

UNO-Bench: A Unified Benchmark for Exploring the Compositional Law Between Uni-modal and Omni-modal in OmniModels

Chen Chen^{*} Zeyang Hu^{*} Fengjiao Chen[†]
Liya Ma Jiaxing Liu Xiaoyu Li[†] Xuezhi Cao

Meituan, Beijing, China.

{chenchen165, huzeyang, chenfengjiao02, maliya06}@meituan.com
{liujiaxing10, lixiaoyu28, caoxuezhi}@meituan.com

Abstract

Multimodal Large Languages models have been progressing from uni-modal understanding toward unifying visual, audio and language modalities, collectively termed omni models. However, the correlation between uni-modal and omni-modal remains unclear, which requires comprehensive evaluation to drive omni model’s intelligence evolution. In this work, we propose a novel, high quality and **UNified Omni** model benchmark, **UNO-Bench**, which effectively assesses both **UNi-modal** and **Omni-modal** capabilities. The benchmark consists of 3730 human curated samples, across 44 task types, and a innovative multi-step open-ended question type that better assess complex reasoning tasks. Besides, a general scoring model supporting 6 question types is proposed for automated evaluation with 95% accuracy. Experimental result shows the compositional law between cross-modal and uni-modal performance and the omni-modal capability manifests as a bottleneck effect on weak models, while exhibiting synergistic promotion on strong models. Our code and data are available at [GitHub](#).

1 Introduction

Multimodal artificial intelligence has undergone extensive researches in visual language model and audio language model, with current advancements progressing toward unifying visual, audio and language modalities, collectively termed omni models. The evaluation paradigm for these models has consequently expanded from assessing uni-modal understanding capabilities (i.e. visual understanding, audio understanding) to the next-level of intelligence (i.e. omni-modal understanding).

Existing omni model evaluation benchmarks remain relatively scarce and focus on different aspects. For example, some prioritize image comprehension([Li et al., 2024b](#)), others emphasize video

understanding([Hong et al., 2025](#)), while a subset concentrates on speech interaction([Gong et al., 2024](#)). Notably, existing datasets are exclusively English-centric, lacking evaluation benchmarks for Chinese linguistic contexts.

The ideal omni model should simultaneously preserve visual understanding capabilities (e.g., MMBench([Liu et al., 2024b](#))/MathVista([Lu et al., 2024a](#))/MVBench([Li et al., 2024a](#))), speech interaction proficiency (e.g., MMAU([Sakshi et al., 2025](#))), and cross-modal integration capacity (e.g., OmniBench([Li et al., 2024b](#))/WorldSense([Hong et al., 2025](#))). Current evaluation paradigms employ disjointed benchmark suites for separate capability testing, creating resource-intensive evaluation processes and disconnected modality assessments. Beyond uni-modal, omni-modal capability introduces advanced challenges across image, video and audio modality. However, 77% questions from WorldSense are solvable without video or audio, and 25% questions from OmniBench contain erroneous answers. These designs limit the evaluation and analysis of omni models’ abilities.

Due to the limited quality and coverage of existing benchmarks, we introduce a novel and unified benchmark **UNO-Bench** which comprises 1250 human curated samples for omni-modal with **98% cross-modality solvable**, and 630 enhanced samples for uni-modal, across **44 task types** and 6 question types. As shown in Figure.1, the materials are collected from **in-house dataset** with manual annotation which prevents data contamination while better aligning with practical application scenarios. Beyond conventional multiple-choice formats, the evaluation incorporates innovative **multi-step open-ended question** tasks to show a more realistic and discriminative evaluation result on complex reasoning. Besides the self-constructed omni-modal dataset, we construct a uni-modal dataset by aggregating public uni-modal datasets systematically and design a clustering-guided sampling

^{*}Equal contribution.

[†]Project Lead.

Dataset	Omni-modal	Uni-modal	Acc.	Solvable	Source	#Tasks	#QA Pairs	QA Type	Language
MMBench	X	I	-	-	80% private	20	3217	MC	EN/CH
MMAU	X	A	-	-	15% private public	27 20	10000 4000	MC	EN
MVBench	X	V	-	-				MC	EN
OmniBench	I+A	X	75%	90%	public	8	1142	MC	EN
AV-Odyssey	I+V+A	X	91%	99%	public	26	4555	MC	EN
WorldSense	V+A	X	99%	23%	public	26	3172	MC	EN
Daily-Omni	V+A	X	94%	59%	public	6	1197	MC	EN
UNO-Bench-omni	I+V+A	-	100%	98%	90% private	44	1250	MC/MO	EN/CH
UNO-Bench-uni	-	I/V/A	86%	97%	42% private	44	2480	MC	EN/CH

Table 1: Comparison of MultiModal Benchmarks, with I, A, V, and T representing image, audio, video, and text modalities, respectively. It reports on the accuracy of question-answer pairs and the percentage of questions requiring omni-modal solutions, labeled as Acc. and Solvable. The Source category specifies the origin of the materials. Private sources, as opposed to public ones, can prevent data contamination. QA types include MC for multi-choice questions and MO for multi-step open-ended questions. EN and CH denote English and Chinese languages. UNO-Bench includes 1250 human-curated samples in the omni-modal section (referred to as -omni) and 2480 enhanced samples in the uni-modal section (referred to as -uni).

method to achieves both evaluation efficiency and consistency. In this way, our benchmark can have a thoroughly evaluation that requires omni models to keep uni-modal capability and obtain omni-modal capability at the same time.

Main Contributions:

1. Propose the first **UNified Omni** model benchmark, which efficiently assesses both **UNi-modal** and **Omni-modal** understanding capabilities. UNO-Bench verifies the **compositional law between omni-modal and uni-modal capability**. For weaker models, this manifests itself as a "shortboard effect" (where performance is constrained by the weakest modality), while stronger models demonstrate emergent ability across modalities.

2. Establish a **systematic and unified dataset framework** and build a complete manual construction pipeline. This strategy ensures manual participation from the material to the final data, fundamentally avoiding the potential risks of data contamination in automated processing and better preserving the complexity of real-world scenarios. Meanwhile, the Chinese data more accurately reflect the actual application needs of the users in a Chinese context. Therefore, UNO-Bench has achieved a 100% quality inspection pass rate across the 44 task types it covers, and its comprehensive quality significantly surpasses existing datasets.

3. Beyond conventional multiple-choice question type, the evaluation incorporates innovative **Multi-Step Open-Ended Question (MO)** to show a more realistic and discriminative evaluation result on complex reasoning especially for multi-step reasoning across modalities. For automated evaluation, a general scoring model is proposed to support 6 kinds of question types with 95% accuracy on

OOD models and benchmarks.

2 Related Work

2.1 Uni-Modal Benchmarks

Based on large language models, vision language models (VLM) (Bai et al., 2025; Xiaomi, 2025; Zeng et al., 2025) and audio language models (ALM) (Ding et al., 2025; Wu et al., 2025) introduce the general intelligence to vision modality and audio modality respectively. Various uni-modal benchmarks conduct comprehensive evaluations on VLMs (Liu et al., 2024b; Lu et al., 2024a; Wang et al., 2024b,c; Liu et al., 2024d; Mathew et al., 2021; Ouyang et al., 2024; Li et al., 2024a; Wu et al., 2024; Liu et al., 2024c; xAI, 2023; Xiao et al., 2021; Huang et al., 2025; Hu et al., 2025; Fu et al., 2024) and ALMs (Ardila et al., 2019; Wang et al., 2021; Yang et al., 2024; Ao et al., 2024). For image modality, MMBench(Liu et al., 2024b) proposed a systematically designed benchmark to evaluate general image understanding on 20 different tasks. Focused on mathematic, MathVision(Wang et al., 2024b) collected questions from 19 mathematic competitions to evaluate VLMs complex reasoning ability. In addition to above, OCRBench(Liu et al., 2024d) supplied the evaluation on text recognition and document understanding. For video modality, MVBench(Li et al., 2024a) aggregated 11 public video benchmarks and incorporated data enhancement process to cover 20 dynamic video understanding tasks. To complement the long video understanding area, LongVideoBench(Wu et al., 2024) introduces hourly video materials to evaluate the information retrieval ability from long context. For audio modality, MMAU(Sakshi et al., 2025) provides general audio understanding assessment

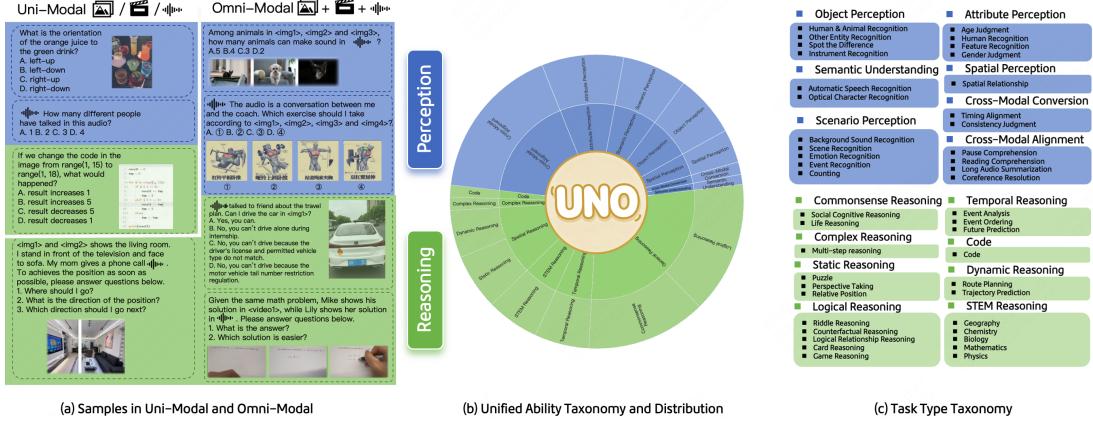


Figure 1: Illustration of the unified ability taxonomy proposed in UNO-Bench. (a) exemplarily presents the way to unify uni-modal and omni-modal samples. (b) shows the three-layer hierarchical structure of the unified ability taxonomy and its distribution in UNO-Bench. Among the 13 abilities, (c) demonstrates the 44 diverse task types that cover a large range of real world scenarios.

across speech, sounds and music domains, featuring diverse audio samples. There's massive uni-modal benchmarks covering diverse model abilities on vision modality and audio modality separately.

