurrently process image, audio, and text as a starting point since these are the most well-explored modalities in the field, but the "omni" concept is extendable.



[Audio Content]
Sound of a man: "You take the blue pill,the story ends, you wake up in your bed and believe whatever you want to believe. You take the red pill, you stay in Wonderland, and I show you how deep the rabbit hole goes."

What is the man holding between his fingers?
A. A red candy.
B. A blue pill.
C. A red pill. ✓
D. A red gem.

[Audio Content]
A:so what are you doing here
B:well I was wondering maybe I can play with you guys
A:play?I don't think this is such a good idea

Will they play together?
A.They will play together
B.They won't play together ✓
C.The girl will play with him
D.The girls refuse to play with him

Figure 2. OmniBench Annotation Scheme. The annotation example shows flawed data that does not pass inspection because the information in audio *alone* is sufficient to answer. The audio in the flawed sample will then be sent back to annotators to edit.

narrative drivers, *Current Action & Activity* to understand ongoing dynamics, and *Future Plot and Purpose Inference* to anticipate subsequent developments; and (3) *abstract concept*, involving *Identity & Relationship* to identify and relate entities, *Text & Symbols* for symbolic interpretation, and *Count & Quantity* for numerical reasoning. This taxonomy is designed to evaluate both foundational perceptual skills and complex cognitive processes, thereby providing a comprehensive assessment of multimodal language models' (MLLMs) abilities to integrate and interpret diverse information sources. OmniBench includes **1142** question-answer pairs, with details on task types, text length, and the characteristics of images and audio presented in Table 1. The audio content of the dataset is categorized into speech, sound events, and music, enriching the diversity of stimuli for evaluating the models' tri-modal capabilities and aiding in the development of future omni-language models.

3.2. Annotation Protocol

Annotation Scheme. Our annotation scheme is built upon a fundamental principle: the correct answer to each question must require information from both the image and audio components. This ensures that the benchmark effectively evaluates the model's ability to analyze information across modalities. As shown in Figure 2, we implemented a rigorous annotation pipeline consisting of three stages: initial annotation, human inspection, and model inspection. Data samples that failed to meet our criteria at any stage were returned to annotators for revision, ensuring high-quality, multimodal-dependent samples. Through the whole process, 16 *annotators and* 5 *quality inspectors* are involved, all are full-time industrial data annotation employee with higher education backgrounds.

The questions are formalized as multi-choice question-answering (MCQ) but try to maintain a consistent logic that suggests the only one possible and accurate answer, *i.e.*, they can be potentially further re-organized into blank filling questions. Furthermore, when constructing the options, the annotators need to ensure at least one confusing wrong option. To ensure question difficulty, annotators were required to verify that questions and options were not trivially easy, lacked distinguishable patterns, and could not be answered by state-of-the-art MLLMs using image information alone. GPT-4 are allowed to use to provide initial annotator self-assessments of question quality. We restrict the images with a minimum resolution of 480P (854x480 pixels) and audio clips with a maximum duration of 30 seconds.

We implemented strict measures to maintain diversity across the dataset. This includes varying image and audio sources, limiting the frequency of individual speakers in audio clips to no more than five occurrences, and restricting the replication of similar instructions or questions. For instance, questions about specific environmental contexts were limited up to three samples. Importantly, annotators were required to provide **rationales** for correct answers, detailing the specific information that should be derived from the image and audio modalities respectively. This approach not only aided in quality inspection but also laid the groundwork for future fine-grained evaluation.

Quality Control. Our quality control process was two-fold, including human inspection round and automatic inspection round assisted by MLLM. First, all annotated instruction-response pairs undergo cross-inspection by human annotators. Inspectors provide detailed reasons for any samples that failed to meet our stringent criteria, allowing for targeted revisions. Samples that pass human inspection are then subjected to a secondary check using a vision-language model LLaVA-1.6-34B (Liu et al., 2024a), where the the automatic quality inspection model is selected by considering trade-off between efficiency and performance. This automated process evaluates each sample under various ablation settings: *image and text only*, *audio transcript*

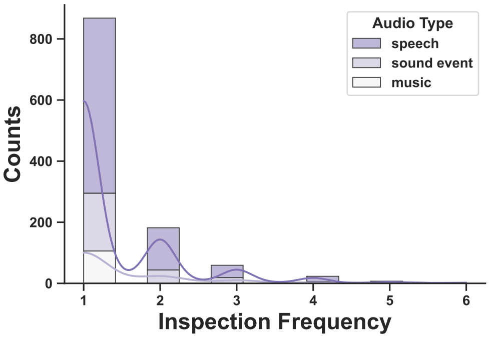


Figure 3. The Distribution of Inspection Frequency of the Passed Samples in OmniBench.

and text only, and *text only* (repeated three times). Samples are only accepted if the model either rejected the task or made mistakes under these limited-information scenarios, confirming the necessity of both visual and auditory information for correct responses. We plot the distribution of the inspection frequency of the passed samples in Figure 3, where we could find that 76% (868) of the passed samples do not require further modification under a well-defined annotation framework and 21.1% of them requiring 1-2 times of revision. During the iterative quality checking, 9.58% (121) QA pairs are defined as "hard to recycle" and dumped after revisions and discussions.

3.3. OmniInstruct

To improve the model capability of tri-modal reasoning, we develop the **OmniInstruct** dataset to facilitate the supervised fine-tuning of models. This dataset leverages the following data sources: the MSRVTT-QA (Xu et al., 2017), AVQA (Yang et al., 2022) and Music-AVQA2.0(Liu et al., 2024c), all of which contain visual, audio and corresponding QA text resources. MSRVTT-QA and AVQA consist of short video clips, typically ranging from 10 to 20 seconds, with minimal scene changes, and music-AVQA 2.0 dataset include 1 minute music performance video. We only adopt the train and validation split of this dataset and regard the whole OmniBench as the test set of the task.

