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

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have been progressing from uni-modal understanding toward unifying visual, audio, and language modalities, collectively termed omni models. However, the relationship between uni-modal and omni-modal capabilities remains poorly understood, requiring comprehensive evaluation to drive omni models' intelligence evolution. In this work, we introduce **UNO-Bench**, a unified benchmark that jointly evaluates **UN**i-modal and **O**mni-modal capabilities under a shared ability taxonomy spanning 44 task types and 5 modality combinations. UNO-Bench comprises 1,250 human-curated omni-modal samples with 98% cross-modality solvability—a strict constraint eliminating single-modal shortcuts—and 2,480 uni-modal samples distilled from 18 public benchmarks at 90% cost reduction. Beyond multiple-choice questions, we propose a multi-step open-ended format assessed by a general scoring model (6 question types, 95% accuracy). Experimentally, we discover a **Compositional Law**: omni-modal performance follows a super-linear power-law of the product of uni-modal scores, manifesting as a bottleneck in weaker models and accelerating cross-modal gains in stronger ones. These findings offer quantitative insights into the path toward advanced omni-modal intelligence.

1 Introduction

Multimodal artificial intelligence has undergone extensive research in visual language models and audio language models, with current advancements progressing toward unifying visual, audio, and language modalities, collectively termed omni models. The ideal omni model should simultaneously preserve visual understanding capabilities, speech interaction proficiency, and cross-modal integration capacity, requiring evaluation paradigms that go beyond isolated modality assessment.

However, current benchmarks fall short on two fronts. First, existing evaluation suites assess uni-modal and omni-modal capabilities in isolation, creating disconnected and resource-intensive evaluation processes. Second, and more critically, existing omni-modal benchmarks fail to ensure that questions genuinely require cross-modal reasoning: 77% of questions in WorldSense(Hong et al., 2025) are solvable using a single modality alone, and 25% of questions in OmniBench(Li et al., 2024b) contain erroneous answers. Without guaranteeing that questions require joint cross-modal reasoning, high benchmark scores may simply reflect single-modal proficiency rather than true cross-modal integration—making it impossible to reliably evaluate or analyze genuine omni-modal capability.

Due to these limitations in quality and coverage, we introduce UNO-Bench, a novel unified benchmark designed to provide rigorous and comprehensive evaluation of omni models. UNO-Bench is built around three design principles. First, it adopts a unified ability taxonomy that jointly covers uni-modal and omni-modal capabilities under a shared two-dimensional framework of Perception and Reasoning, spanning 44 task types and 5 modality combinations (image, video, audio, and their combinations). This unified structure enables direct comparison of how models perform across modality boundaries and avoids the disconnected assessment common in existing evaluation suites. Second, the omni-modal dataset is constructed through a human-centric pipeline emphasizing data authenticity and cross-modal necessity: all questions are verified to require joint reasoning across modalities, and audio content is recorded by native speakers in real-world Chinese contexts to ensure ecological validity. Third, beyond conventional multiple-choice questions, UNO-Bench introduces multi-step open-ended questions that decompose complex problems into interdependent sub-questions, enabling fine-grained measurement of multi-step

Dataset	Omni-modal	Uni-modal	Acc.	Solvable	Source	#Tasks	#QA Pairs	QA Type	Language
MMBench	✗	I	-	-	80% private	20	3217	MC	EN/CH
MMAU	✗	A	-	-	15% private	27	10000	MC	EN
MVBench	✗	V	-	-	public	20	4000	MC	EN
OmniBench	I+A	✗	75%	90%	public	8	1142	MC	EN
AV-Odyssey	I+V+A	✗	91%	99%	public	26	4555	MC	EN
WorldSense	V+A	✗	99%	23%	public	26	3172	MC	EN
Daily-Omni	V+A	✗	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	-	-	40% 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).

cross-modal reasoning chains. A comparison with existing benchmarks is provided in Table 1.

Our main contributions are summarized as follows:

- We propose **UNO-Bench**, the first unified benchmark that jointly assesses **UN**i-modal and **O**mni-modal understanding under a shared ability taxonomy. A central design principle is that 98% of omni-modal questions require both modalities jointly for resolution, eliminating the single-modal shortcuts that plague existing benchmarks. This rigorous design enables a clean empirical investigation of how uni-modal capabilities compose into omni-modal performance, leading to our discovery of the Compositional Law: omni-modal performance follows a super-linear power-law function of the product of uni-modal scores, where the super-linear exponent quantifies genuine cross-modal synergy—stronger models exhibit accelerating omni-modal gains beyond what their individual uni-modal scores would linearly predict.

- We establish a high-quality and diverse dataset through a pipeline integrating human curation and automated compression, comprising 1,250 omni-modal samples across 44 task types and 5 modality combinations, with particular coverage of real-world Chinese contexts. The automated uni-modal compression reduces evaluation costs by 90% while maintaining 98% consistency with full-scale benchmarks, substantially improving upon existing datasets in both quality and efficiency.

- Beyond conventional multiple-choice questions, we introduce a Multi-Step Open-Ended Question (MO) format for more realistic and discriminative assessment of complex cross-modal reasoning, supported by a General Scoring Model covering 6 question types with 95% accuracy on unseen models and benchmarks.

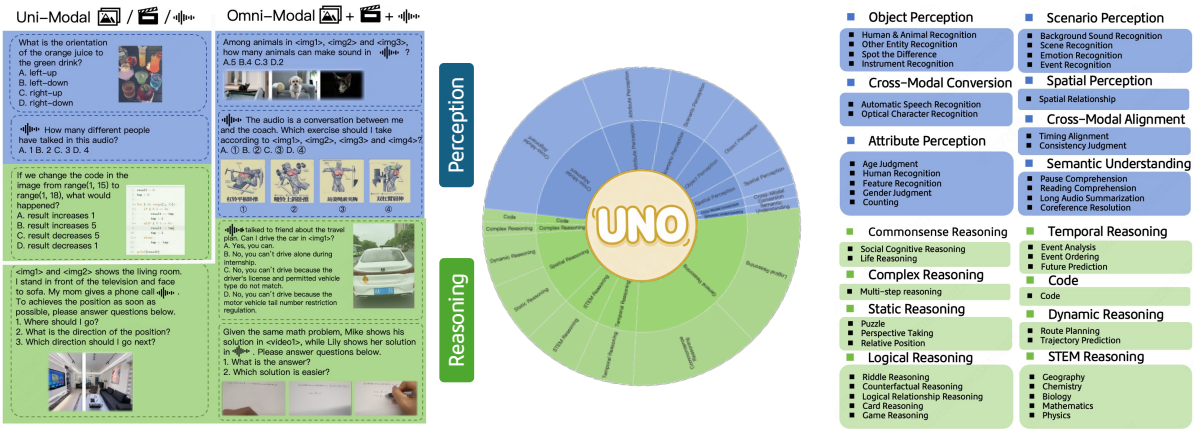
2 Related Work

2.1 Uni-Modal Benchmarks

Vision language models (VLMs) (Bai et al., 2025; Xiaomi, 2025; Zeng et al., 2025) and audio language models (ALMs) (Ding et al., 2025; Wu et al., 2025) have advanced multimodal understanding across vision and audio modalities respectively. Corresponding uni-modal benchmarks cover diverse capabilities: for image understanding, representative works include MMBench (Liu et al., 2024b), OCRBench (Liu et al., 2024c), and DocVQA (Mathew et al., 2021); for video, MVBench (Li et al., 2024a) and Video-MME (Fu et al., 2024); for audio, MMAU (Sakshi et al., 2025), AIR-Bench (Yang et al., 2024), and SD-Eval (Ao et al., 2024). Despite broad coverage, these benchmarks evaluate modalities in isolation, leaving cross-modal interaction unaddressed.