2.2 Omni-Modal Benchmarks

Omni models have arisen in recent years(Comanici et al., 2025; Xu et al., 2025b; AI et al., 2025; Li et al., 2025), as the pioneer, Gemini(Comanici et al., 2025) shows a strong ability in understanding both vision and audio, while qwen-3-omni(Xu et al., 2025b) provides leading performance in open-source models. However, there's less omni-modal benchmarks that can evaluate the modality combination across image, video and audio. OmniBench(Li et al., 2024b) inserted audio as a context into the image understanding task and made up an omni-modal benchmark, while the data quality needs further improvement. WorldSence(Hong et al., 2025) emphasized audio-visual data in real world scenarios with high data quality, while most audio-visual questions can be solved by audio or video alone, which can't assess the cross-modality ability. Other datasets focus on audio (Gong et al., 2024) or video (Zhou et al., 2025) and cover limited task types.

Addressing these limitations, we propose a novel and unified benchmark, UNO-Bench, that enables comprehensive model assessment and pushing omni model to the next-level of intelligence.

3 Method

In this section, we first introduce the omni-modal dataset construction pipeline in Section.3.1. The

uni-modal dataset improvement method and a general evaluation dataset compression method to improve the evaluation efficiency are introduced in Section.3.2. Finally, the multi-step open-ended questions are introduced in Section.3.3.

3.1 Omni-modal Dataset Construction

We have established a human-centric data construction pipeline (Figure.2) that efficiently empower human intelligence to produce high-quality and high-diversity dataset.

3.1.1 Model Ability Taxonomy

Through cumulative experiences on multimodal evaluation from both model-side and user-side, we summarize the capabilities of uni-modal and omni-modal into a unified model ability taxonomy. As shown in Figure.1, the omni model's capabilities are systematically categorized into two primary dimensions: Perception and Reasoning. Detailed definitions and examples can be found in the Appendix.D.

Perception dimension structured through seven recognition types: Object Perception, Attribute Perception, Scenario Perception, Spatial Perception, Cross-Modal Conversion, Semantic Understanding. In addition, we incorporate Cross-Modal Alignment evaluation to assess information synchronization across modalities.

The perception dimension fuses visual and audio modalities. For example, Scenario Perception includes the recognition of visual scenes and the judgment of audio scenes.

Reasoning dimension extends conventional reasoning categories (General, STEM, Code) with:

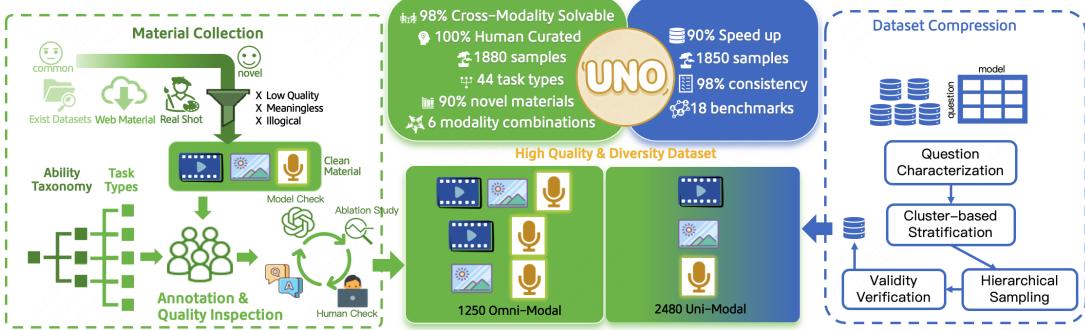


Figure 2: Dataset Construction Pipeline includes human crafted process (left side) and automated data compression (right side). First, we collect diverse and novel materials to prevent data contamination. Second, with the proposed unified ability taxonomy, human annotators including experts will craft questions, answers and record audios in real-world scenarios. Finally, with model checking, ablation study and human experts revision, we achieves high quality and diversity dataset. Regarding automated data compression, we use model performances as a feature of the sample, employing clustering and resampling to achieve a 90% dataset compression.

Spatial Reasoning (including Static Reasoning and Dynamic Reasoning), Temporal Reasoning, and Complex Reasoning (multi-conditional, multi-step problem solving).

The unified ability taxonomy combines uni-modal and omni-modal abilities which provides a comprehensive measurement that is particularly critical for advanced omni models. Based on this taxonomy, we create a diversity dataset with 44 task types distributed as shown in Figure.1.

3.1.2 Material Collection

In both data quality checks and experimental results, we found that natural audio-visual synchronized video data contains a large amount of information redundancy, only a few videos require both audio and vision modality simultaneously. Therefore, we begin with carefully designed material collection. Our materials have the following three characteristics: **a. Diverse Sources**. The majority of our materials are real-world photos and videos collected through crowdsourcing, and another portion sourced from copyright-free website. Additionally, a small fraction comes from high-quality public datasets such as MMVU([Zhao et al., 2025](#)), LongVideoBench([Wu et al., 2024](#)), and VideoVista([Chen et al., 2025](#)). **b. Rich and Diverse Topics**. Our materials cover a broad spectrum of subjects, including society, culture, art, life, literature, science, and so on. **c. Live-Recorded Audio**. Apart from background sounds and music, all dialogue is recorded by human speakers. With over 20 participants in the recording process, the audio features are rich and closely reflect the diverse vocal characteristics of the real world.

Finally, we conduct material filtering. Eliminate meaningless, illogical, and low-quality materials, and categorize the remaining materials by theme to create a material library. Additionally, label the materials with more detailed information such as subject, event, scene, and style to facilitate subsequent annotators to quickly find matching materials.

3.1.3 QA Annotation

Our annotators consist of human experts and high-quality crowd-sourced users. Human experts have extensive experience in cross-modal data construction and annotation, a deeper understanding of model capabilities, and thus ensure higher professionalism and specificity in the data they construct. Most crowd-sourced users are college students with rich experience in multi-modal model interactions and diverse professional backgrounds, providing data with better authenticity and diversity.

First, annotators clarify the required image/video features based on task type definitions and filter appropriate materials from existing libraries using tags. **Second**, following data construction requirements, they then design prompts and corresponding answers. **Third**, to enhance data authenticity, all dialogue audio is recorded manually. Through this workflow, we ultimately generate complete QA pairs encompassing three modalities: visual, auditory, and textual.

Compared to conventional methods limited to human intervention only during the quality assurance phase, our pipeline integrates a "Human-in-the-Loop" approach, ensuring continuous manual involvement from the initial data sourcing to the final output. This methodology not only prevents

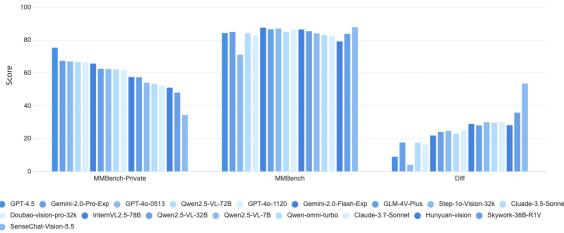


Figure 3: Comparison result of privatization improvement. After improvement, the performances among models are more distinguishable.

data leakage but also more accurately simulates real-world scenarios. Furthermore, the manually curated Chinese dataset genuinely captures user requirements in a Chinese linguistic context, compensating for the shortcomings of most existing English-centric datasets.

3.1.4 Quality Inspection

To ensure the data quality, we have established a multi-stage, cyclically validated quality assurance system composed of automated tools and manual review. Each question undergoes at least three rounds of independent quality inspection to maximize data quality. **First**, a preliminary model check is conducted to filter out cases with ambiguous questions, non-unique answers, or those that do not conform to the task type. **Second**, through modality ablation experiments, we remove one modality of information from the QA pair to see if the model can answer based solely on the remaining information. If the question becomes unsolvable or ambiguous after removing any one modality, it proves the cross-modality solvability of the data. **Finally**, manual quality inspection and revision are performed.

3.2 Uni-modal Dataset Improvement

3.2.1 Quality Improvement

Existing public uni-modal datasets are bothered by data leakage issue(Xu et al., 2024). To verify the influence, we adopt privatization improvement on the widely used public dataset MMBench(Liu et al., 2024b). As shown in Figure.3, the performance of models have better distinguishability after dataset improvement, reflecting the true capability differences between models. Therefore, for uni-modal data, we also follow the aforementioned construction process for self-construction datasets. In addition to self-constructed data, we also selected some

multimodal data from public datasets to supplement in terms of capability items and data types. (Data mainly comes from AV-Odyssey(Gong et al., 2024) and WorldSense(Hong et al., 2025), accounting for 11% of the total). The specific selection logic is as follows:

Comprehensiveness: In terms of capabilities, focus on supplementing the perception part with a relatively low self-construction proportion, while also adding some reasoning questions; in terms of data types, prioritize selecting the video plus audio modality combination with a lower self-construction proportion for supplementation, followed by image plus audio.

Diversity: Supplement material types not covered in self-construction data to enhance diversity.

High Quality: Pay attention to the quality of datasets (whether uni-modal answers are reasonable and accurate).

Discriminative: Pay attention to the performance of this dataset on the model, and remove overly difficult subsets with little discrimination.

3.2.2 Dataset Compression

Regarding the existing large-scale uni-modal benchmarks, to reduce the evaluation cost of large-scale models, we designed a **clustering-guided hierarchical sampling (CGHS)** method as shown in Figure.2. This method achieves efficient compression and high evaluation consistency through the following steps:

Question Characterization: Represent each question as an x -dimensional vector, where dimensions correspond to scores from different models on that question.

Cluster-based Stratification: Utilize the Kmeans++(Arthur and Vassilvitskii, 2007) algorithm to categorize questions into k clusters, each representing a "model performance similar" question type (e.g., easy questions, difficult questions, etc.).

Hierarchical Sampling: Determine the sample size for each stratum based on cluster size proportions, and construct the final evaluation subset through simple random sampling.

Validity Verification: Use Margin of Error (MoE) and Root Mean Square Error (RMSE) to measure theoretical and practical errors respectively. Verify the consistency between model rankings and scores using Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC).

To ensure statistical stability, we repeat the above steps by using 5 random splits and performing 10-fold cross-validation. This approach identifies the optimal sample size via cost-benefit curve analysis, leading to a reduction in evaluation costs by over 90% while preserving accuracy, as shown in Figure.9.

3.3 Multi-Step Open-Ended Questions

3.3.1 Question Type Definition

Evaluating the multi-step reasoning capabilities of omni models presents a significant challenge. Real-world problems require models to integrate multimodal information and execute a sequence of logical steps. However, current automated benchmarks, often relying on Outcome Reward Models (ORMs), typically provide only a binary pass/fail judgment. This approach fails to distinguish between a model that completes 80% of a task and one that fails at 20%, a crucial gap that human evaluators easily perceive. While alternatives like Process Reward Models (PRMs)(Lightman et al., 2023) or multi-turn dialogues(Reddy et al., 2019) exist, they are hampered by high implementation difficulty, low accuracy, or poor efficiency. Moreover, the prevalence of multiple-choice formats in existing benchmarks is unrepresentative of real-world, open-ended user queries and can mask model deficiencies.

