Omni-RewardBench: Toward a Comprehensive Evaluation of Generative Reward Models Across Modalities

Chi-Min Chan[1*] **Yujin Zhou**[1*] **Pengcheng Wen**[1] **Boqin Yin**[1]
Jiaming Ji[2] **Juntao Dai**[2] **Wei Xue**[1] **Sirui Han**[1†] **Yike Guo**[1†]
[1]The Hong Kong University of Science and Technology
[2]Peking University
cchanbc@connect.ust.hk

Abstract

The rise of Omni-modality Large Language Models (OLLMs) capable of jointly processing text, audio, and visual inputs marks a major step toward general intelligence. Ensuring their alignment with human preferences requires effective Omni-modality Reward Models (ORMs), which serve as surrogates for human judgment to guide OLLMs behavior. However, ORMs evaluation remains underdeveloped in the previous literature. Existing benchmarks are largely text-centric or limited to bimodal tasks, restricting comprehensive assessment for ORMs. To bridge this gap, we introduce *Omni-RewardBench*, the first benchmark for comprehensive evaluation of ORMs across modalities. In short, our contributions are threefold: (1) a hybrid automatic-annotation and human-verification pipeline to construct high-quality evaluation data; (2) extensive experiments on 20+ models, including inherently omni-modal and modality-bridged systems. Our experimental results demonstrate that current OLLMs fall short as reward models, revealing several common failure modes such as **perception failure**, **modality dominance failure**, and **cross-modal fusion failure**. and (3) strong correlations between *Omni-RewardBench* scores and downstream performance (IID r = 0.94, OOD r = 0.72), validating its reliability as a predictor of real-world capability and alignment quality.

1 Introduction

The rapid evolution of artificial intelligence has led to the emergence of multimodal large language models (MLLMs) capable of processing heterogeneous data modalities, including text, audio, and image/videos (Hurst et al., 2024; Team et al., 2023; Anthropic, 2025). This advancement reflects the multimodal nature of human perception, where textual, auditory, and visual cues jointly support richer understanding (Turk, 2014; Jaimes and Sebe, 2007). Building upon the transformative success of large language models (LLMs) (Jaech et al., 2024; Grattafiori et al., 2024; Guo et al., 2025; Li et al., 2024a), recent research extends these capabilities to multimodal reasoning, emphasizing cross-modal alignment and joint representation learning (Bai et al., 2025; Xu et al., 2025a; Wu et al., 2024). Although significant progress has been achieved in bi-modal systems such as vision-language and audio-language models, the unified integration of omni-modalities (textual, audio and visual) remains a promising yet insufficiently explored frontier.

Advancing OLLMs requires not only architectural innovation but also robust evaluation methodologies that comprehensively assess model responses (Bommasani et al., 2021). Rigorous evaluation is essential for ensuring deployment reliability and guiding iterative optimization (Wu, 2025). Reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) exemplifies this principle by aligning model outputs with human intentions. The growing need for reliable evaluation has driven the development of specialized benchmarks to assess reward models' ability to distinguish response quality (Lambert et al., 2024; Wang et al., 2024d; Li et al., 2025a; Yasunaga et al., 2025; Ji et al., 2025). However, current benchmarks remain constrained to unimodal or bimodal domains, underscoring the necessity for a comprehensive omni-modal reward model benchmark that can effectively evaluate and advance the next generation of ORMs.

To address this gap, we propose *Omni-RewardBench*, a comprehensive benchmark specifically designed to evaluate reward models within omni-modal contexts. Adopting the standard triplet (question, chosen response, rejected response) format (Lambert et al., 2024; Li et al., 2025a; Yasunaga et al., 2025), our benchmark requires re-

[*] Equal contribution.
[†] Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13962–13984
July 2-7, 2026