Table 2. Summarization of OmniInstruct Data Distribution across Modalities and Question Types.

	image	audio	how many	what	who	when&where	others
Train	84,580	-	1,131	56,168	22,449	959	3,873
Valid	11,525	-	80	4,498	1,884	66	4,997
Total	96,105	-	1,211	60,666	24,333	1,025	8,870

To construct a dataset that aligns with the challenges proposed in OmniBench, we enhanced each question to connect with an audio track and an image extracted from the corresponding video and filter it with VLMs for better quality. Notably, we avoid the first and last five frames of each video to exclude transitional or obscure incomplete scenes that might distort the task's focus. For MSRVTT-QA train and valid subset, we discard videos without audio tracks and retain a dataset comprised 6,176 videos from the original set that include audio tracks alongside 151.7k QA pairs directly related to these videos. Then we use InternVL-2-76B to filter the questions from the three aforementioned datasets to *delete* (1) questions that can be answered only with an image, (2) questions irrelevant with image, potentially answerable only with audio, and (3) ambiguous or non-logical questions, where the detailed prompt and statistics could be found at Figure 5 and Table 9, Appendix B. After the processing pipeline, only 96k data samples remain for training and validation.

4. Experiment Settings

Baseline Systems We select three groups of MLLM baselines according to the modalities available: *(i) omni-language models*: MIO-Instruct (Wang et al., 2024b), AnyGPT (Zhan et al., 2024), Video-SALMONN (Sun et al., 2024), UnifiedIO2 series (Lu et al., 2024b), the VITA series (Fu et al., 2024; 2025), OpenOmni (Luo et al., 2025), Baichuan-Omni-1.5 (Li et al., 2025), Qwen-2.5-Omni (Xu et al., 2025); *(ii) vision-language models*: InternVL-2 series (Chen et al., 2024b), Qwen2-VL series (Wang et al., 2024a), Deepseek-VL (Lu et al., 2024a), LLaVA-One-Vision series (Li et al., 2024), Cambrian series (Tong et al., 2024), Xcomposer2-4KHD (Dong et al., 2024), Idefics2 (Laurençon et al., 2024) as well as the derived Mantis-Idefics2 (Jiang et al., 2024); *(iii) audio-language models*: LTU series (Gong et al., 2023), Mu-LLaMA (Liu et al., 2023b), MusiLingo (Deng et al., 2023), Qwen-Audio series (Chu et al., 2023a), SALMONN-Audio (Sun et al., 2024) and Audio-Flamingo (Kong et al., 2024). We also include the API calls from proprietary models that could support image-text or audio-text inputs, including GPT4-o, Gemini Pro, Reka and Claude-3.5-Sonnet (Achiam et al., 2023; Team et al., 2023; Ormazabal et al., 2024; Anthropic, 2024). We do not conclude them as in the group of VLMs, ALMs or OLMs (even not a single model) in our context at the moment since the mechanisms behind these models are not revealed[2]. Besides, we invite three musician with higher education background to test on the benchmark and use the average accuracy as a human expert baseline.

Omni-Understanding Evaluation. The main focus of OmniBench is to evaluate how well could the MLLMs understand and reconstruct the context given information from image (I), audio (A) and text (T) modalities. Setting up

[2] The authors conclude from an investigation on September 22, 2024.

Table 3. Overall Omni-Undesratnding Results on Baseline Omni-Language Models. The overall (Image & Audio), image-ablated and audio-ablated results on all samples are provided.

Input Context	Image & Audio	Audio	Image
MIO-Instruct (7B)	24.80%	25.39%	27.93%
AnyGPT (7B)	18.04%	16.20%	20.05%
video-SALMONN (13B)	35.64%	**35.90%**	**34.94%**
VITA (8 × 7B)	33.10%	27.06%	33.54%
VITA-1.5 (7B)	33.40%	-	-
UnifiedIO2-large (1.1B)	27.06%	29.07%	29.07%
UnifiedIO2-xlarge (3.2B)	38.00%	31.17%	34.76%
UnifiedIO2-xxlarge (6.8B)	33.98%	32.49%	33.45%
OpenOmni	37.40%	-	-
Baichuan-Omni-1.5	42.90%	-	-
Qwen-2.5-Omni	**56.13%**	-	-
Gemini-1.5-Pro	42.91%	27.93%	26.09%
Reka-core-20240501	30.39%	23.12%	30.65%
Human Evaluator	63.19%	-	-

Table 4. OLM Baselines Overall Results Grouped by Audio Type.

Model	Speech	Sound Event	Music
AnyGPT (7B)	17.77%	20.75%	13.21%
Video-SALMONN (13B)	34.11%	31.70%	**56.60%**
VITA (8 × 7B)	31.52%	32.45%	46.23%
UnifiedIO2-large (1.1B)	25.94%	29.06%	30.19%
UnifiedIO2-xlarge (3.2B)	39.56%	36.98%	29.25%
UnifiedIO2-xxlarge (6.8B)	34.24%	36.98%	24.53%
Qwen-2.5-Omni	**55.25%**	**60.00%**	52.83%
Gemini-1.5-Pro	42.67%	42.26%	46.23%
Reka-core-20240501	31.52%	26.04%	33.02%
Human Evaluator	58.71%	75.85%	64.15%

questions with four available options for the models, we use accuracy, *i.e.*, the ratio matched letter of the correct option and model response, as the evaluation metric (*n.b.*, the accuracy of a random guess model is 25% under this setting). Additionally, we test the models in an ablation setting of removing one of the image or audio inputs to further reveal a more comprehensive reasoning capability of the baselines and verify the robustness of our benchmark.

Textual Approximation of Image and Audio. For most of the existing MLLMs that only support two input modalities ((I, T) or (A, T)), we build up a simulated evaluation setting allowing us to explore the potential of these models to become omni-language models in the future. We use the audio transcript (A') annotated by human as the alternative of the audios to enable the evaluation on vision-language models. Regarding the audio-language models, we generate high-quality detailed captions of images (I') automatically with a state-of-the-art VLM, InternVL-2-76B. In such an approximated evaluation setting, models go through the same process of inference and metric calculation as the vanilla one with textual alternatives of images or audios.

5. Findings

We evaluate the selected baseline systems in multiple meticulously designed settings, and provide insights on the development of OLMs based on the results. All the claims stated in this section are supported by statistic significance verification, for more information please refer to Appendix C.