To address these issues, we propose an innovative Multi-Step Open-Ended Question (MO) type, designed for granular and realistic assessment. In the construction of MO dataset, complex problems are first deconstructed by human experts into a series of progressive, interdependent sub-questions. Each sub-question is assigned a score based on its importance, summing to a total of 10 points. During testing, all sub-questions are posed in a single turn, requiring the model to generate a step-by-step open-ended response. This method allows us to precisely quantify how far along a complex reasoning chain a model can proceed, offering a more accurate and insightful measure of its true capabilities. An example is shown in Figure.4.

3.3.2 General Scoring Model

Beside the dataset construction, multi-step open-ended question introduces a new challenge of automated evaluation. To overcome this obstacle, we propose a general scoring model that supports multi-choice question, single-step open-ended question and multi-step open-ended question at the same time. Since the task is to compare the

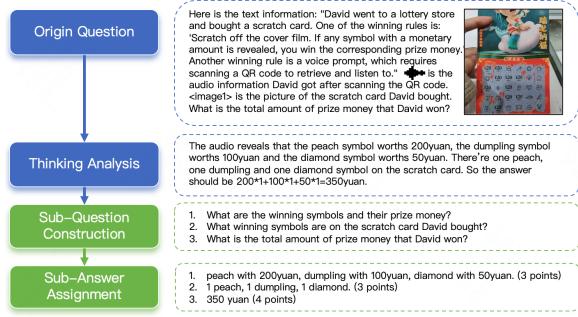


Figure 4: Example of multi-step open-ended questions.

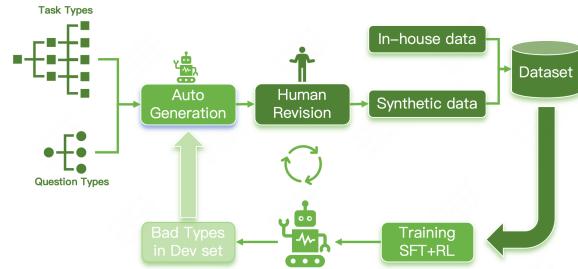


Figure 5: Training dataset construction and model training pipeline for general scoring model.

target answer and the predicted answer, we use Qwen3-14B(Yang et al., 2025) as backbone and curate a training dataset as illustrated in Figure.5. One of the critical way to improve accuracy is to group questions into finer types and define appropriate criteria for each types. See details in Appendix.C. Through the human-in-the-loop dataset curation, the scoring model achieves 95% accuracy in out-of-distribution models and benchmarks.

Experiments in Section.4.4.1 show that compared with single-step evaluation method (e.g. multiple-choice questions), multi-step open-ended questions can effectively observe the ability decay of models in long-chain reasoning, providing a more realistic difficulty for advanced models with stronger discrimination.

4 Experiment and Analysis

4.1 Experiment Setting

We evaluate omni models that support text, visual, and audio inputs simultaneously, including open-source models: Qwen-3-Omni-30B-A3B-Instruct(Xu et al., 2025b), Qwen-2.5-Omni-3B, Qwen-2.5-Omni-7B(Xu et al., 2025a), Baichuan-Omni-1.5(Li et al., 2025), MiniCPM-O-2.6(Yao et al., 2024), and Ming-lite-Omni-1.5(AI et al., 2025), as well as closed-source models: Gemini-2.5-Pro, Gemini-2.5-Flash, and Gemini-

Table 2: Cross-modal benchmark results on multimodal models.

Model	Audio	Visual	Omni-MC	Omni-MO
Qwen2.5-Omni-3B	54.40	42.67	27.80	24.76
MiniCPM-O-2.6	56.50	42.27	28.60	23.76
Ming-lite-Omni-1.5	58.30	46.28	28.90	25.48
Baichuan-Omni-1.5	54.10	44.66	29.70	21.04
Qwen2.5-Omni-7B	60.20	50.68	32.60	27.72
Qwen3-Omni-30B-A3B	79.40	63.29	42.10	37.08
Gemini-2.0-Flash	70.70	62.76	44.90	38.56
Gemini-2.5-Flash	79.50	69.54	54.30	47.08
Gemini-2.5-Pro	88.40	78.67	70.90	57.32

Table 3: Cross-modal analysis on model ability taxonomy.

Model	Perception			Reasoning				Overall
	Cross-modal Alignment	Cross-modal Recognition	Overall	General Reasoning	Sequential Reasoning	Spatial Reasoning	Overall	
Qwen2.5-Omni-3B	29.84	35.94	33.09	20.65	50.00	20.83	23.98	27.80
MiniCPM-O-2.6	26.70	30.88	28.92	26.62	42.42	26.67	28.40	28.60
Ming-lite-Omni-1.5	28.80	35.94	32.60	24.38	43.94	24.17	26.53	28.90
Baichuan-Omni-1.5	30.89	32.26	31.62	25.87	45.45	28.33	28.57	29.70
Qwen2.5-Omni-7B	38.22	36.41	37.25	28.11	43.94	26.67	29.59	32.60
Qwen3-Omni-30B-A3B	53.40	45.16	49.02	38.06	53.03	26.67	37.41	42.10
Gemini-2.0-Flash	43.98	49.77	47.06	45.02	57.58	31.67	43.71	44.90
Gemini-2.5-Flash	56.02	50.69	53.19	61.44	68.18	27.50	55.27	54.30
Gemini-2.5-Pro	74.35	70.05	72.06	75.62	84.85	45.00	70.41	70.90

2.0-Flash(Comanici et al., 2025). To have a fair comparison between instruct model and thinking model, we adopt similar way in Qwen3(Xu et al., 2025b) that limits thinking budget to 128 tokens. We apply this restriction to Gemini-2.5-Pro and disable the thinking mode for both Gemini-2.5-Flash and Gemini-2.0-Flash. All the other model integrations strictly adhere to official implementations. In video processing, each model receives raw video and performs frame sampling according to its own sampling strategy.

In the following sections, we conduct comprehensive experiments on UNO-Bench and try to answer the following questions:

1. How about the performance of current omni models and their shortage abilities?
2. What is the relationship of uni-modal ability and omni-modal ability?
3. Could the UNO-Bench provide an effective "All in One" evaluation for the omni model?

4.2 Model Performance

4.2.1 Overall Analysis

Our main evaluation, summarized in Table 2, reveals a clear performance hierarchy where proprietary models, particularly Gemini-2.5-Pro, establish the state-of-the-art across all benchmarks.

Meanwhile, progress within the open-source community is notable, with increased model scale and more training data, exemplified by Qwen3-Omni-30B, leading to substantial improvements. Furthermore, we observe a strong positive correlation between a model’s performance on the foundational Audio and Visual tasks and its scores on the more demanding Omni benchmarks, suggesting that robust uni-modal perception is a prerequisite for advanced omni-modal understanding.

On the **Omni-MC** (Multiple-choice) benchmark, which evaluates omni-modal comprehension, smaller open-source models exhibit performance marginally surpassing the random guess baseline (25.00), achieving scores between 27.80 and 29.70. The larger Qwen3-Omni-30B marks a significant leap, with a score of 42.10 that approaches the performance of entry-level proprietary models like Gemini-2.0-Flash (44.90). Nevertheless, a substantial performance deficit persists when compared to the leading Gemini-2.5-Pro (70.90). This gap highlights the profound difficulty of advanced omni-modal comprehension, even in a multiple-choice format.

The **Omni-MO** (Multi-step Open-ended) benchmark presents a considerably greater challenge, as evidenced by the universal and marked degradation in performance for all models relative to their Omni-MC scores. This degradation reveals a sys-

temic limitation in multi-step omni-modal reasoning. For instance, the leading model, Gemini-2.5-Pro, attained a score of merely 57.32 on this benchmark, reflecting a decline of 13.58 points relative to its performance on the Omni-MC task. In comparison, the highest-scoring open-source model, Qwen3-Omni-30B, achieved only 37.08 points.

To dissect the models’ core capabilities, we perform a fine-grained analysis based on our proposed ability taxonomy, with detailed results presented in Table 3.

In perception, a notable trend emerges: while smaller models find Recognition easier than Alignment, more powerful models like Qwen-3-Omni-30B-A3B and the Gemini-2.5 series exhibit stronger Alignment capabilities. This suggests that advanced models develop a more sophisticated grasp of inter-modal relationships. Among open-source models, Qwen-3-Omni-30B-A3B achieves the highest perception score (49.02). Gemini-2.5-Pro significantly leads overall, with both its Alignment (74.35) and Recognition (70.05) scores surpassing 70.

In reasoning, Spatial Reasoning is consistently the most challenging task across all models. Even the top-performing Gemini-2.5-Pro only achieves 45.00. Notably, Baichuan-Omni-1.5 demonstrates the best spatial reasoning among open-source models with a score of 28.33. For General and Sequential Reasoning, the new Qwen-3-Omni-30B-A3B establishes itself as the open-source leader.

Overall, reasoning proves to be a more challenging frontier than perception. This is highlighted by the performance gap between the leading proprietary model, Gemini-2.5-Pro, and the best open-source model, Qwen-3-Omni-30B-A3B. The disparity is 23.04 points in Perception (72.06 vs. 49.02) but widens to a more substantial 33.00 points in Reasoning (70.41 vs. 37.41). This indicates that advanced reasoning remains a key differentiator and a primary bottleneck for current multimodal models.

4.2.2 Top-tier Analysis

Why Gemini-2.5-Pro performances significantly better than other models? We attempt to provide an analysis and some hypotheses. On one hand, it stems from the leading uni-modal understanding ability. On the other hand, regarding the reasoning mechanism, Gemini is equipped with audio captioning functionalities as indicated in the technical report(Comanici et al., 2025), and illustrated in

Question

Based on the audio and image content, determine which of the following foods might be my favorite? Please select the correct answer from the options below:

A. Fish meat burger
B. Deep-sea cod burger
C. Juicy Wagyu beef burger
D. Fish meat wrap



Gemini-2.5-Pro

Based on the audio and image content, the following conclusions can be drawn:
Audio Information: The speaker mentions that the top of the picture shows her favorite type of food. However, she does not eat beef and believes that fish tastes the best.
Image Information: The top of the picture shows a hamburger (typically a beef burger).
Comprehensive Judgment: ... Answer: B.