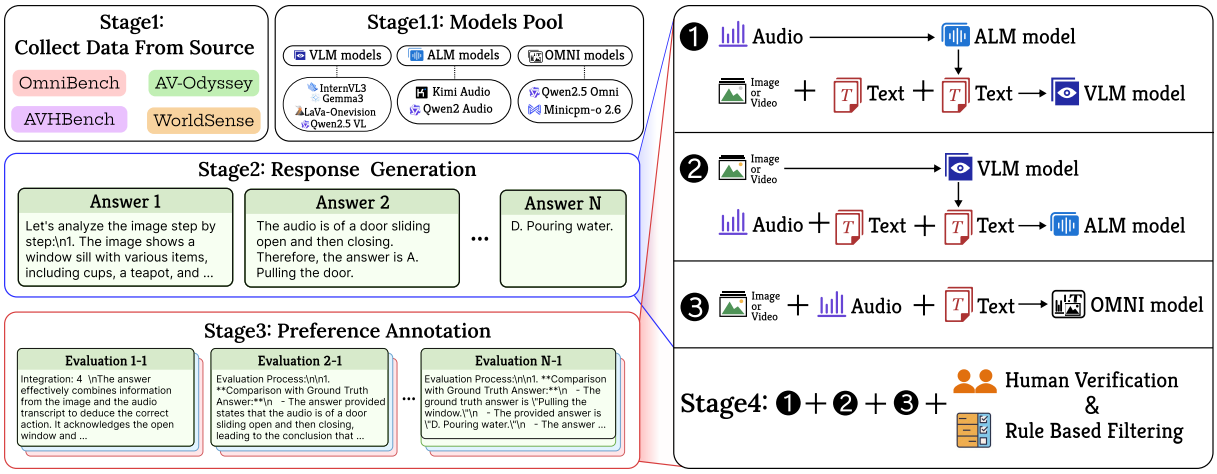

Figure 1: Pipeline for constructing *Omni-RewardBench*.

ward models to correctly rank responses by assigning higher rewards to chosen response and lower rewards to rejected response, thereby identifying their preference ordering. In order to acquire high-quality omni-modal triplet, we propose a novel data annotation pipeline that combines scalable automated annotation with rigorous human verification, as detailed in Section 3.2 and curate **1,375** triplets in total. These samples enable comprehensive evaluation of ORMs' discriminative capabilities across diverse categories (**8** categories and **22** subcategories), thereby establishing a foundational benchmark to guide the development of future ORMs.

Due to the current scarcity of models capable of inherently processing omni-modal inputs, our evaluation further incorporates a modality bridging framework as a practical interim solution. Specifically, this framework addresses unsupported modalities by converting them into textual descriptions through specialized models—such as transcribing audio using audio-language models (ALMs) or captioning visual inputs using vision-language models (VLMs). This enables bi-modal models to effectively participate in omni-modal evaluations.

Empirically, we conduct extensive experiments across 20+ cutting-edge proprietary and open-source models including Gemini (Comanici et al., 2025), GPT-4o (Hurst et al., 2024), Qwen (Xu et al., 2025a), InternVL (Chen et al., 2024b), Kimi (Ding et al., 2025a) and Deepseek series (Wu et al., 2024). Through both quantitatively and qualitatively analysis, our work makes following key contributions:

1. **First omni-modal reward model benchmark**: A curated collection of **1,375** high-quality triplets spanning **8** categories and **22** subcategories, enabling fine-grained evaluation for ORMs.

2. **Modality bridging annotation pipeline**: Combines automated processing with human verification to overcome data scarcity challenges caused by limited OLLMs availability.

3. **Comprehensive performance analysis**: We reveal that current OLLMs fall short as reward models, while demonstrating the effectiveness of a modality-bridging framework as a viable interim solution. Moreover, we identify several common failure modes of ORMs, providing insights and guidance for further optimization of ORMs.

4. **High correlation with downstream tasks**: Demonstrates that performance on *Omni-RewardBench* strongly correlates with downstream tasks performance under inference-time scaling paradigm, validating its reliability as a predictor of real-world capability and alignment quality.