5.1. Results on Omni-Language Models

Overall Table 3 demonstrates that most open-source OLM baselines surpass random guessing accuracy across various setting, and the latest developed one (Qwen-2.5-Omni) could even outperform the proprietary models. Notably, the UnifiedIO2 series demonstrates inconsistent performance scaling with model size, indicating challenges in effectively leveraging increased capacity for multimodal understanding. In contrast, Gemini-1.5-Pro and Reka-core-20240501, the two available proprietary models evaluated in this tri-modal setting, shows more promising results. Regarding the scores across audio types, the Gemini-1.5-Pro shows a more balanced performances while Reka-core-20240501 showing a lag on modeling the sound events. Moreover, the comparison of Gemini-1.5-Pro's performance across full input context and ablated settings (image-removed and audio-removed) suggests that it effectively leverages information from all modalities to enhance its reasoning capabilities. While it demonstrates superior overall performance and balanced accuracy across audio types compared to most of the open-source alternatives, its accuracy remains below 50%.

Breakdown Results. We present the breakdown of the performance of open-source omni-language model baselines across different audio types and task categories in the OmniBench evaluation in Table 4 and Table 5. The results reveal inconsistent performance patterns across audio types. For instance, despite overall poor performance, open-source baselines generally exhibit higher accuracy on speech audio, indicating a potential bias towards speech data. Besides, Video-Salmonn and Gemini-1.5-Pro provide better results on music subsets compared to their performance on speech and music, potentially due to their large corpus of music videos, though the music ethics of training foundation models are still under-discussion (Ma et al., 2024). Across task categories, many of the models, such as Gemini-1.5-Pro, tend to perform better on object identification and description tasks while struggling with more reasoning tasks such as plot inference and story description. This might be because visual entity recognition task is an essential component for image captioning and other type of pre-training dataset. Furthermore, some models like Gemini-1.5-Pro, Reka-core-20240501, and Video-SALMONN perform significantly badly on quantity & counting tasks compared with the rest of the tasks. But scaling up of UnifiedIO model contributes a lot to this type of task.

Table 5. OLM Baselines Overall Results Grouped by Task Category.

Accuracy ↑	Causal Inference			(Temporal-)Spatial Entity		Abstract Concept		
Sub-class of QA	Action & Activity	Story Description	Plot Inference	Object Identification & Description	Contextual & Environmental	Identity & Relationship	Text & Symbols	Count & Quantity
AnyGPT (7B)	19.52%	16.52%	14.77%	22.27%	15.60%	21.88%	12.00%	33.33%
Video-SALMONN (13B)	31.47%	28.26%	25.74%	62.56%	36.88%	**37.50%**	20.00%	6.67%
VITA (8 × 7B)	35.86%	33.04%	29.54%	33.65%	41.84%	21.88%	16.00%	6.67%
UnifiedIO2-large (1.1B)	29.88%	20.87%	31.65%	30.81%	23.40%	18.75%	24.00%	6.67%
UnifiedIO2-xlarge (3.2B)	32.27%	**33.48%**	31.65%	**63.03%**	34.04%	34.38%	24.00%	20.00%
UnifiedIO2-xxlarge (6.8B)	32.27%	29.13%	29.96%	48.82%	34.75%	25.00%	8.00%	**46.67%**
Gemini-1.5-Pro	**41.83%**	30.87%	**32.91%**	62.56%	**60.28%**	31.25%	**28.00%**	13.33%
Reka-core-20240501	25.50%	24.78%	20.68%	49.76%	39.01%	28.12%	**28.00%**	6.67%
Human Evaluator	72.91%	55.51%	53.87%	71.41%	65.48%	59.38%	45.33%	66.67%

Human Evaluation We invite three human annotators to evaluate on the OmniBench questions, deriving a much higher accuracy 63.19% compared to all the OLMs[3]. The human evaluator results hold a Fleiss' Kappa value of 0.421 suggests, suggesting a high level of inter-annotator agreement. Different from models, human evaluation shows a particularly good performance on ”Sound Event” among the audio types, which is reasonable since the sounds are short and straightforward without requiring complex reasoning to recognize. Moreover, human shows higher scores on ”Abstract Concept” tasks, potentially because these tasks require more reasoning.

5.2. The Effectiveness of OmniInstruct

To validate the effectiveness of OmniInstruct, we conducted experiments using MIO-instruct, a 7B parameter omni-language model previously trained only on audio-language and vision-language pairs. We evaluated the impact of supervised fine-tuning using OmniInstruct through two experimental settings. First, we created a compact subset of OmniInstruct containing 6.4K samples (approximately 7.5% of the full dataset) to enable efficient experimentation. In our first setting (MIO-Instruct-Omni-V1), we directly fine-tuned the model using its vanilla multimodal tokenizers. In the second setting (MIO-Instruct-Omni-V1-voice-filtered), we refined the approach by filtering out non-speech tokens from the audio RVQ tokenizer that only supports speech, aligning with MIO-Instruct's original training to avoid mode collapse. The results demonstrate meaningful improvements: while the baseline MIO-Instruct achieved 24.8% accuracy on OmniBench, fine-tuning with the vanilla setting improved performance to 25.7%, and the voice-filtered approach further enhanced accuracy to 29.2%. These results indicate that even a small subset of OmniInstruct can effectively adapt existing multimodal language models to tri-modal scenarios. Notably, the superior performance of the voice-filtered approach suggests that aligning fine-tuning data distribution with a model's distribution can yield better results.

Table 6. Results on Textual Audio Approximation Experiments. All the audios are represented in text transcript. The results are divided into groups of vision-language models and omni-models. We use the text transcript to approximate the audios in this setting. Boldface shows the best model performance, and underline shows the best open-source model.