Qwen3-Omni

D. Fish meat wrap

Qwen2.5-Omni

Hmm... since you like fish the most, it should probably be a fish burger or a fish wrap. But judging from the picture, the top part looks like a hamburger, and the bottom part is a roujiamo (Chinese meat sandwich), which has meat inside. A fish roujiamo or fish wrap would better suit your preference for fish. So I think the answer might be A. Fish roujiamo. If you have any other thoughts, feel free to share them with me.

Figure 6: An example of Omni-modal Evaluation Result. Gemini-2.5-Pro displays audio captions to assist in solving the problem.

Figure 6. It can also naturally incorporate audio content as a foundation for reasoning. Existing open-source models, due to their smaller size, lack reasoning processes in a multimodal context. Limited reasoning mostly relies on text or images, with few involving specific audio content.

However, does Gemini performance good enough? To answer this question, we invited human experts to conduct a competition. It is important to note that, unlike the dataset annotators, the invited human experts have not been exposed to the question or the answer.

Finding 1. Gemini-2.5-Pro has reached human comparable perception ability in omni-modal understanding, yet there remains a gap in its reasoning performance. Compared to human experts, Gemini-2.5-Pro only has 8.3% less performance, which shows humanity comparable intelligence. The comparison of scores for specific ability items can be seen in Figure 7. Upon examining ability analysis, we observe an intriguing phenomenon: humans are more proficient in reasoning as opposed to perception (81.3% compared to 74.3%), which contrasts with the model’s performance. By interviewing various human experts, it becomes evident that humans might miss some information in video or audio formats, and their world knowledge is more limited compared to large language models.

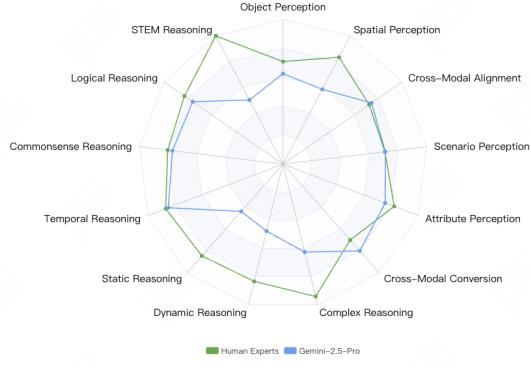


Figure 7: The competition between Human experts and Gemini-2.5-Pro. Gemini-2.5-Pro shows comparable perception capability but lower reasoning capability.

4.3 Uni-Modal v.s. Cross-Modal

To investigate the relationship of uni-modal and omni-modal understanding ability, we conduct regression analysis and ablation experiments. Thanks to the unified ability taxonomy and the high quality of omni-modal samples in UNO-Bench, we find some interesting observations.

Finding 2. Compositional Law: the effectiveness of omni-modal capability is related to the product of the performances of individual modalities by a power-law. Observing the results in Table 2, we identify a strong correlation between a model’s omni-modal performance and its uni-modal capabilities.

We derive the Compositional Law from a general functional form by applying two simplifying principles dictated by the omni-modal tasks proposed in our UNO-Bench.

1. General Model & Task Constraints. We begin by positing that omni-modal performance $\mathcal{P}_{\text{Omni}}$ is a function of uni-modal performances \mathcal{P}_A and \mathcal{P}_V . A general, separable form can be written as:

$$\mathcal{P}_{\text{Omni}} = f_A(\mathcal{P}_A) + f_V(\mathcal{P}_V) + f_I(\mathcal{P}_A, \mathcal{P}_V) + b \quad (1)$$

where f_A, f_V represent independent path contributions and f_I the interaction. Our benchmark is designed around an **indivisibility constraint**: a task is unsolvable if either modality is absent. This imposes a boundary condition:

$$\lim_{\mathcal{P}_A \rightarrow 0} \mathcal{P}_{\text{Omni}} = b \quad \text{and} \quad \lim_{\mathcal{P}_V \rightarrow 0} \mathcal{P}_{\text{Omni}} = b \quad (2)$$

Satisfying Eq. 2 requires that the independent path contributions be null, i.e., $f_A \equiv 0$ and $f_V \equiv 0$.

Consequently, Eq. 1 reduces to a form governed solely by interaction:

$$\mathcal{P}_{\text{Omni}} = f_I(\mathcal{P}_A, \mathcal{P}_V) + b \quad (3)$$

2. Interaction Form & Symmetry. We model the interaction term f_I using a general power-law form, which is effective for capturing scaling relationships:

$$f_I(\mathcal{P}_A, \mathcal{P}_V) = C \cdot \mathcal{P}_A^{\alpha_a} \mathcal{P}_V^{\alpha_v} \quad (4)$$

We then posit a **fusion symmetry** assumption: in end-to-end omni models, the fusion mechanism does not inherently favor one modality over another(Xu et al., 2025b; Yao et al., 2024), implying symmetric scaling behavior. This leads to $\alpha_a = \alpha_v = \alpha$. Substituting this into Eq. 4 yields:

$$f_I = C \cdot (\mathcal{P}_A \times \mathcal{P}_V)^\alpha \quad (5)$$

Combining this with Eq. 3, we arrive at the **Omni-modal Compositional Law**:

$$\mathcal{P}_{\text{Omni}} = C \cdot (\mathcal{P}_A \times \mathcal{P}_V)^\alpha + b \quad (6)$$

where α is the synergistic exponent, C is a scaling coefficient, and b is a baseline bias.

A non-linear regression on data from leading models (Figure 8) yields the precise empirical formula:

$$\mathcal{P}_{\text{Omni}} \approx 1.0332 \cdot (\mathcal{P}_A \times \mathcal{P}_V)^{2.1918} + 0.2422 \quad (7)$$

This model demonstrates an exceptional fit, with a coefficient of determination (R^2) of **0.9759**. This indicates our law explains over 97% of the variance in omni-modal performance. Analysis of the fitted parameters reveals a clear transition from limited gains to emergent capabilities, driven by the super-linear nature of the law.

Power-law Synergy and Emergent Ability. The exponent $\alpha \approx 2.19$ is the most critical discovery, revealing a powerful **Power-law synergy**. Because $\alpha > 1$, the function is convex, meaning the rate of performance gain accelerates. This explains the transition from a "short-board effect" to an "emergent ability" seen in Figure 8:

- **Limited Gains at Low Performance:** For models with weaker uni-modal abilities (e.g., MiniCPM-O), the curve is relatively flat. Small improvements in the product of uni-modal scores yield only marginal gains in

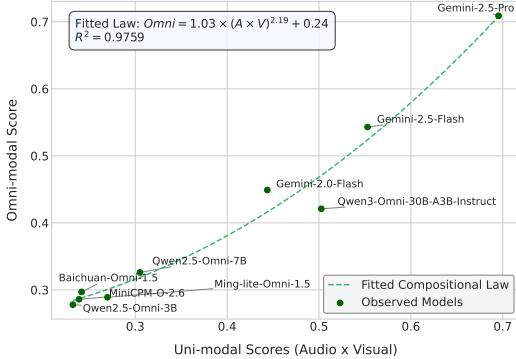


Figure 8: The Compositional Law of Omni-modal Performance. Observed omni-modal scores (dots) versus the product of their uni-modal scores. The dashed line represents our fitted law (Eq. 7), which shows a remarkable alignment with the empirical data ($R^2 = 0.9759$). The convex, accelerating curve visualizes the Power-law synergy.

omni-modal performance. This can be seen as a "short-board effect," where the system is not yet capable of effectively leveraging the combined inputs.

- **Emergent Ability at High Performance:** As uni-modal abilities strengthen (e.g., moving towards Gemini-2.5-Pro), the curve steepens dramatically. The same amount of improvement in the uni-modal product now yields a much larger increase in omni-modal performance. This accelerating return on investment is the quantitative signature of **emergence**, where stronger foundational skills unlock disproportionately powerful combined capabilities.

Interpreting the Coefficients and Benchmark Coherence. The other parameters complete the picture. The bias term $b \approx 0.2422$ acts as a performance floor. As uni-modal performances approach zero, the system's output converges to this value, which is strikingly close to the 0.25 random-guess accuracy of our benchmark. The scaling coefficient $C \approx 1.0332$, being remarkably close to unity, indicates a harmonious and naturally scaled system. We attribute this harmony not only to the models' intrinsic fusion mechanisms but also to the coherent design of our benchmark itself.

Additional fitted models are presented in Appendix B. We argue that our proposed model is the most natural and interpretable among them. Interestingly, most models indicate that the visual modality tends to offer greater benefits than the

audio modality. This phenomenon may be attributable to the relatively weaker visual capabilities of models at the current stage of development.

Next, we conduct ablation studies to dig dive about the enhancement from vision and audio modality respectively.

4.3.1 Ablation Visual Understanding

To quantify the contribution of visual information, we conducted an ablation study with three settings: audio-only (Audio), audio plus high-quality textual captions of the visual scene (+ Caption), and the full audio-visual input (+ Visual). The captions were generated by Gemini-2.5-Pro to ensure descriptive richness. Results are detailed in Table 4.

With only audio input, most models' performance drops to a level near random guessing (around 20-28%), confirming the critical role of visual context. A notable exception is Gemini-2.5-Pro, which scores 40.34, suggesting an ability to leverage linguistic cues or shortcuts within the questions even without visual data.

The introduction of Caption information yields significant but highly variable performance gains. Powerful models like the Gemini series and Qwen-3-Omni-30B-A3B demonstrate a substantial leap in performance (gains of 20-25 points), showcasing their strong ability to reconstruct scenes from textual descriptions. In contrast, models like MiniCPM-O-2.6 and Ming-lite-Omni-1.5 show minimal improvement, indicating a weaker capacity for this text-to-vision reasoning.

Comparing Caption against full Visual input reveals a fascinating dichotomy. For the most capable model, Gemini-2.5-Pro, direct visual information provides a clear advantage over captions (70.90 vs. 65.10), proving that raw visual data contains nuances that text cannot fully capture. However, for several other models, including Gemini-2.0-Flash and the powerful Qwen-3-Omni-30B-A3B, performance with captions is surprisingly on par with, or even slightly exceeds, that with direct visual input. This suggests that for these models, the language processing pathway may be more adept at extracting semantic meaning than their own visual encoders, highlighting a potential imbalance in their multimodal processing capabilities.

4.3.2 Ablation Audio Understanding

To isolate the impact of auditory information, we evaluated models under three conditions:

Table 4: Ablation of visual understanding ability.