2.2 Omni-Modal Benchmarks

Omni models capable of jointly processing vision, audio, and language have emerged in recent years (Comanici et al., 2025; Xu et al., 2025b; AI et al., 2025; Li et al., 2025), with Gemini (Comanici et al.,



(a) Samples in Uni-Modal and Omni-Modal (b) Unified Ability Taxonomy and Distribution (c) Task Type Taxonomy

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.

2025) and Qwen-3-Omni (Xu et al., 2025b) representing the state of the art in proprietary and open-source settings respectively. However, existing omni-modal benchmarks have notable limitations. OmniBench (Li et al., 2024b) constructs omni-modal questions by inserting audio as context into image understanding tasks, but suffers from data quality issues. WorldSense (Hong et al., 2025) provides high-quality audio-visual data in real-world scenarios, yet most questions can be solved using a single modality alone, making it unable to assess genuine cross-modal ability. Other datasets (Gong et al., 2024; Zhou et al., 2025) cover limited task types and modality combinations.

As illustrated in Figure 3, questions solvable by a single modality (b, c) tend to overestimate omni-modal capability; only questions requiring joint cross-modal reasoning (a) can reliably probe true omni-modal understanding (see Section 4.3). Addressing these limitations, we propose UNO-Bench to enable rigorous and comprehensive omni-modal evaluation.

3 Method

3.1 Omni-modal Dataset Construction

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

3.1.1 Model Ability Taxonomy

Through cumulative experience in multimodal evaluation, we summarize the capabilities of uni-modal and omni-modal models into a unified ability taxonomy. As shown in Figure 1, capabilities are categorized into two primary dimensions: Perception and Reasoning.

The Perception dimension covers Object Perception, Attribute Perception, Scenario Perception, Spatial Perception, Cross-Modal Conversion, Semantic Understanding, and Cross-Modal Alignment. The Reasoning dimension extends conventional categories (General, STEM, Code) with Spatial Reasoning, Temporal Reasoning, and Complex Reasoning. Based on this taxonomy, we create a diverse dataset with 44 task types (Figure 1(c)); detailed definitions are in Appendix I.

3.1.2 Annotation Pipeline

Based on the unified taxonomy, we construct the dataset through three steps: Material Collection, QA Annotation, and Quality Inspection.

- **Material Collection.** To eliminate cross-modal information redundancy and the risk of data leakage, we collect diverse novel data sources and replace all original audio dialogues with annotator recordings tailored to each task. All dialogue audio is recorded by over 20 native speakers, covering Mandarin and regional dialects (e.g., Sichuan dialect); materials span a broad range of topics including society, culture, science, and daily life, with a focus on Chinese real-world contexts.

- **QA Annotation.** Drawing on human experts and high-quality crowd-sourced annotators, we construct QA pairs that simulate

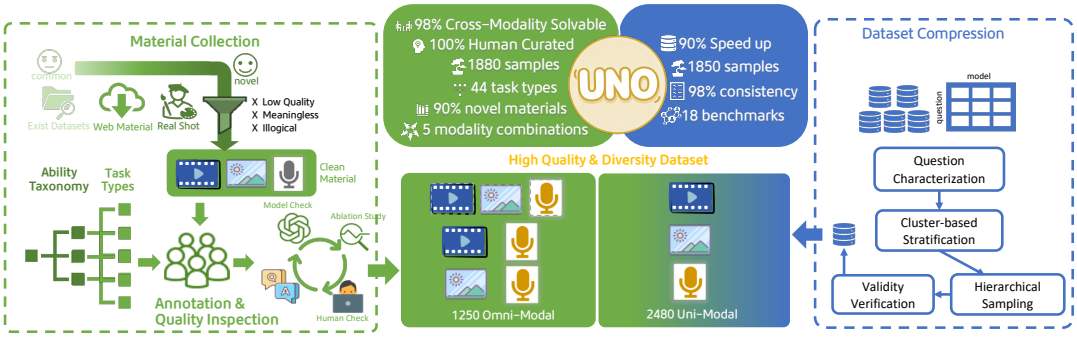


Figure 2: Dataset Construction Pipeline includes human crafted process (left side) and automated data compression (right side). First, we collect diverse and novel materials to prevent data contamination. Second, with the proposed unified ability taxonomy, human annotators including experts will craft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

real-world scenarios through a human-centric approach.

- **Quality Inspection & Annotator Training.** We employ an iterative cross-review pipeline with 13 annotators and 4 inspectors. Quality is tracked via Expert-Annotator Alignment Rate and Critical Defect Rate (CDR): training drove alignment from 44% to 83% and reduced CDR from 20% to 4%. Each item underwent 2–3 rounds of independent inspection with rotated reviewers, achieving a final acceptance rate exceeding 99%.

3.2 Uni-modal Dataset Improvement

3.2.1 Quality Improvement

Existing public uni-modal datasets suffer from data leakage issues (Xu et al., 2024). To verify the influence, we apply privatization improvement to MMBench (Liu et al., 2024b). As shown in Figure 9, model performances become more distinguishable after improvement, better reflecting true capability differences.

3.2.2 Dataset Compression

To reduce evaluation costs while maintaining consistency with full-scale benchmarks, we design a Clustering-Guided Hierarchical Sampling (CGHS) method (Figure 2). CGHS uses model performance metrics as sample features—rather than question content—to select representative samples. The pipeline proceeds as follows:

Question Characterization & Clustering: Each question is represented as a vector of model scores. Kmeans++ (Arthur and Vassilvitskii, 2007) then categorizes questions into k clusters representing performance-similar types (e.g., easy, difficult).

Hierarchical Sampling & Verification: Sample sizes per stratum are determined by cluster proportions. Compression quality is measured by Spearman's Rank Correlation Coefficient (SRCC), Pearson's Linear Correlation Coefficient (PLCC), Root Mean Square Error (RMSE), Margin of Error (MoE), and Confidence Interval Coverage (CIC).

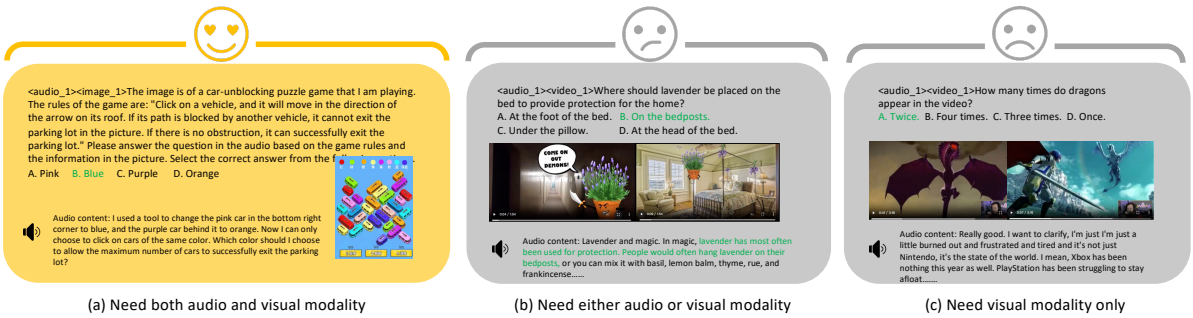
To ensure statistical stability, we use 5 random splits and 10-fold cross-validation. At a 10% sampling rate, CGHS achieves SRCC > 0.98 and PLCC > 0.98 with RMSE below 0.02 and CIC $\approx 95\%$, reducing evaluation costs by over 90% while preserving benchmark rankings. Full curves are shown in Figure 10.