Input Context	Image & Audio Transcript	Audio Transcript	Image
InternVL-2-2B	42.29%	27.32%	28.11%
InternVL-2-8B	50.79%	33.63%	33.36%
InternVL-2-26B	51.75%	31.87%	33.89%
InternVL-2-40B	54.29%	31.96%	34.76%
Qwen2-VL-Chat-2B	42.47%	31.44%	38.09%
Qwen2-VL-Chat-7B	48.60%	32.05%	36.87%
Deepseek-VL-Chat-7B	39.67%	29.51%	26.27%
Idefics2-8B	45.10%	32.31%	34.41%
Mantis-Idefics-8B	46.15%	36.43%	32.57%
LLaVA-OneVision-0.5B	38.00%	31.79%	31.17%
LLaVA-OneVision-7B	47.02%	31.70%	29.68%
Cambrian-8B	42.12%	31.35%	32.22%
Cambrian-13B	45.01%	31.96%	33.98%
Cambrian-34B	46.76%	30.12%	33.01%
XComposer2-4KHD (7B)	43.96%	29.25%	30.65%
GPT4-o (0513)	57.62%	45.71%	42.21%
GPT4-o (0806)	51.14%	47.55%	31.44%
GPT4-o-mini	49.04%	39.23%	34.06%
Gemini-1.5-Pro	44.40%	22.50%	26.09%
Reka-core-20240501	46.58%	34.59%	30.65%
Claude-3.5-Sonnet	**59.37%**	33.54%	**43.08%**
GPT-4V-Preview	38.18%	41.24%	25.57%
GPT-4V-0409	33.36%	**45.80%**	32.75%
UnifiedIO2-large (1.1B)	34.33%	31.96%	29.07%
UnifiedIO2-xlarge (3.2B)	43.17%	34.50%	34.76%
UnifiedIO2-xxlarge (6.8B)	40.81%	29.77%	33.45%

5.3. Textual Approximation on Images and Audios

As the absence of strong OLM baselines, we further introduce the text alternatives of images (I') and audios (A') to embrace more dual-modal MLLMs to analyze the current research progress in the field.

Performance Changes of Open OLMs. We select the UnifiedIO-2 series to conduct the textual approximation experiments due to their relatively robust performances in the vanilla evaluation setting suggested in Table 3. Compared with the vanilla setting, all three UnifiedIO-2 models

[3] Human performance is limited by time constraints, annotators' non-native language or music knowledge gaps, and their associate-level education.

Table 7. Results on Textual Image Approximation Experiments. All the images are represented in text caption. The results are divided into groups of audio-language models and omni-models.

Accuracy ↑	All Audio Types			Speech	Sound Event	Music
Input Context	Image Caption & Audio	Audio	Image Caption	Image Caption & Audio		
LTU (7B)	23.29%	23.91%	23.12%	25.42%	20.00%	16.04%
Mu-LLaMA (7B)	1.58%	1.84%	1.84%	1.56%	1.13%	2.83%
MusiLingo-long-v1	13.66%	11.03%	9.02%	11.93%	13.96%	25.47%
Audio-SALMONN (13B)	34.76%	32.66%	33.36%	34.50%	29.43%	**50.00%**
Qwen-Audio-Chat (7B)	17.51%	16.64%	18.39%	14.66%	22.64%	25.47%
Qwen2-Audio-7B-Instruct	**40.72%**	**35.20%**	**35.29%**	**40.60%**	**41.89%**	38.68%
Audio-Flamingo (1.3B)	24.78%	23.82%	24.78%	26.98%	21.51%	16.98%
Gemini-1.5-Pro	38.62%	28.02%	21.02%	39.82%	33.96%	41.51%
Reka-core-20240501	29.42%	23.12%	26.27%	28.53%	29.43%	35.85%
UnifiedIO2-large (1.1B)	29.16%	29.07%	29.33%	28.40%	32.45%	26.42%
UnifiedIO2-xlarge (3.2B)	32.22%	31.17%	30.21%	32.43%	32.45%	30.19%
UnifiedIO2-xxlarge (6.8B)	32.05%	32.49%	27.15%	31.13%	38.87%	21.70%

show performance gains, averagely at 6.42%, in the audio replacement setting and average performance drops in the replaced-image (1.87%) and both-repaced settings (0.12%). This indicates the shortcoming of existing OLMs on modeling the audio on the one hand, and the potential noise in the generated image captions compared to the human-written audio transcripts on the other hand.

Performances of Dual-modal MLLMs. In the setting of using text as the alternatives of audios and images, the VLMs show generally better results than ALMs (Table 6 vs Table 7) even compared with open-source model with similar model size[4]. This could be caused by : 1) more available research resources have been put in VLMs to develop datasets and cross-modality alignment architectures, leading to higher instruction following rate and accuracy compared to ALMs; 2) the audio data are naturally harder (and hence more expensive) to annotate; and 3) audio typically has longer sequence tokens and requires more computational resource compared to text and image, making it harder to scale up. If I' and A' have the information loss ratio when converted from I and A, it seems to be easier for the researchers to train the future omni-language models from exisiting VLMs rather than ALMs. Besides, we can observe Claude-3.5 and GPT-4o are generally the best two VLMs, significantly better compared to open-source VLMs. Qwen2- audio and Gemini are the two best ALMs in speech, Qwen2- audio is the best in sound, and Audio-SALMONN is music. Moreover, we can see significant differences in different types of audio, i.e., LTU and audio-flamingo are worse for music compared to speech, while Qwen-audio, which includes music on pre-training, provides better results on music compared to speech. And MusiLingo only uses music for pre-training performs worse in speech and audio.

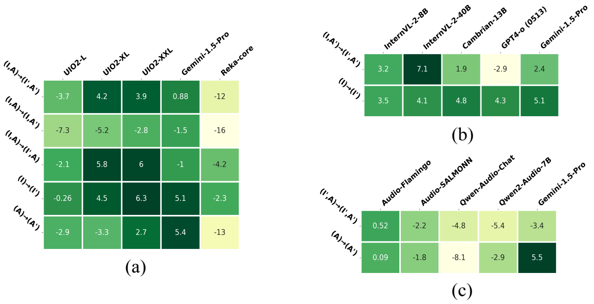


Figure 4. The Performance Changes Brought By Textual Alternatives. The numbers in the cell suggest the accuracy change. *(a)* includes the UnifiedIO2 OLMs and the proprietary models supporting tri-modal inputs. *(b)* and *(c)* consists of VLMs and ALMs grouped with Gemini-1.5-Pro for comparison.

Pure Textual Evaluation. The performance gaps brought by the replaced textual image and audio descriptions are in revealed in Figure 4. Notably, the majority of models demonstrate improved accuracy when processing textual representations of multimodal data compared to their performance on either image captions or audio transcripts alone. This suggests that these models show stronger reasoning capability when equipped with information from multiple textual sources rather than handling raw multimodal inputs. For instance, Qwen2-Audio-7B-Instruct shows a significant jump in accuracy from 39.05% (audio transcript only) and 39.67% (image caption only) to 47.02% when given both textual representations. But the gains of best proprietary models GPT4-o (0513) andClaude-3.5-Sonnet exhibit is not significant, though GPT4-o (0513) achieved an impressive 60.60% accuracy in the pure textual setting.