Model	Perception			Reasoning			Overall		
	Audio	+ Caption	+ Visual	Audio	+ Caption	+ Visual	Audio	+ Caption	+ Visual
Qwen-2.5-Omni-3B	17.76	29.13	33.09	20.07	21.43	23.98	19.12	24.60	27.80
MiniCPM-O-2.6	29.44	29.61	28.92	27.21	29.93	28.40	28.13	29.80	28.60
Ming-lite-Omni-1.5	26.28	31.07	32.60	23.13	21.43	26.53	24.42	25.40	28.90
Baichuan-Omni-1.5	22.14	32.04	31.62	23.81	26.70	28.57	23.12	28.90	29.70
Qwen-2.5-Omni-7B	22.14	30.10	37.25	20.41	25.34	29.59	21.12	27.30	32.60
Qwen-3-Omni-30B-A3B	27.01	46.84	49.02	18.71	39.63	37.41	22.12	42.60	42.10
Gemini-2.0-Flash	25.55	44.17	47.06	29.76	45.58	43.71	28.03	45.00	44.90
Gemini-2.5-Flash	22.63	49.03	53.19	29.08	53.23	55.27	26.43	51.50	54.30
Gemini-2.5-Pro	37.71	63.83	72.06	42.18	65.99	70.41	40.34	65.10	70.90

visual-only (Visual), visual plus transcribed audio (+Caption), and the full audio-visual input (+Audio). We further divided the audio into two categories: Environment sounds, for which we provided textual descriptions, and Speech, for which we provided ASR transcripts. All transcriptions were generated by a powerful multimodal model. The results are presented in Table 5.

The Visual-only setting results in significantly lower performance across all models, with Overall scores ranging from 21.20 to 33.70. This confirms the critical role of auditory context in multimodal understanding. The introduction of textual audio descriptions (+Caption) substantially boosts performance across the board. The improvement is particularly dramatic for high-capacity models like Gemini-2.5-Pro (+31.0 points Overall) and Qwen-3-Omni-30B-A3B (+17.4 points Overall), demonstrating their strong ability to integrate textual information.

The comparison between +Caption and +Audio reveals crucial insights into the models’ raw audio processing capabilities. In environmental sound scenarios, understanding raw audio remains a significant challenge for most open-source models. For instance, Qwen-2.5-Omni-3B, MiniCPM-O-2.6, and Ming-lite-Omni-1.5 all exhibit considerably higher performance with textual descriptions (+Caption) than with the original audio (+Audio). This suggests that their audio encoders struggle to extract meaningful features from complex non-speech sounds, making them prefer clean textual summaries. In contrast, the most capable models—Gemini-2.5-Pro, Gemini-2.5-Flash, and Qwen-3-Omni-30B-A3B—demonstrate superior audio understanding by scoring higher in the +Audio setting, indicating they can extract richer information directly from the audio signal than is present in the provided caption.

In conversational (Speech) scenarios, the results are more nuanced. The top-performing Gemini-2.5-

Pro shows a substantial advantage with raw audio over ASR transcripts (+Audio 72.16 vs. +Caption 66.00), indicating it effectively leverages paralinguistic cues such as tone, emotion, and prosody that are lost in transcription. Conversely, several other models, including the Qwen series and MiniCPM-O-2.6, perform slightly better with ASR transcripts (+Caption) than with raw audio. This points to a common bottleneck where imperfections in their audio encoders are a greater liability than the information lost during ASR, making clean text a more reliable input. Notably, Gemini-2.5-Flash achieves nearly identical scores in both settings, suggesting its ASR and audio understanding capabilities are exceptionally well-aligned.

4.4 Benchmark Analysis

In this section, we verify the effectiveness of UNO-Bench on three aspects, the performance of multi-step open-ended question, the performance of dataset compression and the benchmark comparison with other open-source benchmarks.

4.4.1 Multi-Step Open-Ended Question Analysis

In this work, we introduce a new type of evaluation method, multi-step open-ended question, which effectively assess the complex reasoning ability, especially appears in cross-modality understanding.

As shown in Table 6, the experimental results on our multi-step open-ended questions reveal a clear performance stratification among models. Gemini-2.5-Pro establishes itself as the top-tier model with an overall score of 57.32, with Gemini-2.5-Flash (47.08) and Gemini-2.0-Flash (38.56) forming a distinct second tier. Among open-source models, Qwen-3-Omni-30B-A3B emerges as the clear leader with a score of 37.08, significantly outperforming smaller-scale models like Qwen-2.5-omni-7B (27.72). This starkly illustrates that both advanced architecture and model scale are pivotal fac-

Table 5: Ablation of audio understanding ability.

Model	Environment			Speech			Overall		
	Visual	+Caption	+Audio	Visual	+Caption	+Audio	Visual	+Caption	+Audio
Qwen-2.5-Omni-3B	26.28	41.03	34.62	24.76	26.66	26.54	25.00	28.90	27.80
MiniCPM-O-2.6	26.92	39.74	34.62	28.08	28.44	27.49	27.90	30.20	28.60
Ming-lite-Omni-1.5	31.41	43.59	35.26	22.27	25.59	27.73	23.70	28.40	28.90
Baichuan-Omni-1.5	25.64	32.05	28.85	23.70	23.58	29.86	24.00	24.90	29.70
Qwen-2.5-Omni-7B	30.77	41.03	37.18	24.41	33.06	31.75	25.40	34.30	32.60
Qwen-3-Omni-30B-A3B	32.05	48.08	48.72	23.58	41.23	40.88	24.90	42.30	42.10
Gemini-2.0-Flash	25.00	48.08	45.51	22.87	48.93	44.79	23.20	48.80	44.90
Gemini-2.5-Flash	17.95	48.72	49.36	21.80	55.09	55.21	21.20	54.10	54.30
Gemini-2.5-Pro	32.69	57.69	64.10	33.89	66.00	72.16	33.70	64.70	70.90

tors for success in complex, multi-turn multimodal tasks.

As the depth of questions increases from Q1 to Q3+, most models exhibit a general decline in performance, confirming the effectiveness of our dataset’s progressive difficulty. For instance, the leading open-source model, Qwen-3-Omni-30B-A3B, sees its overall score drop from 18.08 on the first question (Q1) to 14.18 (Q2) and further to 11.42 (Q3+). This decay highlights a common challenge for current models in handling long-range dependencies, maintaining conversational context, and performing multi-step reasoning. However, a notable exception is Gemini-2.5-Pro, whose performance on the second question (Q2) surpasses its score on the first (24.48 vs. 23.44), before declining on subsequent questions. This unique pattern suggests a superior ability to utilize the context from the initial turn to enhance its understanding and response in the subsequent turn, a capability not observed in other models.

Reasoning ability remains the key bottleneck that differentiates model performance. For all open-source models and the lower-tier Gemini models, scores on Perception tasks are considerably higher than on Reasoning tasks. The gap is particularly pronounced for Qwen-3-Omni-30B-A3B, which scores 53.8 in Perception but only 32.9 in Reasoning. This indicates that while these models have developed solid foundational perception capabilities, converting this perceptual input into complex logical or causal reasoning remains a major hurdle. Interestingly, Gemini-2.5-Pro is the only model that defies this trend, achieving a higher score in Reasoning (58.1) than in Perception (54.2). This exceptional result demonstrates that state-of-the-art models are beginning to overcome the reasoning bottleneck, showcasing advanced cognitive abilities that are on par with, or even exceed, their perceptual skills. The design of our

dataset successfully magnifies this critical capability gap between the SOTA and other models.

4.4.2 Dataset Compression

We design a cluster-guided stratified sampling to compress the scale of benchmark. To evaluate the consistency of model ranking and the best size of compression data size, we conduct several experiments to analysis.

The baseline data set consists of 8000 samples including 18 open-source benchmarks(e.g. MathVista and MMAU, details see appendix A) and 20 models evaluation results on them, which split into 12/8 on models as training/test set. Kmeans++(Arthur and Vassilvitskii, 2007) is used to cluster with K=48. To eliminate the random factor, we conduct 5-fold settings and evaluate 10 times on each setting.

To verify the compression performance, we define these metrics: Spearman’s Rank Correlation Coefficient(SRCC) for ranking consistency, Pearson’s Linear Correlation Coefficient(PLCC) for linear value consistency, Root Mean Square Error(RMSE) for numerical precision, Margin of Error(MoE) for quantifying estimation uncertainty, and Confidence Interval Coverage(CIC) for statistical reliability.

The experimental result is shown in Figure.9. At a 10% sampling rate, our method achieved excellent results on test-set. Both **SRCC** and **PLCC** exceeded 0.98, indicating near-perfect preservation of ranking and value relationships. The **RMSE** was below 0.02 with a corresponding **MoE** of 0.024; together, these values signify high numerical precision and a tight estimation range. Furthermore, the **CIC** was approximately 95%, confirming the statistical unbiasedness of the sample.

4.4.3 Benchmark Comparison

To ensure the quality of dataset, we conduct quality check on 10%-20% random samples in each bench-

Table 6: Performance on Multi-Step Open-Ended Questions.

Model	Perception				Reasoning				Overall			
	Q1	Q2	Q3+	All	Q1	Q2	Q3+	All	Q1	Q2	Q3+	All
Baichuan-Omni-1.5	15.4	8.2	5.33	25.2	9	7.25	5.75	18.9	10.28	7.44	5.7	20.16
MiniCPM-O-2.6	20.0	6.2	11.33	29.6	9.05	9.55	8.02	22.3	11.24	8.88	8.43	23.76
Qwen-2.5-Omni-3B	19.8	12.2	5.33	33.6	10.7	7.2	8.86	22.55	12.52	8.2	8.42	24.76
Ming-lite-Omni-1.5	19.6	12.4	4.67	33.4	10.9	8.4	7.92	23.5	12.64	9.2	7.52	25.48
Qwen-2.5-Omni-7B	20.2	15.0	12.0	38.8	12.15	8.99	7.83	24.95	13.76	10.2	8.35	27.72
Qwen-3-Omni-30B-A3B	25.0	22.8	20.0	53.8	16.35	12.01	10.19	32.9	18.08	14.18	11.42	37.08
Gemini-2.0-Flash	25.2	19.4	14.67	49.0	15.5	14.05	13.02	35.95	17.44	15.12	13.22	38.56
Gemini-2.5-Flash	31.6	22.6	12.0	57.8	18.35	17.35	16.42	44.4	21.0	18.4	15.87	47.08
Gemini-2.5-Pro	25.6	22.2	21.33	54.2	22.9	25.05	19.43	58.1	23.44	24.48	19.67	57.32

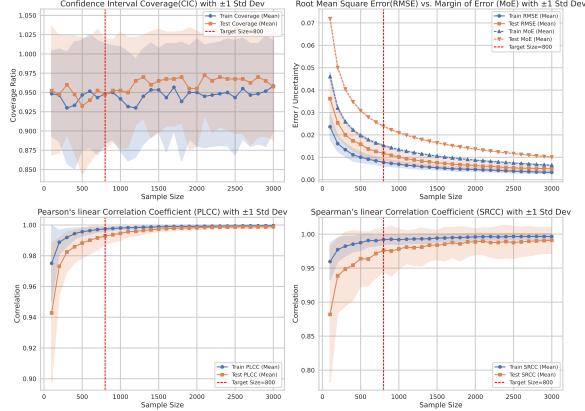


Figure 9: Data compression Performance.

marks. As shown in Table.1, An effective benchmark must provide both a clear performance ladder and a meaningful difficulty range, and our UNO-Bench is engineered to deliver on both fronts. It excels in discriminability, establishing substantial and remarkably linear intervals of ~10-12 points between adjacent models. This superior discriminability comes from a well-calibrated difficulty. UNO-Bench creates a vast 31.9-point performance gap between the top and bottom models, effectively separating their capabilities. This approach avoids the pitfall of being universally difficult, a problem seen in AVOdyssey where all models are compressed into a narrow, low-scoring band (34-45). By combining a structured performance ladder with a balanced challenge, UNO-Bench serves as a more reliable and insightful tool for gauging genuine progress in the field.