3.3 Multi-Step Open-Ended Questions

3.3.1 Question Type Definition

Evaluating multi-step reasoning in omni models is challenging. Current automated benchmarks relying on Outcome Reward Models (ORMs) provide only binary pass/fail judgments, failing to distinguish a model that completes 80% of a task from one that fails at 20%. Alternatives such as Process Reward Models (Lightman et al., 2023) or multi-turn dialogues (Reddy et al., 2019) are hampered by high implementation cost or poor efficiency. Moreover, the prevalence of multiple-choice formats is unrepresentative of real-world open-ended queries and may conceal model weaknesses.

To address these issues, we propose the **Multi-Step Open-Ended Question (MO)** format. Complex problems are deconstructed by human experts into progressive, interdependent sub-questions,



(a) Need both audio and visual modality (b) Need either audio or visual modality (c) Need visual modality only

Figure 3: Illustration of the cross-modality solvable sample.

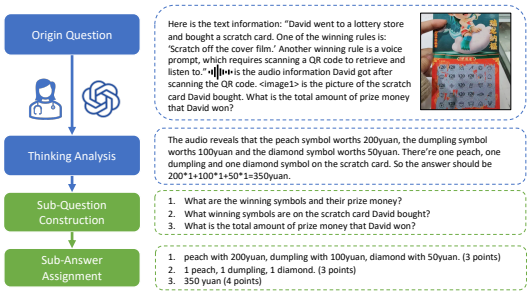


Figure 4: Construction Pipeline of Multi-Step Open-Ended Questions.

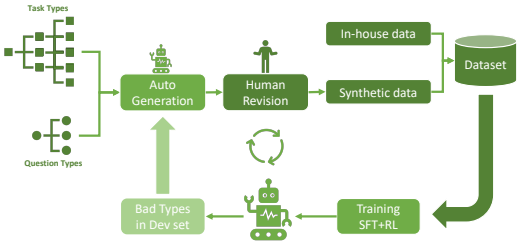


Figure 5: Training dataset construction and model training pipeline for general scoring model.

each assigned a score by importance (total: 10 points). During testing, all sub-questions are posed in a single turn, requiring step-by-step open-ended responses. This method precisely quantifies how far along a complex reasoning chain a model can proceed. Importantly, all MO items are grounded in inherently omni-modal contexts: the interdependency between sub-questions means that any failure in cross-modal integration propagates through the subsequent reasoning chain, making MO particularly diagnostic of omni-modal reasoning weaknesses that single-step multiple-choice questions may mask. An example is shown in Figure 4.

3.3.2 General Scoring Model

Multi-step open-ended questions introduce a new challenge for automated evaluation. We propose a general scoring model based on Qwen3-14B (Yang et al., 2025) that takes a *(question, reference, prediction)* triplet as input and outputs a scalar score. One critical improvement is grouping questions into finer types with tailored scoring criteria; we define six question types covering numerical, enumeration, multiple-choice, judgment, short-answer, and discursive formats. The training pipeline is illustrated in Figure 5.

Empirical evaluations show our model achieves 95.05% accuracy on the internal MO benchmark, outperforming GPT-4.1 (OpenAI, 2023), while maintaining robust performance on the public benchmark EVOUNA (Wang et al., 2023). As shown in Section C.1, MO questions effectively reveal ability decay in long-chain reasoning, providing stronger model discrimination than single-step evaluation.

4 Experiment and Analysis

4.1 Experiment Setting

We evaluate omni models that support text, visual, and audio inputs simultaneously, including open-source models: Longcat-Flash-Omni(Team, 2025), 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), Ming-lite-Omni-1.5(AI et al., 2025), and MiniCPM-O-4.5(Cui et al., 2026), as well as closed-source models: Qwen3.5-Omni-Flash, Qwen3.5-Omni-Plus(Team, 2026), Gemini-3-Pro-Preview, Gemini-3-Flash-Preview, Gemini-2.5-Pro, Gemini-2.5-Flash, and Gemini-2.0-Flash(Comanici et al., 2025). For thinking models, we limit the thinking budget to 128 tokens following (Xu et al., 2025b), and disable the thinking mode for Gemini-2.5-Flash and Gemini-2.0-Flash.

Model	Audio	Visual	Omni-MC	Omni-MO
Qwen-2.5-Omni-3B	54.40	42.67	27.80	24.76
MiniCPM-O-2.6	56.50	42.27	28.60	23.76
MiniCPM-O-4.5	69.00	55.89	40.60	30.20
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
Qwen-2.5-Omni-7B	60.20	50.68	32.60	27.72
Qwen-3-Omni-30B-A3B	79.40	63.29	42.10	37.08
LongCat-Flash-Omni	**80.20**	**67.06**	**49.90**	**42.68**
Qwen3.5-Omni-Flash	71.80	71.64	52.80	45.36
Qwen3.5-Omni-Plus	84.20	75.64	65.30	54.96
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
Gemini-3.0-Flash-Preview	87.80	83.61	74.70	68.16
Gemini-3.0-Pro-Preview	**89.80**	**84.30**	**76.30**	**76.68**

Table 2: General performance of omni models in UNO-Bench for both uni-modal capability and omni-modal capability, where omni-modal benchmark includes multi-choice questions (Omni-MC) and multi-step open-ended questions (Omni-MO). Metrics are accuracy for all tasks except Omni-MO, which employs keypoint-based scoring.

4.2 Model Performance

4.2.1 Overall Analysis

Table 2 reveals a clear performance hierarchy. Among open-source models, smaller models perform only marginally above the random-guess baseline (25%) on Omni-MC, while Qwen-3-Omni-30B marks a significant leap at 42.10% and LongCat-Flash-Omni leads open-source models at 49.90%. Among closed-source models, Qwen3.5-Omni-Plus achieves 65.30%, though a substantial gap persists relative to Gemini-3.0-Pro (76.30%). A strong positive correlation between uni-modal scores and omni-modal performance suggests that robust uni-modal perception is a prerequisite for advanced omni-modal understanding.

The Omni-MO benchmark presents a considerably greater challenge, with universal performance degradation relative to Omni-MC scores across all models, revealing a systemic limitation in multi-step omni-modal reasoning. Among open-source models, LongCat-Flash-Omni achieves the highest MO score of 42.68%, while Gemini-3.0-Pro leads all models at 76.68%—a gap we attribute primarily to the scarcity of high-quality, modality-synergistic reasoning data within the open-source community.

Ablation studies confirm that visual information is critical for omni-modal performance—removing it reduces most models to near random-guess level—while audio contributes especially to speech-based tasks (See Appendix B.1 and Appendix B.2).

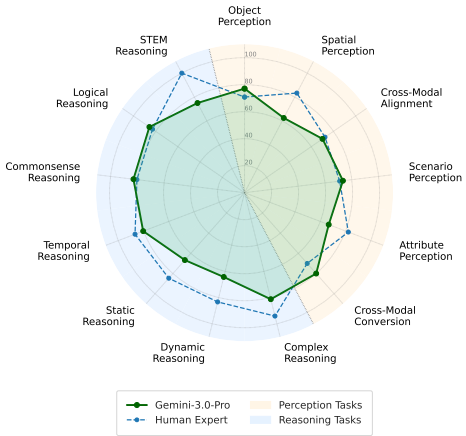


Figure 6: The competition between human experts and Gemini-3.0-Pro. Gemini-3.0-Pro shows comparable perception capability but lower reasoning capability.