6. Conclusion and Future Study

The proposed novel multimodal benchmark, OmniBench, reveals that current open-source multimodal large language models struggle with simultaneous processing of visual,

[4] The results of pure text (I', A') setting are placed at Table 8, Appendix A.

acoustic, and textual inputs. We observed a general bias towards speech audio and superior performance of vision-language models over audio-language models when using textual approximations. These findings underscore the need for more appropriate architecture designs for multimodal integration, diverse datasets for training, and techniques to reduce modality bias. OmniBench serves as a crucial tool for guiding advancements in multimodal language models, driving progress towards more advanced and versatile models towards human-like multimodal understanding and reasoning.

Impact Statement

This paper introduces OmniBench, a crucial step towards developing truly multimodal AI. By rigorously evaluating models on their ability to integrate visual, acoustic, and textual information, OmniBench exposes critical limitations in current approaches and highlights the need for dedicated research in tri-modal reasoning. This benchmark has significant potential societal impact, as robust OLMs could revolutionize fields like healthcare (improved diagnostics), accessibility (enhanced assistive technologies), and human-computer interaction (more natural interfaces). Furthermore, the accompanying OmniInstruct dataset provides a valuable resource for training and fine-tuning these models. While the development of OLMs raises ethical considerations regarding bias and fairness, OmniBench's focus on comprehensive evaluation enables researchers to identify and address these issues proactively. By fostering transparency and reproducibility, this work paves the way for responsible development and deployment of powerful multimodal AI systems that benefit society.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. Claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Bu, H., Du, J., Na, X., Wu, B., and Zheng, H. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pp. 1–5. IEEE, 2017.

Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., and Wei, F. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J., and Liu, J. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.

Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024b.

Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pp. 493–507, 1952.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023a.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023b.

Deng, Z., Ma, Y., Liu, Y., Guo, R., Zhang, G., Chen, W., Huang, W., and Benetos, E. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. In *NAACL-HLT*, 2023. URL `https://api.semanticscholar.org/CorpusID:262043691`.

Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Zhang, S., Duan, H., Zhang, W., Li, Y., et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024.

Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Drossos, K., Lipping, S., and Virtanen, T. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Du, J., Na, X., Liu, X., and Bu, H. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*, 2018.

Fu, C., Lin, H., Long, Z., Shen, Y., Zhao, M., Zhang, Y., Dong, S., Wang, X., Yin, D., Ma, L., et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Fu, C., Lin, H., Wang, X., Zhang, Y.-F., Shen, Y., Liu, X., Li, Y., Long, Z., Gao, H., Li, K., et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Gong, Y., Yu, J., and Glass, J. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155. IEEE, 2022.

Gong, Y., Luo, H., Liu, A. H., Karlinsky, L., and Glass, J. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

Guyon, I., Makhoul, J., Schwartz, R., and Vapnik, V. What size test set gives good error rate estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (1):52–64, 1998.

Hemdan, E. E.-D., El-Shafai, W., and Sayed, A. Cr19: A framework for preliminary detection of covid-19 in cough audio signals using machine learning algorithms for automated medical diagnosis applications. *Journal of Ambient Intelligence and Humanized Computing*, 14(9): 11715–11727, 2023.

Jiang, D., He, X., Zeng, H., Wei, C., Ku, M., Liu, Q., and Chen, W. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Kong, Z., Goel, A., Badlani, R., Ping, W., Valle, R., and Catanzaro, B. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*, 2024.

Laurençon, H., Tronchon, L., Cord, M., and Sanh, V. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.-R., and Hu, D. 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.

Li, J. and Lu, W. A survey on benchmarks of multimodal large language models. *ArXiv*, abs/2408.08632, 2024. URL https://api.semanticscholar.org/CorpusID:271892136.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E., et al. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*, 2023b.

Li, Y., Liu, J., Zhang, T., Chen, S., Li, T., Li, Z., Liu, L., Ming, L., Dong, G., Pan, D., et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Liang, J., Nolasco, I., Ghani, B., Phan, H., Benetos, E., and Stowell, D. Mind the domain gap: a systematic analysis on bioacoustic sound event detection. *arXiv preprint arXiv:2403.18638*, 2024a.

Liang, J., Phan, H., and Benetos, E. Learning from taxonomy: Multi-label few-shot classification for everyday sound recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 771–775. IEEE, 2024b.

Liu, F., Zhu, T., Wu, X., Yang, B., You, C., Wang, C., Lu, L., Liu, Z., Zheng, Y., Sun, X., et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226, 2023a.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Liu, S., Hussain, A. S., Sun, C., and Shan, Y. Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning. *arXiv preprint arXiv:2308.11276*, 2023b.

Liu, X., Dong, Z., and Zhang, P. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4478–4487, 2024c.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023c.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Sun, Y., et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024a.

Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. 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, 2024b.

Luo, R., Lin, T.-E., Zhang, H., Wu, Y., Liu, X., Yang, M., Li, Y., Chen, L., Li, J., Zhang, L., et al. Openomni: Large language models pivot zero-shot omnimodal alignment across language with real-time self-aware emotional speech synthesis. *arXiv preprint arXiv:2501.04561*, 2025.

Ma, Y., Øland, A., Ragni, A., Del Sette, B. M., Saitis, C., Donahue, C., Lin, C., Plachouras, C., Benetos, E., Quinton, E., et al. Foundation models for music: A survey. *arXiv preprint arXiv:2408.14340*, 2024.

Meskó, B. The impact of multimodal large language models on health care's future. *Journal of medical Internet research*, 25:e52865, 2023.

Ormazabal, A., Zheng, C., d'Autume, C. d. M., Yogatama, D., Fu, D., Ong, D., Chen, E., Lamprecht, E., Pham, H., Ong, I., et al. Reka core, flash, and edge: A series of powerful multimodal language models. *arXiv preprint arXiv:2404.12387*, 2024.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. arxiv 2022. *arXiv preprint arXiv:2212.04356*, 10, 2022.