5 Conclusion

In this work, we introduce a high quality and diversity benchmark to evaluate omni models comprehensively. The main findings are summarized as below.

1. With unified data framework in UNO-Bench, we found that the multimodal capability may not

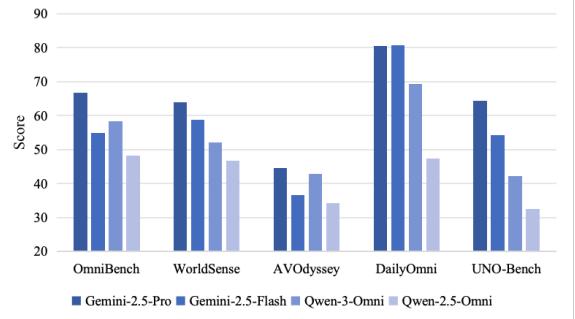


Figure 10: Comparison between omni benchmarks.

simply be a linear superposition of uni-modalities, but rather follows a significant multiplicative relationship. The evaluation results show that it manifests as a bottleneck effect on weak models, while exhibiting synergistic promotion on strong models.

2. Both uni-modal and omni-modal understanding capability of the Gemini series far surpasses existing open-source omni-models. The Gemini-2.5-Pro shows comparable perception capability with human experts but still has a performance gap in reasoning aspect.

3. Besides better dataset quality and evaluation efficiency, UNO-Bench can provide sufficient metric discriminability and a progressive difficulty scale to drive model capability growth.

In the future work, we will extend the dataset’s scale by human-in-the-loop automated pipeline and hold a private test set to avoid hacking. The ability coverage also needs to extend to more difficult reasoning tasks like STEM and code. At the same time, the relationship among cross-modals understandings and how to improve them are still exciting problem to explore.

Acknowledgement

We hereby express our appreciation to the LongCat Team EVA Committee for their valuable assistance, guidance, and suggestions throughout the course

of this work.

References

- Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, GuangMing Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, and 39 others. 2025. Ming-omni: A unified multimodal model for perception and generation.
- Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *Advances in Neural Information Processing Systems*, 37:56898–56918.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.
- David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3.
- Xinyu Chen, Yunxin Li, Haoyuan Shi, Baotian Hu, Wenhan Luo, Yaowei Wang, and Min Zhang. 2025. Videovista-culturallingo: 360 horizons-bridging cultures, languages, and domains in video comprehension. *arXiv preprint arXiv:2504.17821*.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.
- Fahim Faisal, Sharlina Keshava, Md Mahfuz ibn Alam, and Antonios Anastasopoulos. 2021. **SD-QA: Spoken Dialectal Question Answering for the Real World**. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP Findings)*. Association for Computational Linguistics.
- Chaoyou Fu, Yuhua Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.
- Zhang Ge, Du Xinrun, Chen Bei, Liang Yiming, Luo Tongxu, Zheng Tianyu, Zhu Kang, Cheng Yuyang, Xu Chunpu, Guo Shuyue, Zhang Haoran, Qu Xingwei, Wang Junjie, Yuan Ruibin, Li Yizhi, Wang Zekun, Liu Yudong, Tsai Yu-Hsuan, Zhang Fengji, and 3 others. 2024. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.20847*.
- Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. 2024. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *abs/2412.02611*.
- Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *abs/2502.04326*.
- Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. **Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos**.
- Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. 2025. **Online video understanding: Ovbench and videochat-online**. *Preprint*, *arXiv:2501.00584*.
- Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjali Awasthi, Sreenivas Gollapudi, Dee Guo, and 1 others. 2024. Remi: A dataset for reasoning with multiple images. *Advances in Neural Information Processing Systems*, 37:60088–60109.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024a. **Mvbench: A comprehensive multimodal video understanding benchmark**. *Preprint*, *arXiv:2311.17005*.

- Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, and 74 others. 2025. *Baichuan-omni-1.5 technical report*. Preprint, arXiv:2501.15368.
- Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, and 2 others. 2024b. Omnibench: Towards the future of universal omni-language models. abs/2409.15272.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Yike Yuan, Wangbo Zhao, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahu Lin. 2024b. Mmbench: is your multi-modal model an all-around player?
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024c. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024d. *Ocr-bench: on the hidden mystery of ocr in large multi-modal models*. *Science China Information Sciences*, 67(12).
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun-yuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun-yuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. *Docvqa: A dataset for vqa on document images*. Preprint, arXiv:2007.00398.
- Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2024. *Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations*. Preprint, arXiv:2412.07626.
- Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaseswaran S, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. Mmau: A massive multi-task audio understanding and reasoning benchmark.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.
- Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *Interspeech*, volume 2021, pages 2247–2251.
- Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.
- Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024b. *Measuring multimodal mathematical reasoning with math-vision dataset*. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*.
- Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. *Longvideobench: A benchmark for long-context interleaved video-language understanding*. Preprint, arXiv:2407.15754.

- xAI. 2023. Realworldqa. <https://huggingface.co/datasets/xai-org/RealworldQA>. Version 1.0, Accessed: 2024.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.
- LLM-Core-Team Xiaomi. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report.
- Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *Preprint*, arXiv:2404.18824.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, and 1 others. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.
- Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, and 1 others. 2024. Gmai-mmmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *Preprint*, arXiv:2409.02813.
- Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *arXiv preprint arXiv:2508.06471*.
- Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. *Preprint*, arXiv:2501.12380.
- Ziwei Zhou, Rui Wang, and Zuxuan Wu. 2025. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *Preprint*, arXiv:2505.17862.

A Benchmarks Utilized in Dataset Compression

To construct our compressed datasets, we utilized a variety of benchmarks for both visual and audio modalities. For the visual component, we curated data from 15 public and several private benchmarks that assess a range of capabilities, including general visual question answering, document and chart comprehension, STEM/scientific reasoning, and multi-image understanding. For the audio component, we used 3 audio question answering benchmarks. The detailed composition of the resulting uni-modal dataset is presented in Figure 11.

- **General visual question answering**, RealWorldQA(xAI, 2023), MME(Chaoyou et al., 2023), SeedBench(Li et al., 2023).
- **Document and chart understanding**, OCR-Bench (Liu et al., 2024d), Fox(Liu et al., 2024a), DocLocal4k(Ye et al., 2023).
- **Stem & reasoning**, MMMU(Yue et al., 2024), MMMU-Pro(Yue et al., 2025), CMMU(Ge et al., 2024), MathVista(Lu et al.,

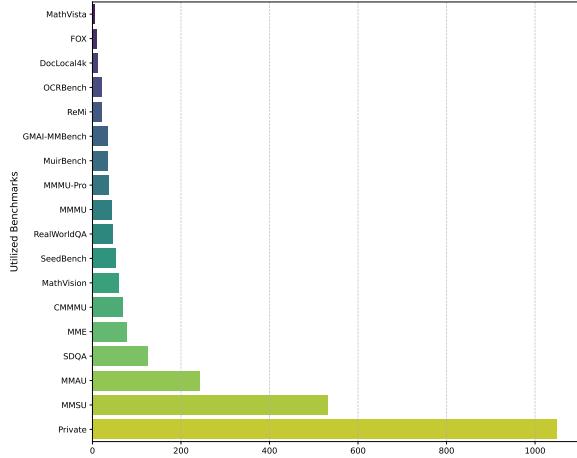


Figure 11: The distribution of the uni-modal benchmarks in UNO-Bench.

2024b), MathVision(Wang et al., 2024b), ScienceVista(Team et al., 2025), GMAI-MMBench(Ye et al., 2024).

- **Multi-image Understanding**, ReMi(Kazemi et al., 2024), MuirBench(Wang et al., 2024a).
- **Audio question answering**, MMAU(Sakshi et al., 2025), MMSU(Wang et al., 2025), SDQA(Faisal et al., 2021).

B Model Selection for the Compositional Law

To validate our choice of the Compositional Law, we compared its performance against several alternative models. The fitting results for all candidate models on our 9-model dataset are summarized in Table 7.

As shown in Table 7, more complex models like the ‘Generalized Power Law’ achieve a near-perfect fit on the training data. However, this superior performance is misleading. These models yield parameters that are physically implausible, such as negative exponents (e.g., $P_{\text{Audio}}^{-1.59}$) or negative weights. Such parameters would illogically imply that improving a model’s uni-modal capability could degrade its omni-modal performance. This is a classic symptom of overfitting, where a model with high capacity learns the noise in a small dataset rather than a generalizable underlying trend.

In contrast, our proposed **Symmetric Power Law** provides an excellent fit ($R^2 = 0.976$) while maintaining theoretical coherence. All its parameters are positive and have clear interpretations:

a super-linear synergy ($\alpha = 2.19 > 1$) between modalities, a positive scaling factor ($C = 1.03$), and a reasonable baseline score ($b = 0.24$). Following the principle of Occam’s Razor, we select this model as it offers the most parsimonious, robust, and interpretable explanation for the observed phenomenon.

Interestingly, while the parameters from the overfitted models are invalid, they consistently suggest a stronger influence from the visual modality (e.g., the large positive exponent for P_{Visual} in the ‘Generalized Power Law’). This hints that while our symmetric law captures the primary collaborative effect, the visual component may play a slightly more dominant role, a direction for future investigation.

C General Scoring Model

One of the critical way to improve accuracy is to group questions into finer types and define appropriate criteria for each types. We define six different question types in Figure.12.

D Model Ability Taxonomy

This section will provide specific definitions for each ability item and present examples of various task types.

The specific model abilities and task types of the Perception dimension can be seen in Figure.13, and the Reasoning dimension can be seen in Figure.14.