4.2.2 Top-tier Analysis

In which direction should the state-of-the-art model be improved? To answer this question, we invited human experts for a competition; unlike the dataset annotators, these experts had no prior exposure to the questions or answers.

Finding 1. Gemini-3.0-Pro has reached human-comparable perception ability in omni-modal tasks, yet a gap remains in reasoning performance. Compared to human experts, Gemini-3.0-Pro exhibits similar perception scores but falls behind in reasoning (Figure 6). Notably, humans outperform the model in reasoning relative to perception (81.3% vs. 74.3%), an inversion of the model's pattern. Interviews with human experts suggest this reflects humans' tendency to miss fine-

grained details in video or audio, while benefiting from stronger causal and commonsense reasoning relative to their perceptual bandwidth.

4.2.3 Multi-Step Open-Ended Evaluation

Table 3 presents the complete results. Key observations include:

Harder than MC. The MO benchmark is generally more challenging than MC, with most models showing a performance drop. The gap widens for weaker models, indicating that MO captures multi-step reasoning demands beyond simple pattern matching.

Progressive difficulty. As sub-question depth increases (Q1 $\rightarrow$ Q3+), most models show clear performance decay. For example, Qwen-3-Omni-30B drops from 18.08% (Q1) to 11.42% (Q3+). Notably, Gemini-2.5-Pro improves from Q1 to Q2 (23.44% $\rightarrow$ 24.48%), suggesting better context utilization.

Reasoning bottleneck. The gap between perception and reasoning is substantially larger in open-source models than in top proprietary models, indicating that complex cross-modal reasoning remains a key differentiator.

4.3 Uni-Modal v.s. Cross-Modal

To investigate the relationship between uni-modal and omni-modal understanding, we conduct regression analysis and ablation experiments. The analysis is grounded in a key design property of UNO-Bench: 98% of omni-modal questions require the joint use of both modalities for resolution (Table 1), ensuring that observed performance reflects genuine cross-modal integration rather than single-modal shortcuts. In contrast, only 23% of questions in WorldSense satisfy this property, meaning models can achieve high scores there without performing any cross-modal reasoning. This design enables a controlled empirical investigation of how uni-modal capabilities compose into omni-modal performance.

Finding 2. Compositional Law: omni-modal capability is related to the product of individual uni-modal performances by a power-law. Observing Table 2, we identify a strong correlation between omni-modal performance and uni-modal capabilities. We formalize this through the Omni-modal Compositional Law, derived from a general functional form under two task constraints; full derivation is in Appendix F.

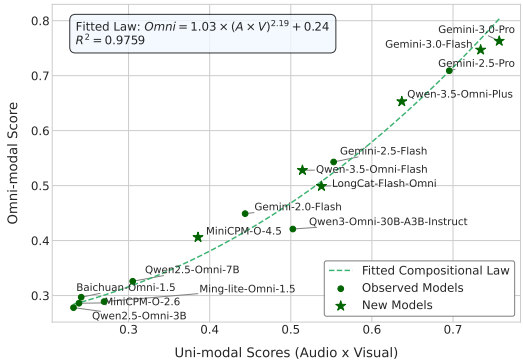


Figure 7: The Compositional Law of Omni-modal Performance. Observed omni-modal scores versus the product of their uni-modal scores. The dashed line represents our fitted law (Eq. 2), derived from 9 observed omni-modal models, and 6 additional held-out models are plotted to evaluate fitting generalizability. The convex, accelerating curve illustrates the Power-law synergy.

General Model & Task Constraints. We posit that omni-modal performance $\mathcal{P}_{\text{Omni}}$ is a function of uni-modal performances $\mathcal{P}_{\text{A}}$ and $\mathcal{P}_{\text{V}}$:

$$\mathcal{P}_{\text{Omni}}(\mathcal{P}_{\text{A}}, \mathcal{P}_{\text{V}}) = f_{\text{A}}(\mathcal{P}_{\text{A}}) + f_{\text{V}}(\mathcal{P}_{\text{V}}) + f_{\text{I}}(\mathcal{P}_{\text{A}}, \mathcal{P}_{\text{V}}) + b \tag{1}$$

where $f_{\text{A}}, f_{\text{V}}$ represent modality-independent contributions, f_{I} the cross-modal interaction, and b a baseline constant (e.g., random-guess performance). This yields:

$$\mathcal{P}_{\text{Omni}} = C \cdot (\mathcal{P}_{\text{A}} \times \mathcal{P}_{\text{V}})^{\alpha} + b \tag{2}$$

As shown in Figure 7, a non-linear regression on 9 observed models yields R^2=0.976; 6 additional held-out models validate the generalizability of the fit. The fitted exponent $\alpha \approx 2.19 > 1$ constitutes direct empirical evidence for cross-modal synergy: since $\alpha > 1$, the relationship is convex—as uni-modal performance improves, omni-modal gains accelerate rather than grow linearly, a quantitative "1+1>2" effect that is empirically observed across 15 omni models spanning a wide capability range.

Power-law Synergy and Emergent Ability. The super-linear exponent explains the transition from a bottleneck effect to emergent capability (Figure 7):

- **Limited Gains at Low Performance:** For weaker models (e.g., MiniCPM-O-2.6), the curve is relatively flat—small improvements in the uni-modal product yield only marginal omni-modal gains, manifesting as a "short-board effect."

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	9.00	7.25	5.75	18.90	10.28	7.44	5.70	20.16
MiniCPM-O-2.6	20.0	6.2	11.33	29.6	9.05	9.55	8.02	22.30	11.24	8.88	8.43	23.76
MiniCPM-O-4.5	18.8	13.47	8.00	34.4	13.60	9.80	10.28	29.15	14.64	10.52	10.00	30.20
Qwen-2.5-Omni-3B	19.8	12.2	5.33	33.6	10.70	7.20	8.86	22.55	12.52	8.20	8.42	24.76
Ming-lite-Omni-1.5	19.6	12.4	4.67	33.4	10.90	8.40	7.92	23.50	12.64	9.20	7.52	25.48
Qwen-2.5-Omni-7B	20.2	15.0	12.00	38.8	12.15	8.99	7.83	24.95	13.76	10.20	8.35	27.72
Qwen-3-Omni-30B-A3B	**25.0**	**22.8**	**20.00**	**53.8**	16.35	12.01	10.19	32.90	18.08	14.18	11.42	37.08
LongCat-Flash-Omni	**25.0**	19.2	18.00	49.6	**16.60**	**16.65**	**14.53**	**40.95**	**18.28**	**17.16**	**14.96**	**42.68**
Qwen3.5-Omni-Flash	29.4	22.20	12.00	55.2	18.45	17.25	13.58	42.90	20.64	18.24	13.39	45.36
Qwen3.5-Omni-Plus	36.6	27.60	22.00	68.8	21.10	22.75	16.32	51.50	24.20	23.72	17.02	54.96
Gemini-2.0-Flash	25.2	19.4	14.67	49.0	15.50	14.05	13.02	35.95	17.44	15.12	13.22	38.56
Gemini-2.5-Flash	31.6	22.6	12.00	57.8	18.35	17.35	16.42	44.40	21.00	18.40	15.87	47.08
Gemini-2.5-Pro	25.6	22.2	21.33	54.2	22.90	25.05	19.43	58.10	23.44	24.48	19.67	57.32
Gemini-3.0-Flash-Preview	36.2	32.2	**24.00**	75.6	26.45	28.85	21.13	66.30	28.40	29.52	21.49	68.16
Gemini-3.0-Pro-Preview	**40.4**	**33.8**	22.67	**81.0**	**28.50**	**33.95**	**24.81**	**75.60**	**30.88**	**33.92**	**24.55**	**76.68**

Table 3: Performance on Multi-Step Open-Ended (MO) Questions. Each MO question comprises multiple sub-questions. We denote the first sub-question as Q1, the second as Q2, and the third and subsequent sub-questions as Q3+. Results were evaluated using our general scoring model.