Su, H., Qi, W., Chen, J., Yang, C., Sandoval, J., and Laribi, M. A. Recent advancements in multimodal human–robot interaction. *Frontiers in Neurorobotics*, 17:1084000, 2023.

Sun, G., Yu, W., Tang, C., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Wang, Y., and Zhang, C. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

Sun, H., Liu, X., Xu, K., Miao, J., and Luo, Q. Emergency vehicles audio detection and localization in autonomous driving. *arXiv preprint arXiv:2109.14797*, 2021.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., and Zhang, C. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Terenzi, A., Ortolani, N., Nolasco, I., Benetos, E., and Cecchi, S. Comparison of feature extraction methods for sound-based classification of honey bee activity. *IEEE/ACM transactions on audio, speech, and language processing*, 30:112–122, 2021.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., Wang, A., Fergus, R., LeCun, Y., and Xie, S. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. arxiv 2020. *arXiv preprint arXiv:2012.12877*, 2(3), 2020.

Wang, C., Liao, M., Huang, Z., Lu, J., Wu, J., Liu, Y., Zong, C., and Zhang, J. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*, 2023a.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL `https://arxiv.org/abs/2409.12191`.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., and Tang, J. Cogvlm: Visual expert for pretrained language models, 2023b.

Wang, Z., Zhu, K., Xu, C., Zhou, W., Liu, J., Zhang, Y., Wang, J., Shi, N., Li, S., Li, Y., Que, H., Zhang, Z., Zhang, Y., Zhang, G., Xu, K., Fu, J., and Huang, W. Mio: A foundation model on multimodal tokens, 2024b. URL `https://arxiv.org/abs/2409.17692`.

Wu, J., Gaur, Y., Chen, Z., Zhou, L., Zhu, Y., Wang, T., Li, J., Liu, S., Ren, B., Liu, L., et al. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023a.

Wu, S., Li, Y., Zhu, K., Zhang, G., Liang, Y., Ma, K., Xiao, C., Zhang, H., Yang, B., Chen, W., et al. Scimmir: Benchmarking scientific multi-modal information retrieval. *arXiv preprint arXiv:2401.13478*, 2024a.

Wu, S., Zhu, K., Bai, Y., Liang, Y., Li, Y., Wu, H., Liu, J., Liu, R., Qu, X., Cheng, X., et al. Mmra: A benchmark for multi-granularity multi-image relational association. *arXiv preprint arXiv:2407.17379*, 2024b.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., and Lin, J. Qwen2.5-omni technical report, 2025. URL `https://arxiv.org/abs/2503.20215`.

Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., and Zhu, W. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3480–3491, 2022.

Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., Leng, Y., Lv, Y., Zhao, Z., Zhou, C., and Zhou, J. Air-bench: Benchmarking large audio-language models via generative comprehension. *ArXiv*, abs/2402.07729, 2024. URL `https://api.semanticscholar.org/CorpusID:267626820`.

Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., et al. Yi: Open foundation models by 01. ai, 2024.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 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.

Yun, H., Yu, Y., Yang, W., Lee, K., and Kim, G. 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.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Zhang, G., Du, X., Chen, B., Liang, Y., Luo, T., Zheng, T., Zhu, K., Cheng, Y., Xu, C., Guo, S., Zhang, H., Qu, X., Wang, J., Yuan, R., Li, Y., Wang, Z., Liu, Y., Tsai, Y.-H., Zhang, F., Lin, C., Huang, W., Chen, W., and Fu, J. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *ArXiv*, abs/2401.11944, 2024. URL `https://api.semanticscholar.org/CorpusID:267068665`.

A. More Experiment Results

Table 8. Results on Pure Textual Approximation for *Both Image and Audio*. All the images and audios are represented in texts. The results at the second and third column are taken from the corresponding models in Table 6 and Table 7.

Input Context	Image Caption & Audio Transcript	Audio Transcript	Image Caption
LTU (7B)	22.68%	24.17%	23.12%
Mu-LLaMA (7B)	2.28%	6.57%	1.84%
MusiLIngo-long-v1 (7B)	11.03%	10.51%	9.02%
Audio-SALMONN-13B	36.95%	34.41%	33.36%
Qwen-Audio-Chat	22.33%	24.69%	18.39%
Qwen2-Audio-7B-Instruct	46.15%	38.09%	35.29%
Audio-Flamingo (1.3B)	24.26%	23.73%	24.78%
InternVL-2-8B	47.55%	33.63%	29.86%
InternVL-2-40B	47.20%	31.96%	30.65%
Cambrian-13B	43.08%	31.96%	29.16%
GPT4-o (0513)	60.51%	45.71%	37.92%
GPT4-o (0806)	53.77%	47.55%	29.51%
GPT4-o-mini	51.05%	49.04%	32.84%
Gemini-1.5-Pro	42.03%	22.50%	21.02%
Claude-3.5-Sonnet	56.83%	33.54%	39.05%
Reka-core-20240501	42.23%	36.33%	32.94%
GPT-4V-Preview	33.27%	41.24%	20.32%
GPT-4V-0409	29.95%	45.80%	20.84%
UnifiedIO2-large (1.1B)	30.74%	31.96%	29.33%
UnifiedIO2-xlarge (3.2B)	33.80%	34.50%	30.21%
UnifiedIO2-xxlarge (6.8B)	34.15%	29.77%	27.15%

B. Dataset Development

B.1. Statistics for OmniInstruct Dataset

Table 9. The Statistics of Data Filtering in OmniInstruct. The table shows the number changes of question-answer pairs before and after filtering from each of the data sources.

Source	Original Train	Original Valid	Remained Train	Remained Valid
AVQA	40,182	16,798	4,491 (11.2%)	1,911 (11.4%)
Music-AVQA2.0	42,470	0	11 (0.03%)	0
MSRVTT-QA	140,554	11,143	80,078 (57.0%)	9,614 (86.2%)
Total	233,206	27,941	**84,580**	**11,525**

As demonstrate in Table 9, most of the samples in the dataset are in low quality and therefore abandoned, and only 96k of samples remain for Omni-modality SFT training. This is reasonable because most of the questions are generated from templates, and the image may not sampled from the most relevant part of the questions and, therefore hot high in quality.