Specific examples are provided for each model ability. Examples of Object Perception, Spatial Perception, Cross-Modal Alignment and Attribute Perception can be seen in Figure.15. Examples of Scenario Perception, Cross-Modal Conversion and Semantic Understanding can be seen in Figure.16. Examples of Complex Reasoning, Temporal Reasoning, Spatial Reasoning, Life Reasoning, STEM Reasoning and Code can be respectively seen in Figure.17, Figure.18 and Figure.19.

E Omni-modal Solubility and Uni-modal Solubility Cases

Omni-modal solubility means that both visual and audio information are required to answer the question. While uni-modal solubility means that only visual or audio information is required to answer the question.

There are two uni-modal solubility cases. One is when audio and video present the same amount of information, and only one modality is needed

Question Type	Criteria	Example
Numerical Type	Requires the model's response to exactly match the numerical value in the reference answer, with no margin of error.	Question: In which year was the Beijing Olympics held? Reference Answer: 2008 Model Response: 2004 Scoring Result: Incorrect.
Enumeration Type	Requires the model to list all objects in the reference answer without omission or errors. Synonyms or semantically equivalent expressions are allowed. Order must be maintained if specified.	Question: Which animals appear in the image? Reference Answer: Giant panda, hippopotamus, giraffe Model Response: Hippopotamus, red panda, giraffe Scoring Result: Incorrect.
Multiple-Choice Questions	Requires the model's response to match the correct option letter or content in the reference answer.	Question: Which dynasty did the poet Li Bai belong to? A. Tang Dynasty B. Song Dynasty C. Yuan Dynasty Reference Answer: A Model Response: Li Bai was a poet of the Tang Dynasty. Scoring Result: Correct.
Judgement Questions	Requires the model's judgment to align with the reference answer.	Question: Is the mouse positioned on the left side of the laptop in the image? Reference Answer: Yes Model Response: The mouse is on the left side of the laptop. Scoring Result: Correct.
Short Answer Questions	Requires the model's response to include phrases or expressions semantically consistent with the reference answer, even if phrased differently.	Question: What was the final ingredient added to the pot in the video? Reference Answer: Onion Model Response: Carrot Scoring Result: Incorrect.
Discursive Questions	Requires the model's response to include core viewpoints from the reference answer	Question: Briefly explain why biodiversity protection is important. Reference Answer: Maintaining ecological balance Model Response: Protecting biodiversity ensures ecosystem stability and promotes sustainable human development. Scoring Result: Correct.

Figure 12: Definition of finer question types for general scoring model.

Model Ability Taxonomy	Task Type	Definition
Object Perception	Human and Animal Recognition	Recognize persons or animals by combining information from different modalities.
	Other Entity Recognition	Recognize other entities by combining information from different modalities, for example, plants, daily necessities, electronic products, etc.
	Spot the Difference	Completely identify the differences among multiple images or audio clips by combining information from different modalities.
	Instrument Recognition	Identify different musical instruments through sound by combining information from different modalities.
Spatial Perception	Spatial Relationship	Determine the spatial relationship between people/objects by combining audio and visual information.
Cross-Modal Alignment	Timing Alignment	Examine the matching between information from different modalities, for example, matching a single audio clip with multiple images/videos, or a single image/video with multiple audio clips.
	Consistency Judgment	Determine whether the information within the same modality is consistent.
Scenario Perception	Background Sound Recognition	Identify the background sound in the audio; determine the environment in the image/video based on the background sound.
	Scene Recognition	Recognize the environment in images/videos in conjunction with audio, such as identifying scenic spot names and various indoor/outdoor scenes.
	Emotion Recognition	Determine emotions (fear, anger, happiness, surprise, doubt, hesitation, etc.) based on the tone, pitch, and particles of speech of people/animals in the audio.
	Event Recognition	Recognize the overall scene in a video/image, for example, describing the actions of people in the entire scene and the corresponding scene description; analyzing ongoing events; identifying the chronological order of events.
Cross-Modal Conversion	ASR	Recognize speech content, including the recognition of various dialects.
	OCR	Recognize text, including both short and long texts.
Attribute Perception	Counting	Count entities or actions that appear in audio, images, and videos.
	Age Judgment	Determine a person's age by their timbre.
	Human Recognition	Identify the number of people by different timbres.
	Attribute Recognition	Recognize all entity-related attributes, such as color, size, material, etc.
Semantic Understanding	Gender Judgment	Determine a person's gender by different timbres.
	Pause Comprehension	Recognize the different meanings expressed by pauses at different positions in speech within an audio.
	Reading Comprehension	Understand the ultimate meaning conveyed through a person's dialogue.
	Long Audio Summarization	Summarize the content of long audio information.
	Coreference Resolution	Understand the specific referents of various personal pronouns that appear in the audio through dialogue and other supplementary information.

Figure 13: Definition of the Perception Dimension.

Model Ability Taxonomy	Task Type	Definition
Code	Code	Coding problems, including languages such as Python, C++, Java, etc.
Complex Reasoning	Multi-step Reasoning	Reasoning problems that require multiple steps to solve.
	Route Planning	Provide action route planning according to the target by combining information from different modalities.
	Trajectory Prediction	Predict the subsequent action trajectory, direction, and motion state by combining information from different modalities.
Spatial Reasoning	Puzzle	In jigsaw puzzle tasks, complete tasks such as puzzle restoration and fragment searching by combining spatial understanding abilities.
	Perceptive Taking	Examine the model's understanding of the positional relationship of objects in space from different perspectives.
	Relative Position	Determine the relative position, direction, angle, etc., of objects in space by combining information from different modalities.
	Event Analysis	Analyze the causes and effects of events by combining information from different modalities.
Temporal Reasoning	Event Ordering	Sort past events according to a certain objective order; or organize the correct sequence of an event based on fragmented information.
	Future Prediction	Predict future actions or events by combining information from different modalities.
	Social Cognitive Reasoning	Infer personal relationships, social culture, occupations, etc., by combining information from different modalities.
	Life Reasoning	Includes reasoning in various life scenarios, such as intelligent customer service, combining food delivery orders, common life knowledge, etc.
	Riddle Reasoning	Various riddles, escape room puzzles, and other similar questions.
General Reasoning	Counterfactual Reasoning	Given the conditions and result of an event, ask what result will occur if a certain condition is changed.
	Logical Relationship Reasoning	Involves various logical relationships such as causality and analogy, and requires reasoning according to given rules or logic.
	Card Reasoning	Questions related to chess and card games, including poker, mahjong, Chinese chess, etc.
	Game Reasoning	Various game-related questions, including board games, mobile games, computer games, etc.
	Geography	Geography-related disciplinary reasoning, with a difficulty range from middle school to university level.
	Chemistry	Chemistry-related disciplinary reasoning, with a difficulty range from middle school to university level.
STEM Reasoning	Biology	Biology-related disciplinary reasoning, with a difficulty range from middle school to university level.
	Mathematics	Mathematics-related disciplinary reasoning, with a difficulty range from middle school to university level.
	Physics	Physics-related disciplinary reasoning, with a difficulty range from middle school to university level.

Figure 14: Definition of the Reasoning Dimension.

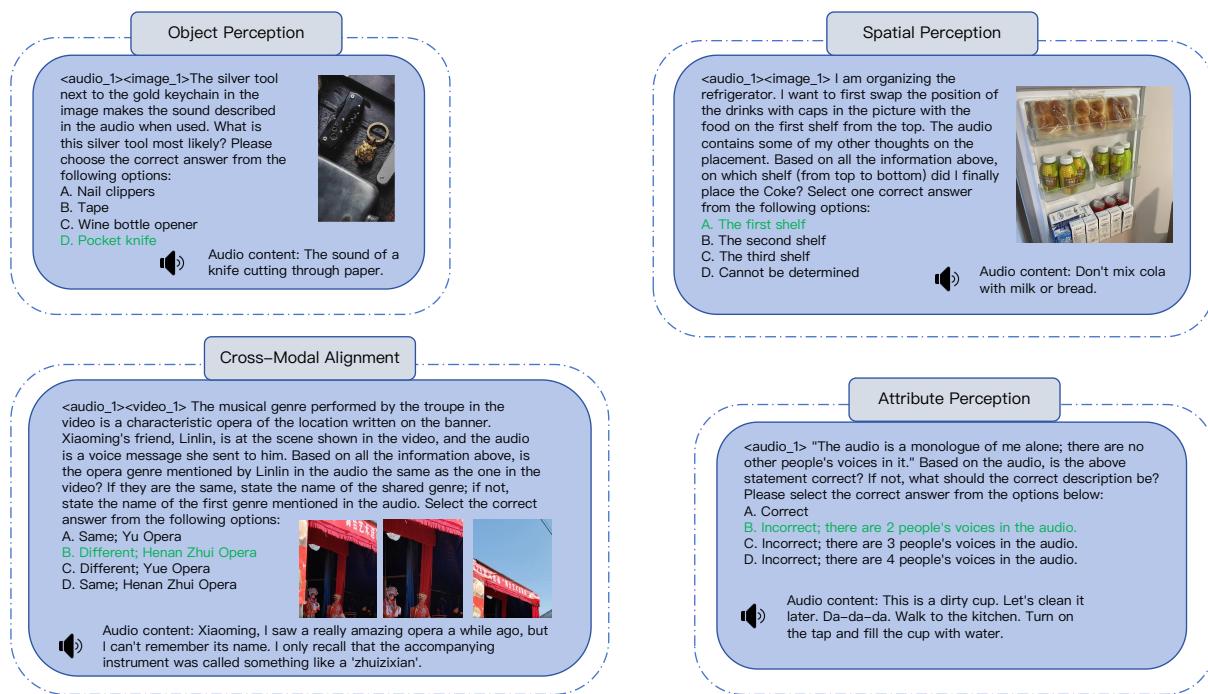


Figure 15: Several cases in the Object Perception, Spatial Perception, Cross-Modal Alignment and Attribute Perception.

Scenario Perception

<audio_1><image_1> Here is a piece of text: "Xiaohong was hiking in the wild and passed by a stream..." Based on the text, image, and audio, reconstruct the entire event. Please select the correct answer from the four options below:

A. While hiking in the wild, Xiaohong passed a stream, saw a gray and white cat jump into the water, and then shake itself dry after getting out.
 B. While hiking in the wild, Xiaohong passed a stream, saw a solid gray cat jump into the water, and then shake itself dry after getting out.
 C. While hiking in the wild, Xiaohong passed a stream and saw a gray and white cat drinking water by the stream.
 D. While hiking in the wild, Xiaohong saw a stream and a solid gray cat drinking water by the stream.