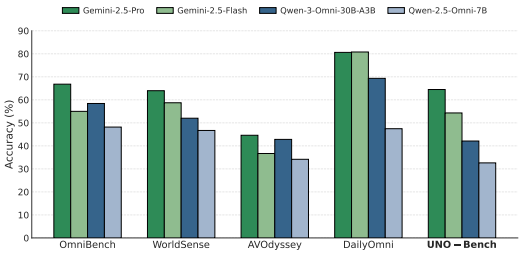


Figure 8: Evaluation of omni-models across five omni-modal benchmarks. The performances among models are more distinguishable on UNO-Bench.

- **Emergent Ability at High Performance:** For stronger models (e.g., Gemini-2.5-Pro, Qwen3.5-Omni-Plus), the curve steepens dramatically. The same uni-modal improvement now yields disproportionately larger omni-modal gains—the quantitative signature of emergent cross-modal capability.

Interpreting the Coefficients. The bias term $b \approx 0.2422$ acts as a performance floor, converging to the 0.25 random-guess accuracy as uni-modal scores approach zero. The scaling coefficient $C \approx 1.0332 \approx 1$ indicates a naturally scaled system, reflecting both the models' intrinsic fusion mechanisms and the coherent design of UNO-Bench.

4.4 Benchmark Analysis

To verify the effectiveness of UNO-Bench, we analyze three aspects. First, as demonstrated in Sec. 4.2.3, MO questions capture the difficulty of complex reasoning and amplify capability gaps between models (full results in Table 3). Second, CGHS achieves SRCC > 0.98 and PLCC > 0.98 at a 10% sampling rate, confirming efficient compression without ranking loss (Appendix C.2). Third, as shown in Figure 8, UNO-Bench maintains 100% answer accuracy with 98% cross-modality solvability, and provides both a clear performance ladder and a meaningful difficulty range among existing omni benchmarks, enabling reliable measurement of genuine omni-modal capability.

5 Conclusion

In this paper, we present UNO-Bench, a unified benchmark that jointly evaluates uni-modal and omni-modal capabilities under a shared ability taxonomy spanning 44 task types and 5 modality combinations. Built on a strict cross-modality solvability design, UNO-Bench enables a controlled investigation of how uni-modal strengths compose into omni-modal understanding, leading to the discovery of a Compositional Law: omni-modal performance follows a super-linear power-law of the product of uni-modal scores, manifesting as a bottleneck effect in weaker models and accelerating cross-modal gains in stronger ones. With strong discriminability, efficient compression, and progressive difficulty, UNO-Bench provides a reliable foundation for measuring and advancing omni-modal intelligence.

Limitations

Scale versus quality. To ensure strict cross-modality dependency and high data quality, we prioritized a human-centric curation process over large-scale automated generation. Although UNO-Bench covers 44 distinct task types, the sample size for certain long-tail tasks may be insufficient to yield statistically significant results at a fine-grained level. To address this, we have subsequently constructed an additional 1,000 high-quality samples following the same annotation protocol, which substantially improves the statistical confidence of fine-grained evaluations; these are included in the released dataset. We further maintain a public leaderboard to continuously incorporate evaluations of newly released omni models. In future work, we aim to develop semi-automated pipelines to further expand the dataset scale without compromising quality.

Generalizability of the Compositional Law. The proposed Compositional Law is empirically derived from 15 representative models. Although the predictive power has been validated on held-out models including Gemini-3-Flash and LongCat-Flash-Omni, the law may require recalibration as model architectures and training paradigms continue to evolve—particularly if future models employ fundamentally different cross-modal fusion mechanisms. We address this through the aforementioned public leaderboard, which will enable periodic recalibration of the law parameters as new models become available. Furthermore, the applicability of this law to other omni-modal data distributions and beyond audio-visual combinations remains to be verified, and the underlying mechanistic interpretation of inter-modal synergy within model representations is left for future work.

Ethics Statement

In the data construction process, we employed LLMs, primarily the Gemini series, to assist in two key aspects. First, we leveraged LLMs to creatively generate certain questions, providing inspiration and guidance for data construction. Second, to ensure appropriate difficulty and solvability, a preliminary model check was conducted on every constructed question.

We employed a professional annotation team from China. All members hold university degrees and possess extensive experience in data annotation. The annotation process was conducted on the AGI-Eval platform. Annotators received an average compensation of $8 USD per hour, a rate significantly exceeding the local average wage. All participants were fully informed that the annotated data would be used for public research release.

To mitigate privacy risks and copyright issues, all multimodal assets (images, audio, and video) were sourced from authorized public benchmarks, royalty-free platforms, or were self-recorded by the annotators. Following data construction, we implemented an enterprise-level content safety audit. This process integrated deep learning algorithms with a professional human review team to rigorously identify and avoid potential violation risks across text, image, video, and audio modalities.

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.

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 Schaekermann, 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*.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, and 1 others. 2026. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*.

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, Yuhan 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*.

Yuying Ge, Sijie Zhao, Jiantao Zhu, Yixiao Ge, Kun Yi, Zehui Hu, Gen Luo, Ying Zhang, Limin Huang, Xiaohu Wang, and 1 others. 2024a. Seed-v1.5: Release of next-generation seed visual language models. *arXiv preprint arXiv:2406.01423*.

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. 2024b. 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?

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

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal 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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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 multi-modal 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.

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 Dahua Lin. 2024b. Mmbench: is your multi-modal model an all-around player?

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. 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.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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, Ramaneswaran 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*.

Meituan LongCat Team. 2025. Longcat-flash-omni technical report. *Preprint*, arXiv:2511.00279.

Qwen Team. 2026. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36:77013–77042.

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*.

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.

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, Xinfa 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-mmbench: 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, Wenhu 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.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

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

Appendix

A Ability Taxonomy Analysis

To dissect the models' core capabilities, we perform a fine-grained analysis based on our proposed ability taxonomy, with detailed results presented in Table.4.

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, LongCat-Flash-Omni achieves the highest perception score 53.68%. Gemini-2.5-Pro leads overall significantly, with both its Alignment 74.35% and Recognition 70.05% exceeding 70%.

In reasoning, Spatial Reasoning is consistently the most challenging task across all models. Even the best-performing Gemini-3.0-Pro only achieves 65.83%, and LongCat-Flash-Omni demonstrates the best spatial reasoning among open-source models with a score of 29.17%.

Overall, reasoning proves to be a more challenging frontier than perception. This is highlighted by the performance gap between the leading proprietary model, Gemini-3.0-Pro, and the best open-source model, LongCat-Flash-Omni. The disparity is 17.15% in Perception (70.83% vs. 53.68%) but widens to a more substantial 32.82% in Reasoning (79.93% vs. 47.11%). This indicates that advanced reasoning remains a key differentiator and a primary bottleneck for current multimodal models.