B.2. Prompt for Quality Control on OmniInstruct Dataset

B.3. Diversity of Music Audios of OmniBenchmark

The music subset of our benchmark reflects a rich diversity of musical traditions, spanning a wide range of genres, styles, and cultural contexts. It encompasses Western classical symphonies, jazz chamber music, and avant-garde compositions alongside popular music from China, England, and France. Traditional forms like Kunqu opera and modern experimental pieces are represented, as well as instrumental music from regions such as India, the Arab world, Africa, and Japan. The benchmark also includes famous film soundtracks with various thematic elements and Asian folk oral traditions, such as chanting, drumming, and Humai. This eclectic collection, enriched by unique instances like famous concert spoofs and iconic YouTube parodies, ensures that each question offers a distinct challenge, showcasing the nuanced intricacies and breadth of global music heritage.

```
Initial Q&A: {question and answer}

The given Q&A is originally designed to answer based on the complementary context
built from an audio and an image together. Please evaluate whether the provided
Q&A is a bad/flawed sample due to one of the following reasons:

1. The answer could be inferred solely from the given image without the assistance
of audio;

2. The Q&A is not relevant to the image;

3. The Q&A is logically inconsistent.

After your evaluation, respond with 'Yes' if the Q&A is a flawed sample should be
removed, else response with 'No'.
```

Figure 5. The Prompt for OmniInstruct Dataset Filtering.

C. Statistic Significance of the Findings

In this paper, we provide about 30 claims on the discussion in section 5. In order to discover whether the observation is due to the bias of a small test set, We utilize multiple types of hypothesis testing for these experimental results.

C.1. Inference if Two Models Provide Significantly Different Results under the Same Small Subset

The method is inspired by (Guyon et al., 1998). Suppose all the data samples $\{x_i\}_{i=1}^{n}$ are i.id, where n is the number of data samples. For each specific model M_j, the performance of the j^{th} model on the i^{th} data is $M_j\left(x_i\right)$ which is equal to 0 if the model predicts the wrong choice and 1 otherwise. Due to the i.id. assumption of the test set, the $\{M_j\left(x_i\right)\}_{i=1}^{n}$ is a random variable sequence (r.v.s.) with binomial distribution with mean p_i and variance $\sigma_i^2 = p_i(1-p_i)$.

Suppose the accuracy p_i of a given model is estimated by computing the average correct rate $\hat{p_i}$ over a finite number n of test examples or patterns. When n is small, $\hat{p_1} > \hat{p_2}$ may hold even when $p_1 < p_2$.

According to the normal law, the whited random variable

$$Z := \frac{p - \hat{p}}{\sigma\sqrt{n}}$$

obeys the standardized Normal law (with mean 0 and variance 1). Define the real number Z_α to be the value such that the probability $\mathbb{P}\left(x > Z_\alpha\right) = \alpha$. Then, according to Chebychev's inequality,

$$\mathbb{P}\left(p - \hat{p} \leq \frac{\sigma z_{alpha}}{\sqrt{n}}\right) \leq \alpha$$

.

We are then calculating the number of test examples n that is needed to guarantee a certain margin of error ϵ_i, e.g., $\epsilon(n, \alpha) = \frac{z_\alpha \sigma}{\sqrt{n}}$for the Normal law). Denote $\beta := \frac{\epsilon_i}{p_i}$. Then

$$p - \hat{p} \leq z_\alpha \sqrt{\frac{p(1-p)}{n}}$$

hold with the probability $1 - \alpha$.

Therefore,

$$n \leq \left(\frac{z_\alpha}{\beta}\right)^2 \frac{1-p}{p}$$

holds with the probability $1 - \alpha$.

According to the cdf of standard normal distribution,

$$n\frac{\beta^2 p_1}{1-p_1} = \frac{n(p_2-p_1)^2}{p_1(1-p_1)} \leq 2.72$$

implies significance 0.05, and the value ≤ 5.29 implies p-value less than 0.01.

SImilarly, with Chernoff bound (Chernoff, 1952), $p - \hat{p} \leq \sqrt{\frac{-2p(\ln\alpha)}{n}}$ hold with the probability $1-\alpha$. And therefore, $n = \frac{-2\ln\alpha}{\beta^2 p}$. The alternative estimation of p-value is $\alpha = \exp\left(\frac{np\beta^2}{2}\right)$

Therefore, we have the following claim with significance value:

- In table2
  - Claim (1): Scaling up may not contribute to the overall performance. With audio and image as input, UnifiedIO2-xlarge's performance 38.00% is better than the scaled-up UnifiedIO2-xxlarge's (33.98%) version on the 1142 samples holds, with a p-value is less than 0.01.
  - Claim (2): Performance f Gemini-1.5-Pro with image and audio input (42.91%) is better than single modality input (e.g. 27.93% for image only) on the 1142 samples holds with p-value 1.19e-20.
- Table 3
  - Claim (3): Scaling up of UnifiedIO can contribute to Count & Quantity tasks, though only 15 samples are selected. UnifiedIO-large (6.67%) is worse than UnifiedIO-xlarge (20%) holds with a p-value less than 0.05, and UnifiedIO-xlarge (20%) is worse than UnifiedIO-xxlarge (46.67%) holds with a p-value less than 0.01.
- In table 4 & table 5
  - - Claim (5): The performance of Claude-3.5 (59.37%) and GPT-4o (57.62%) in Table 4 is better than open-source VLM (less or equal to 54.29%). We can say Claude-3.5 is better than InternVL-2-40B (54.29%) with a p-value less than 0.05, and GPT-4o is better than the second-best open-source VLM InternVL-2-26B (51.75%) with p-value less than 0.01. But we can not say GPT-4o (57.62%) is better than InternVL-2-40B (54.29%).
  - Claim (6): **In speech subset (771 samples)** demonstrated in table 3 with audio and image as input, Gemini-1.5-Pro (42.67%) does not surpass UnifiedIO2-xlarge (3.2B) (39.56%) significantly, but it surpass all other models like the second-best UnifiedIO2-xxlarge (6.8B) (34.24%) with p-value less than 0.01
  - Claim (7): **In general audio subset (265 samples)** demonstrated in Table 3 with audio and image as input, Gemini-1.5-Pro (42.26%) is better than all the others besides UnifiedIO2, such as Video-SALMONN (13B) (31.70%) with p-value less than 0.01
  - Claim (8): **In the music subset (106 samples)** demonstrated in Table 3 with audio and image as input, Video-SALMONN (13B) (56.6%) is better than all the other models such as the best one Gemini-1.5-Pro (46.23%) with p-value less than 0.05.
  - Claim (9): **In the speech subset (771 samples)** demonstrated in Table 5 with audio and image caption as input, Qwen2-audio (40.60%) and Gemini (39.82%) are better than all other ALMs. E.g. Gemini is better than the best of the rest Audio-SALMONN (13B) (34.5%) with p-value less than 0.01
  - Claim (10): **In general audio subset (265 samples)** demonstrated in table 5 with audio and image caption as input, Qwen2-audio (41.89%) is better than every model besides UnifiedIO2-xxlarge (38.87%), e.g. better than Gemini-1.5-Pro (33.96%) with p-value less than 0.05.
  - Claim (11): **In music subset (106 samples)** demonstrated in table 5 with audio and image caption as input, Audio-SALMONN (50%) is better than all the others besides Gemini-1.5-Pro (41.5%). For example, Audio-SALMONN is better than the best of the rest Qwen2-Audio-7B-Instruct (38.68%) with a p-value less than 0.05.
- Compared with Table 6
  - Claim (12): In Table 6, with audio transcription and image caption as inputs, Qwen2-Audio-7B-Instruct (47.02%) is better than with only audio transcription (39.05%) or only image caption (39.67%) with a p-value less than 0.01