Audio content: The sound of falling water.

Cross-Modal Conversion

<audio_1> Which of the following sentences is spoken in the Guangxi dialect in the audio? Please select the correct answer from the options:

A. My favorite fruit is lychee.
 B. Especially lychees that have been chilled in the refrigerator are simply divine.
 C. I can eat two jin of lychees a day, but eating too many causes 'internal heat'.
 D. So, it is still important to eat them in moderation.

Audio content:
 (English): It's June. It's time for lychee.
 (Cantonese): My favorite fruit is lychee.
 (Guangxi dialect): Especially lychees that have been chilled in the refrigerator are simply divine.
 (Shanghai dialect): I can eat two jin of lychees a day, but eating too many causes 'internal heat'.
 (Sichuan dialect): So, it is still important to eat them in moderation.

Semantic Understanding

<audio_1> The audio contains a conversation that takes place at the same time and in the same scene. The three people who speak in chronological order are A, B, and C. Based on the text and audio information, please answer: Who is the "Bro" that B refers to? And who is the "he" that B refers to? Please select the correct answer from the following options:

A. A; C
 B. A; F
 C. F; C
 D. C; A

Audio content:
 A says: You're all here, so let's go.
 B says: Wait, F isn't here yet. What's up with that?
 C says: How come you're so on time today?
 That's a rare sight.
 B says: Bro! Look at him! He's mean to me again.

Figure 16: Several cases in the Scenario Perception, Cross-Modal Conversion and Semantic Understanding.

Complex Reasoning

A, B, C, D, and E, five people in total, visited the four famous attractions shown in the picture. Each person visited only one attraction, and each attraction was visited by at least one person. The following is known:
 The visitor to Figure 1 is neither Egyptian nor Italian.
 The French visitor did not visit Figure 2 or Figure 3.
 The visitor to Figure 4 is younger than the visitor to Figure 3.
 The Chinese visitor is the oldest, and the Indian visitor is the youngest.
 The Egyptian visitor did not visit Figure 1 or Figure 3.
 The visitor to Figure 3 is not the Chinese visitor.
 The visitor to Figure 2 is neither French nor Chinese.
 The Italian visitor is younger than the French visitor.
 The visitor from Egypt is younger than the visitor from France but older than the visitor from Italy.
 Based on all the information above, the images, and the audio content, what is the order of their ages from oldest to youngest? And which attraction did A, B, C, D, and E visit, respectively? Please select the correct answer from the following options:

A. D-C-A-E-B; A visited the Louvre, B visited the Taj Mahal, C and D visited the Pyramids, E visited the Great Wall.
B. D-C-A-E-B; A and E visited the Louvre, B visited the Taj Mahal, C visited the Pyramids, D visited the Great Wall.
C. B-E-A-C-D; A visited the Louvre, B visited the Taj Mahal, C and D visited the Pyramids, E visited the Great Wall.
D. D-C-A-E-B; A visited the Louvre, B visited the Taj Mahal, E visited the Pyramids, C and D visited the Great Wall.

Audio content: It is known that A is from Egypt, B is from India, C is from France, D is from China, and E is from Italy.

Figure 17: The case in the Complex Reasoning.

Temporal Reasoning

<audio_1><image_1> Lili is a sports enthusiast. She participated in some competitions during the May Day holiday. The image contains photos Lili sent me from three of her competition days between May 2nd and May 5th. She participated in exactly one competition each day from the 2nd to the 5th, and she competed in a surfing competition the day after running a marathon. The audio is a recording of Lili talking to a friend.

Based on all the information provided, please list the events Lili participated in and their corresponding dates in chronological order. Select the correct answer from the options below.

- A. May 2nd – Skiing, May 3rd – Marathon, May 4th – Surfing, May 5th – Car Racing
- B. May 2nd – Skiing, May 3rd – Car Racing, May 4th – Marathon, May 5th – Surfing
- C. May 2nd – Marathon, May 3rd – Surfing, May 4th – Skiing, May 5th – Car Racing
- D. May 2nd – Marathon, May 3rd – Surfing, May 4th – Car Racing, May 5th – Skiing



🔊 Audio content: During the May Day holiday, starting from May 2nd, on the first day I did the sport shown in Figure 3 of the image, and then on the second day, I did the sport in Figure 2. Of course, I participated in more than just these sports; these past few days have been so fulfilling.

Spatial Reasoning

<audio_1><image_1> It is known that Xiaoli is a girl wearing a white top today. Xiaohuang has captured both Xiaoli and her good friend Maomao in the picture. The audio is Xiaohuang narrating the WeChat chat history between Xiaoli and Maomao. This conversation took place at the moment the photo was taken. Throughout the conversation in the audio, Xiaohuang remained in the photographer's position, and no one changed their location. Based on all the information above, please answer: From Xiaoli's perspective, in which direction is Maomao? And in which direction is Xiaohuang? (Both answers should be from Xiaoli's point of view). Please select the correct answer from the following options:

- A. To the rear right; to the front left
- B. To the rear left; to the front right**
- C. Directly behind; to the front left
- D. Directly to the left; to the front right



Xiaoli says: Maomao, where are you?
Maomao says: I'm at the entrance of the Chow Tai Seng store.
Xiaoli says: There is a white horse to my left.



Figure 18: Several cases in the Temporal Reasoning and Spatial Reasoning.

Table 7: Fitting results for all candidate models. While more complex models achieve higher fitting scores (R^2), their parameters lack physical interpretability (e.g., negative exponents), indicating severe overfitting on our small dataset. Our chosen **Symmetric Power Law** offers the best balance of a high R^2 value and theoretical soundness.

Model Name	R^2	RMSE	Fitted Equation
Generalized Power Law	0.999	0.005	$P_{\text{Omni}} \approx 1.33 \cdot P_{\text{Audio}}^{-1.59} \cdot P_{\text{Visual}}^{5.09} + 0.24$
Linear Interaction	0.995	0.010	$P_{\text{Omni}} \approx 0.97 - 2.01P_{\text{Audio}} - 0.59P_{\text{Visual}} + 2.85(P_{\text{Audio}} \times P_{\text{Visual}})$
Weighted Sum Power Law	0.995	0.010	$P_{\text{Omni}} \approx 1.19 \cdot (-0.20P_{\text{Audio}} + 1.20P_{\text{Visual}})^{3.83} + 0.24$
Symmetric Power Law	0.976	0.022	$P_{\text{Omni}} \approx 1.03 \cdot (P_{\text{Audio}} \times P_{\text{Visual}})^{2.19} + 0.24$
Simple Linear	0.945	0.033	$P_{\text{Omni}} \approx -0.15 - 0.37P_{\text{Audio}} + 1.43P_{\text{Visual}}$

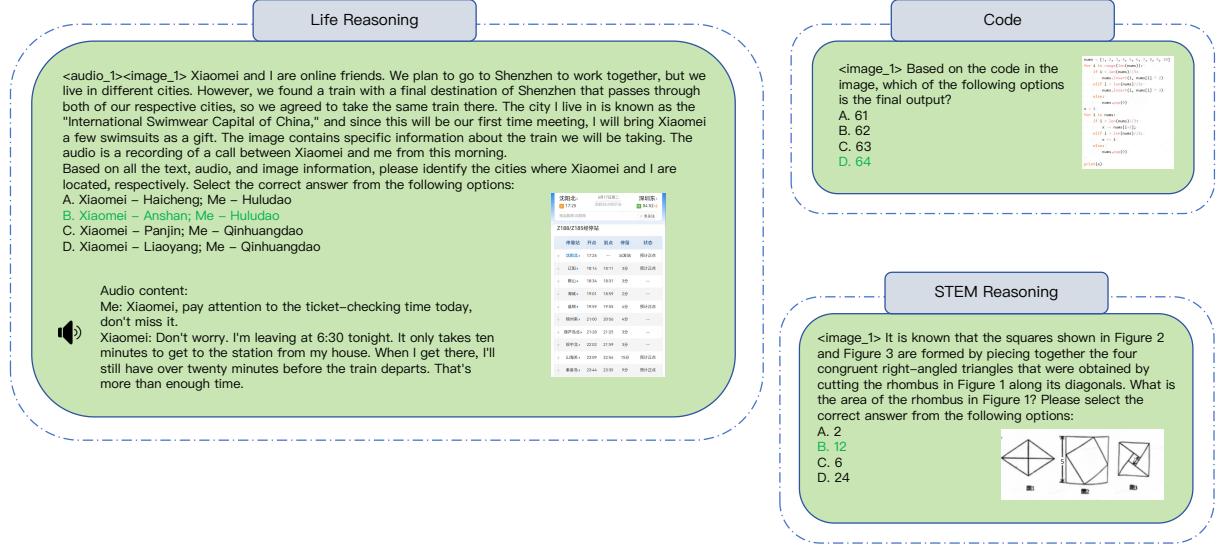


Figure 19: Several cases in the Life Reasoning, STEM Reasoning and Code.

to solve the problem, as shown in Figure.20. The other is when one modality provides useful information, while the other modality is unhelpful, as shown in Figure.21.

The omni-modal solubility case can be seen in Figure.22.

<audio_1><video_1>Where should lavender be placed on the bed to provide protection for the home?

- A. At the foot of the bed.
- B. On the bedposts.**
- C. Under the pillow.
- D. At the head of the bed.



Audio content: Lavender and magic. In magic, lavender has most often been used for protection. People would often hang lavender on their bedposts, or you can mix it with basil, lemon balm, thyme, rue, and frankincense.....

Figure 20: The case that can be solved with only video or audio.

<audio_1><video_1>How many times do dragons appear in the video?

- A. Twice.
- B. Four times.
- C. Three times.
- D. Once.



Audio content: Really good. I want to clarify, I'm just I'm just a little burned out and frustrated and tired and it's not just Nintendo, it's the state of the world. I mean, Xbox has been nothing this year as well. PlayStation has been struggling to stay afloat.....

Figure 21: The case that can be solved with only video.

<audio_1><image_1>The image is of a car-unblocking puzzle game that I am playing. The rules of the game are: "Click on a vehicle, and it will move in the direction of the arrow on its roof. If its path is blocked by another vehicle, it cannot exit the parking lot in the picture. If there is no obstruction, it can successfully exit the parking lot." Please answer the question in the audio based on the game rules and the information in the picture. Select the correct answer from the following options.

- A. Pink
- B. Blue
- C. Purple
- D. Orange

Audio content: I used a tool to change the pink car in the bottom right corner to blue, and the purple car behind it to orange. Now I can only choose to click on cars of the same color. Which color should I choose to allow the maximum number of cars to successfully exit the parking lot?



Figure 22: The case that can be solved with both image and audio.