B Ablation Experiment

B.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.5.

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%), 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 cap-

Model	Perception			Reasoning				Overall
	Cross-modal Alignment	Cross-modal Recognition	Overall	General Reasoning	Temporal Reasoning	Spatial Reasoning	Overall	
Qwen-2.5-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
Qwen-2.5-Omni-7B	38.22	36.41	37.25	28.11	43.94	26.67	29.59	32.60
Qwen-3-Omni-30B-A3B	**53.40**	45.16	49.02	38.06	53.03	26.67	37.41	42.10
LongCat-Flash-Omni	51.31	**55.76**	**53.68**	**50.00**	**62.12**	**29.17**	**47.11**	**49.90**
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
Gemini-3.0-Flash-Preview	67.02	**71.89**	69.61	83.08	83.33	59.17	78.23	74.70
Gemini-3.0-Pro-Preview	70.16	71.43	70.83	**84.08**	80.30	**65.83**	**79.93**	**76.30**

Table 4: Analysis of Omni-MC on ability taxonomy. To simplify the analysis, Cross-modal Recognition refers to the set of other Perception capabilities except Cross-modal Alignment.

ture. 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.

B.2 Ablation Audio Understanding

To isolate the impact of auditory information, we evaluated models under three conditions: visual-only (`Visual`), visual plus transcribed audio (`+Caption`), and the full audio-visual input (`+Audio`). We further divided the audio into three categories: the `Speech` category was annotated with ASR transcripts, while both the `Environment` and `Music` categories received textual descriptions. To ensure the robustness of our analysis and improve statistical reliability, the data-insufficient `Music` class was merged with the `Environment` class. The majority of the transcriptions were manually produced by human annotators, while a smaller subset was generated by a powerful multimodal model. The results are presented in Table.6.

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% Overall) and Qwen-3-Omni-30B-A3B (+17.4% 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.

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	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**

Table 5: Visual modality ablation. Removing visual input (`Audio`-only) collapses performance to near-random levels, confirming cross-modal necessity.

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	**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**

Table 6: Audio modality ablation. Removing audio input (`Visual`-only) collapses performance to near-random levels, confirming cross-modal necessity.

C Benchmark Analysis

C.1 Multi-Step Open-Ended Question Analysis

The complete results of the multi-step open-ended (MO) questions, including the full performance table, are presented in the main paper, Sec. 4.2.3, Table 3. As stated in the main text, each sub-step in our MO questions is bound to information from a specific modality, and the complete reasoning chain requires continuous cross-modal alignment and fusion. Below we provide additional fine-grained observations.

Perception vs. Reasoning breakdown. From the full table, we observe that for all open-source models and lower-tier Gemini models, Perception scores are considerably higher than Reasoning scores. The gap is largest for Qwen-3-Omni-30B (53.8% vs. 32.9%), while it nearly closes for Gemini-3.0-Pro. This indicates that advanced reasoning remains the primary bottleneck for cross-modal understanding.

Depth-wise performance details. The decay from Q1 to Q3+ is consistent across most models. The exception is Gemini-2.5-Pro, which improves from Q1 to Q2 (23.44% $\rightarrow$ 24.48%), suggesting superior context utilization. All other models show monotonic decay.

Model stratification. The MO benchmark clearly separates models into distinct tiers: Gemini-3.0-Pro (top), Gemini-2.5-Pro/Gemini-3.0-Flash (second), Qwen3.5-Omni-Plus (third), and smaller open-source models (bottom), demonstrating its discriminative power.

C.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 sam-

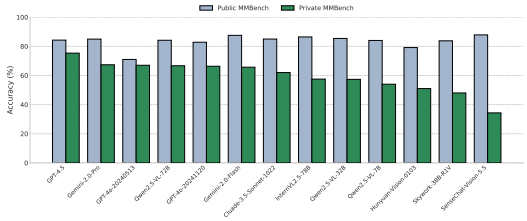


Figure 9: Performance comparison between public MMBench and our private MMBench. With our dataset improvement strategy, the performances among models are more distinguishable.

ples including 18 open-source benchmarks (e.g. MathVista and MMAU, details see Appendix.E) 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.

The experimental result is shown in Figure.10. 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.

C.3 Benchmark Comparison

To ensure the quality of dataset, we conduct quality check on 10%-20% random samples in each benchmarks. As shown in Table.1, UNO-Bench has 100% accuracy on omni-modal dataset while 98% questions requires cross-modality to solve. It shows the highest quality among existing omni benchmarks.

An effective benchmark must provide both a clear performance ladder and a meaningful difficulty range. As shown in Figure.8, UNO-Bench is engineered to deliver on both fronts. It excels in discriminability, establishing substantial and remarkably linear intervals of ~10%-12% between adjacent models. This superior discriminability comes from a well-calibrated difficulty. UNO-Bench creates a vast 31.9% 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 AV-Odyssey 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 to gage genuine progress in the field.

D Dataset Construction Pipeline

D.1 Material Collection

In both data quality checks and experimental results, we found that the natural video with audio-visual synchronized data contains a large amount of information redundancy, only a few videos require both audio and visual modality simultaneously. Therefore, we begin with carefully designed material collection.

Our materials have the following three characteristics:

Diverse Sources. The majority of our materials are real-world photos and videos collected through crowd sourcing, and another portion sourced from copyright-free websites. Additionally, a small fraction comes from high-quality public datasets such as MMVU(Zhao et al., 2025), LongVideoBench(Wu et al., 2024), and Video-Vista(Chen et al., 2025).

Rich and Diverse Topics. Our materials cover a broad spectrum of subjects, including society, culture, art, life, literature, science, and so on.

Live-Recorded Audio. Apart from background sounds and music, all dialogue is recorded by human speakers. With over 20 participants covering Mandarin and regional dialects, the audio features closely reflect real-world vocal diversity (see also Section 3.1).

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.

D.2 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 multimodal model interactions