  - Claim (13): Replace the input of Qwen2-Audio-7B-Instruct from image caption and audio in Table 5 (40.72%) to caption and audio transcription in Table 6 (47.02%), the performance increases significantly with a p-value less than 0.01
  - Claim (14): Replace the input of Claude-3.5-Sonnet from image and audio transcription in Table 4 (59.37%) to image caption and transcription in Table 6 (56.83%) has no significant changes. And GPT4-o (0513) performances in Table 4 (57.62%) and Table 6 (60.51%) has no significant changes either.
  - Claim (15): GPT4-o (0513) performance in Table 6 (60.51%) has no significant difference with human (63.19%)

C.2. Comparasion between Two Different Setting with the Same Samples

We use student-paired t-tests to compare two different experimental settings on the same testset.

Recaall the performance of UnifiedIO2 in different experimental settings are as follows:

- performance with image and audio (25.94%, 39.56%, 34.24% in Table 2)
- performance with image caption & audio (29.16%, 32.22%, 32.05% in Table 5)
- performance with image & audio traiscripton (34.33%, 43.17%, 40.81% in Table 4)
- performance with pure text input (30.74%, 33.80%, 34.15% in Table 6)

Then the t-test results are as follows:

- between Table 2 and Table 4 with audio transcription has a p-value of 0.0471, showing the model has room for improvement in audio understanding capability compared with audio transcription ground truth.
- Claim (17): There is no significant difference between Table 2 and Table 5 (p-value 0.562) or Table 6 (p-value 0.919), showing the capability of OLM on image understanding is similar to the SOTA LLM generated image caption.

C.3. Comparasion on the Same Setting with Two Different Subset Samples

We utilize Wilcoxon rank sum test (i.e. Mann–Whitney U test) to compare the model performance on two different subsets with different but independent samples.

- In table 3 (top):
  - Claim (18): The performance of salmonn on the music subset (56.60%, 106 samples) is better than speech (34.11%, 771 samples) with p-value 6.84e-6; and better than sound events (31.70%, 265 samples) with p-value 8.98e-6.
  - Claim (19): The performance of UnifiedIO2-xlarge (3.2B) music (29.25%) may not be worse than sound (36.98%) with p-value 0.158; but may be worse than speech (39.56%) with p-value 0.0407.
  - Claim (20): The performance of Reka-core-20240501 on sound (26.04%) may not be worse than music (33.02%) with p-value 0.177; and may not be worse than speech (31.52%) with p-value 0.093.
- table3 (bottom)
  - Claim (21): the performance of Gemini-1.5-Pro on Story Description (30.87%, 230 samples) is worse than all the other 7 class with p-value 5.38e-5; Plot Inference (32.91%, 237 samples) worse than others with p-value 6.69e-4; Object Identification & Description (62.56%, 211 samples) is better than others with p-value 9.46e-11. This statement holds for many other models, but not for AnyGPT.
  - Claim (22): Gemini-1.5-Pro on Count & Quantity (13.33%, 15 samples) is worse than all the other 7 classes with a p-value of 0.0209. Reka-core-20240501 on Count & Quantity (6.67%) is worse than the rest with a p-value of 0.0445. Video-SALMONN on Count & Quantity (6.67%) is worse than the rest with a p-value of 0.0239. But not for AnyGPT.
- In table 5

- claim (23): The performance of LTU on music (16.04%) is worse than speech (25.42%) with a p-value of 0.0348, but may not be worse than sound (20.00%) with a p-value of 0.379.
- claim (24): The performance of audio-flamingo on music (16.98%) is worse than speech (26.98%) with a p-value of 0.0275, but may not be worse than sound (21.51%) with a p-value of 0.328.
- claim (25): The performance of Qwen-audio on music (25.47%) is better than speech (14.66%) with a p-value of 0.0044.
- claim (26): The performance of MusiLingo on music (25.47%) is better than speech (11.93%) with a p-value of 1.37e-4, and better than sound (13.96%) with a p-value of 5.96e-3.

Given that we conducted 20-30 hypothesis tests, it is not surprising to observe at least one p-value below 0.05 due to the multiple comparisons, which could suggest a false positive in some cases. However, many of the tests involve the same models or subsets, meaning the p-values are not independent. As a result, the observed number of significant p-values (¡ 0.04, in fact, ¡ 0.01 for many cases) is less indicative of false positives than it would be under the assumption of independent tests. This reduces the likelihood of false positives in our findings.

D. Human Evaluation

Human performance is lower than expected due to several factors: (1) Annotation was conducted under time constraints, limiting thorough reasoning. (2) Annotators may lack relevant background knowledge—some were non-native English/French speakers, potentially affecting speech or lyric comprehension. (3) Annotators held only associate degrees, which may impact annotation quality in complex reasoning tasks such as music expertise.