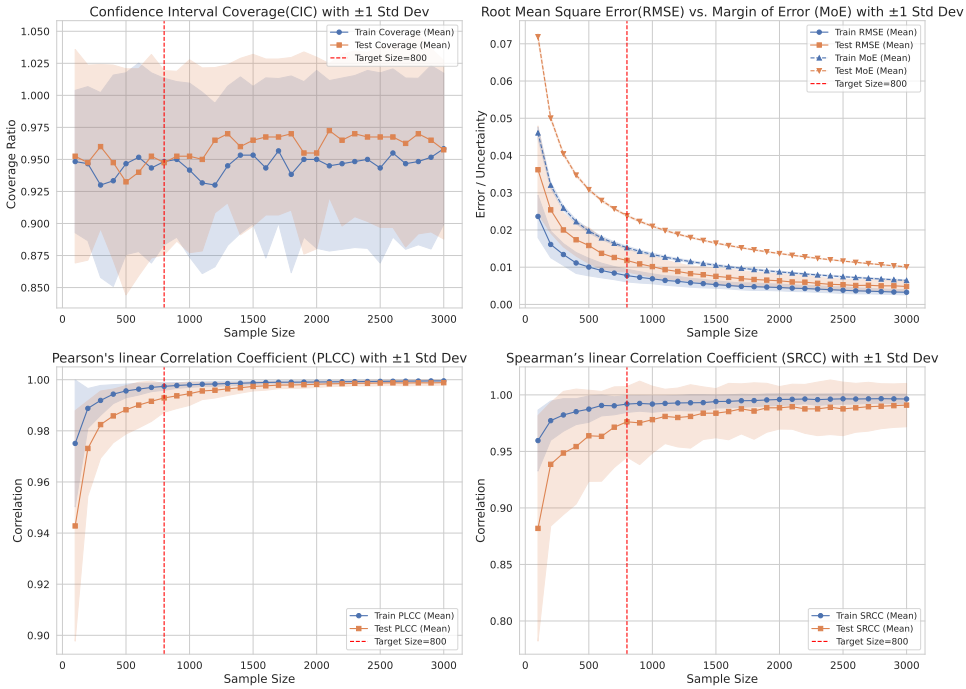


Figure 10: Data compression performance. The four subplots illustrate the evolution of CIC, RMSE/MoE, PLCC, and SRCC metrics as the sample size increases. The vertical red line highlights the chosen sample size (N=800), balancing data reduction with ranking consistency. Shaded areas denote ±1 standard deviation for training (blue) and test (orange) sets.

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. Finally, 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-centric** approach, ensuring continuous manual involvement from the initial data sourcing to the final output. This methodology not only prevents data leakage but also more accurately simulates real-world scenarios. The dataset further includes Chinese-specific tasks such as Mandarin ASR, Chinese OCR, and dialogue coreference resolution, addressing the English-centric bias of most existing benchmarks.

D.3 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. **Model Check**, 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. **Ablation Study**, 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. **Human Check**, finally manual quality inspection and revision are performed.

E 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 in-house benchmarks that assess a range of capabilities, including

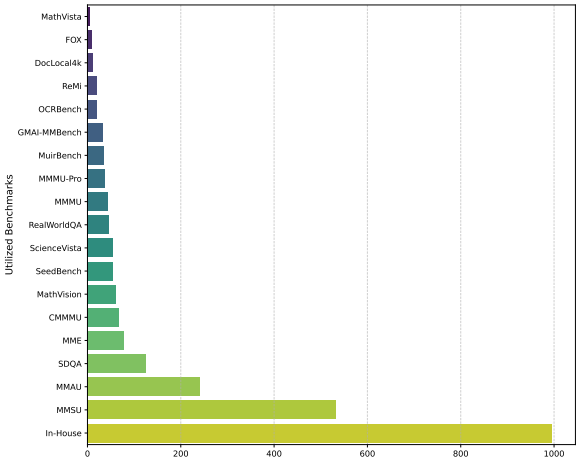


Figure 11: The distribution of the uni-modal benchmarks in UNO-Bench. In addition to publicly available benchmarks, we incorporated several in-house benchmarks both before and after compression to ensure the reasonableness of the data distribution.

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. The public benchmarks utilized in this work are distributed under open-source licenses (e.g., Apache-2.0, MIT, CC-BY, and CC-BY-NC), and we adhere to their specific terms and intended usage. For the in-house benchmarks, we ensure they are constructed and used following standard ethical guidelines for data collection.

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

F Rigorous Derivation of the Compositional Law

Defining the **performance gain** as $\mathcal{P}'_{\text{Omni}} = \mathcal{P}_{\text{Omni}} - b$. From Eq. 1, we have:

$$\mathcal{P}'_{\text{Omni}}(\mathcal{P}_{\text{A}}, \mathcal{P}_{\text{V}}) = f_{\text{A}}(\mathcal{P}_{\text{A}}) + f_{\text{V}}(\mathcal{P}_{\text{V}}) + f_{\text{I}}(\mathcal{P}_{\text{A}}, \mathcal{P}_{\text{V}}) \tag{3}$$

Due to the high quality of our benchmark, where a task is unsolvable if either modality is absent, causing the performance to drop to its baseline (e.g. random guessing). we can have a strict boundary condition:

$$\begin{aligned} \mathcal{P}'_{\text{Omni}}(\mathcal{P}_{\text{A}}, 0) = 0 \quad &\text{and} \\ \mathcal{P}'_{\text{Omni}}(0, \mathcal{P}_{\text{V}}) = 0 \quad &\text{and} \\ \mathcal{P}'_{\text{Omni}}(0, 0) = 0 \quad & \end{aligned} \tag{4}$$

Applying the boundary condition of Eq. 4 to Eq. 3, we find that the gain is a second-order mixed difference of f_{I}:

$$\begin{aligned} \mathcal{P}'_{\text{Omni}}(\mathcal{P}_{\text{A}}, \mathcal{P}_{\text{V}}) = & f_{\text{I}}(\mathcal{P}_{\text{A}}, \mathcal{P}_{\text{V}}) - f_{\text{I}}(\mathcal{P}_{\text{A}}, 0) \\ & - f_{\text{I}}(0, \mathcal{P}_{\text{V}}) + f_{\text{I}}(0, 0) \end{aligned} \tag{5}$$

Substituting the Taylor series of f_{I} around $(0, 0)$ into Eq. 5, the performance gain is thus exactly equal to the sum of all pure interaction terms from f_{I}:

$$\mathcal{P}'_{\text{Omni}}(\mathcal{P}_{\text{A}}, \mathcal{P}_{\text{V}}) = \sum_{i \geq 1, j \geq 1} c_{ij} \mathcal{P}_{\text{A}}^{i} \mathcal{P}_{\text{V}}^{j} \tag{6}$$

where the coefficients c_{ij} are constants derived from the partial derivatives of f_{I} at the origin. For sufficiently small uni-modal performances, we can approximate this series by its leading-order term:

$$\mathcal{P}'_{\text{Omni}}(\mathcal{P}_{\text{A}}, \mathcal{P}_{\text{V}}) \approx c_{11} \mathcal{P}_{\text{A}} \mathcal{P}_{\text{V}} \tag{7}$$

This result strongly motivates modeling the interaction with the general multiplicative Cobb-Douglas form. Re-introducing the baseline b yields our final Compositional Law:

$$\mathcal{P}_{\text{Omni}} = C \cdot \mathcal{P}_{\text{A}}^{\alpha} \mathcal{P}_{\text{V}}^{\beta} + b \tag{8}$$

where C is a scaling constant, and exponents α, β model the interaction's elasticity.

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 $\alpha = \beta$. Substituting this into Eq. 8, we arrive at the **Omni-modal Compositional Law**:

$$\mathcal{P}_{\text{Omni}} = C \cdot (\mathcal{P}_{\text{A}} \times \mathcal{P}_{\text{V}})^{\alpha} + b \tag{9}$$

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

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.

G 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., $P_{Audio}^{-1.59}$) 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.

Table 8: Accuracy comparison across question types. Our scoring model outperforms baselines in the multi-step open-ended question type.

Model	Single-step		Multi-step
	EVOUNA-TQ	EVOUNA-NQ	In-house
Seed-1.5-VL	**95.90**	89.30	91.18
GPT-4.1	94.50	88.60	94.57
Ours	95.50	**89.30**	**95.05**

H General Scoring Model

We employ Qwen3-14B (Yang et al., 2025) as the backbone for our general scoring model. This model is distributed under the Apache-2.0 license, and our usage is consistent with its intended terms. 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 Table.9. We constructed a fine-tuning dataset consisting of 13,000 samples. The model was trained for 3 epochs using the Llama-Factory framework (Zheng et al., 2024). The training process was conducted on a compute node equipped with 8 NVIDIA A100 (80GB) GPUs and was completed in approximately 3 hours. For inference, we deployed the scoring service using vLLM (Kwon et al., 2023) to optimize throughput, achieving a processing speed of approximately 1 sample per second. To facilitate reproducibility, the complete evaluation code for UNO-Bench and the scoring model are publicly available in our github repository.

To validate the precision of our proposed general scoring model, we conducted a comparative analysis against state-of-the-art models, including Seed-1.5-VL(Ge et al., 2024a) and GPT-4.1(OpenAI, 2023). The evaluation covers two distinct domains:

- **Single-step:** We utilize EVOUNA to evaluate the scoring accuracy on single-step questions(e.g., multiple-choice questions).

- **Multi-step:** We utilize a curated internal dataset to evaluate the model's performance on complex, multi-step open-ended questions.

The detailed accuracy comparison is presented in Table 8. Our model demonstrates superior performance, achieving **95.05%** accuracy on the internal multi-step dataset and consistently competitive results on public benchmarks. Furthermore, our scoring model benefits from a smaller size, offering significant cost advantages.

I 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 Table.10, and the Reasoning dimension can be seen in Table.11.

Specific examples are provided for each model ability. Examples of Perception ability including Object Perception, Spatial Perception, Cross-Modal Alignment, Attribute Perception, Scenario Perception, Cross-Modal Conversion and Semantic Understanding can be seen in Figure.12. Examples of Reasoning ability including Complex Reasoning, Temporal Reasoning, Spatial Reasoning, Life Reasoning, STEM Reasoning and Code can be respectively seen in Figure.13.

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.

Table 9: Definition of finer question types for general scoring model.

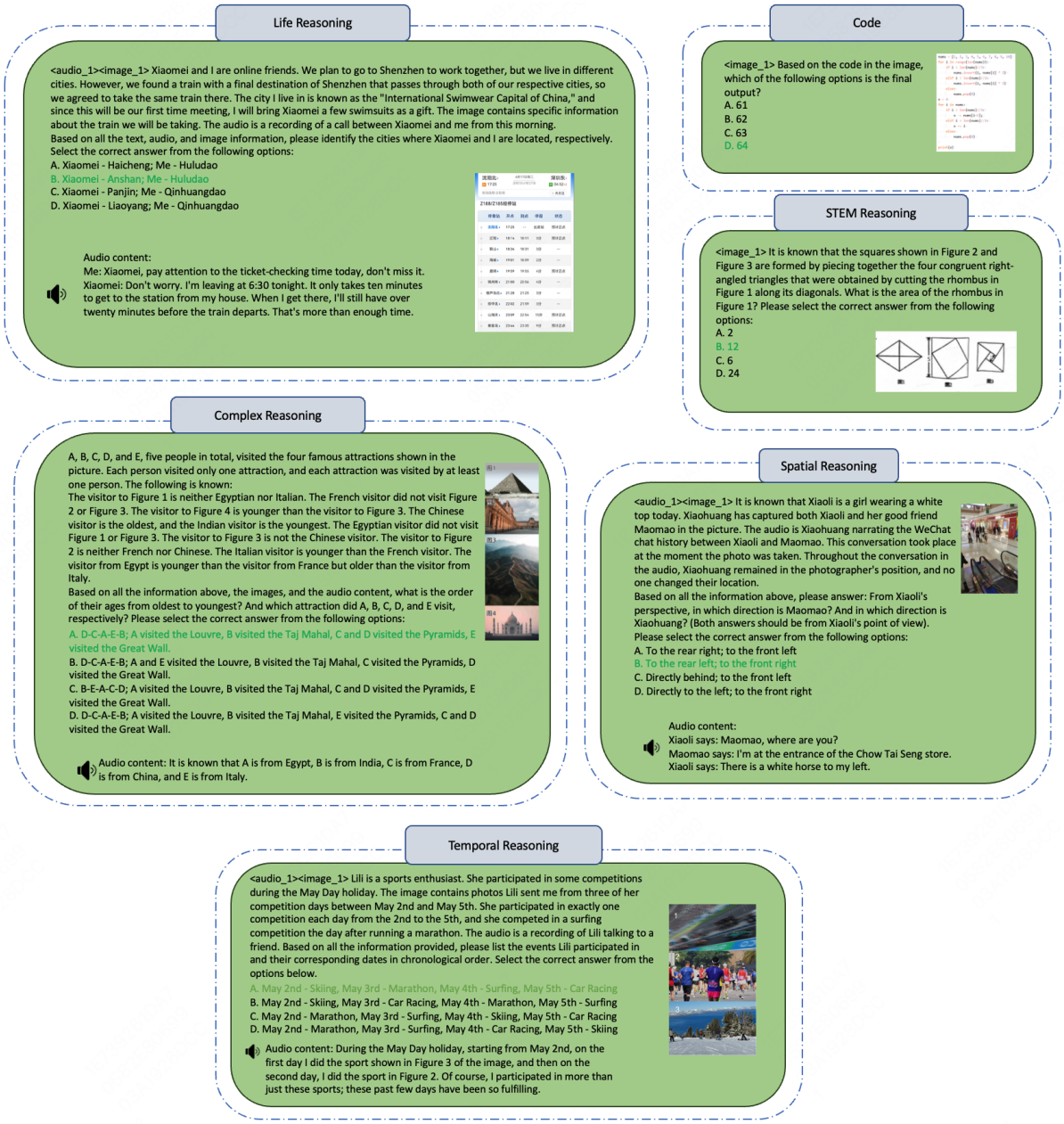
Model Ability Taxonomy	Task Type	Definition
Object Perception	Human and Animal Recognition	Recognize persons or animals by combining information from different modalities.
	Other Entity Recognition	Recognize other entities by combining information from different modalities, for example, plants, daily 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.
	Feature Recognition	Recognize all entity-related attributes, such as color, size, material, etc.
	Gender Judgment	Determine a person's gender by different timbres.
Semantic Understanding	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.

Table 10: 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.
Spatial Reasoning	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.
	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.
Temporal Reasoning	Event Analysis	Analyze the causes and effects of events by combining information from different modalities.
	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.
General Reasoning	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.
	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.
STEM Reasoning	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.
	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.

Table 11: Definition of the Reasoning Dimension.

Figure 12: Example of each ability in perception dimension.


Life Reasoning
<audio_1><image_1> Xiaomei and I are online friends. We plan to go to Shenzhen to work together, but we live in different cities. However, we found a train with a final destination of Shenzhen that passes through both of our respective cities, so we agreed to take the same train there. The city I live in is known as the "International Swimwear Capital of China," and since this will be our first time meeting, I will bring Xiaomei a few swimsuits as a gift. The image contains specific information about the train we will be taking. The audio is a recording of a call between Xiaomei and me from this morning.
Based on all the text, audio, and image information, please identify the cities where Xiaomei and I are located, respectively.
Select the correct answer from the following options:
A. Xiaomei - Haicheng; Me - Huludao
B. Xiaomei - Anshan; Me - Huludao
C. Xiaomei - Panjin; Me - Qinhuangdao
D. Xiaomei - Liaoyang; Me - Qinhuangdao
Audio content:
Me: Xiaomei, pay attention to the ticket-checking time today, don't miss it.
Xiaomei: Don't worry. I'm leaving at 6:30 tonight. It only takes ten minutes to get to the station from my house. When I get there, I'll still have over twenty minutes before the train departs. That's more than enough time.
Code

STEM Reasoning