2 Related Works

2.1 Multi-Modal Models Evaluation Benchmarks

With the rapid advancement of multimodal models, the community now demands more challenging benchmarks to rigorously evaluate the progress of existing multimodal systems. Vision-language benchmarks have seen particularly notable development, covering key areas such as (1) spatial perception (Liu et al., 2024; Yu et al., 2023; Li et al., 2024b; Yakun et al., 2025), (2) reasoning skills (Lu

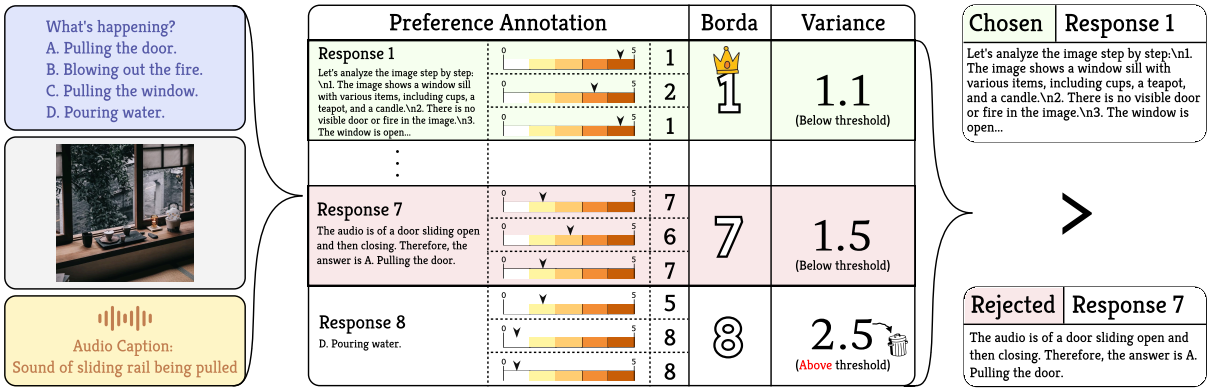

Figure 2: Illustration of variance-controlled borda count ranking method.

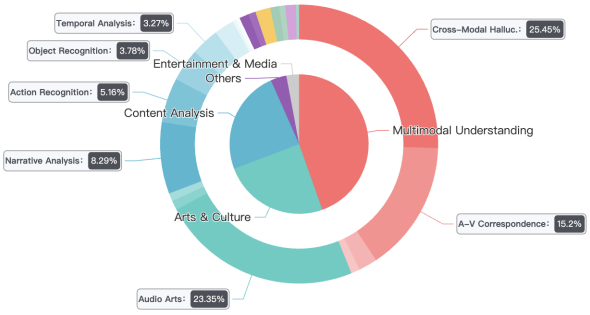

Figure 3: Illustrative taxonomy of *Omni-RewardBench*. Full list shown in Appendix B.2.

et al., 2023; Zhang et al., 2024; Wang et al., 2024b; Yue et al., 2024; Yakun et al., 2026) and (3) instruction following (Bitton et al., 2023; Qian et al., 2024; Ding et al., 2025b). Similarly, benchmarks for audio-language models assess capabilities like (1) speech recognition (Gong et al., 2022; Bu et al., 2017), (2) audio question answering (Penamakuri et al., 2025; Lipping et al., 2022; Wang et al., 2024a) and (3) music understanding (Yang et al., 2024; Weck et al., 2024; Agostinelli et al., 2023). Although these efforts have significantly advanced multimodal benchmarking, most focus on dual-modal interactions (e.g., vision-language or audio-language). With the rise of advanced OLLMs (Chen et al., 2025; Xie and Wu, 2024; Luo et al., 2025b; Fu et al., 2025; Zhang et al., 2025), the development of comprehensive, omni-modality benchmarks has become crucial. Recent initiatives like (Ji et al., 2024; Li et al., 2024c; Chen et al., 2024a) evaluate OLLMs' cross-modal reasoning abilities, primarily focusing on their problem-solving performance. In contrast, *Omni-RewardBench* provides a systematic framework to quantify the reliability of ORMs to provide reward signals, a critical component for training and aligning omni-modal AI systems.

2.2 Reward Models Evaluation Benchmarks

Reward models are critical components in RLHF (Bai et al., 2022); their evaluation serves as a fundamental aspect in validating system performance. Existing literature on benchmarking reward models has predominantly focused on uni-modal (text-only) or bi-modal (vision-language) settings (Lambert et al., 2024; Li et al., 2025a), with current approaches bifurcating into scalar-based reward models (Bradley and Terry, 1952) and generative reward models (Mahan et al., 2024). Scalar-based methods further divide into outcome-oriented models (Lambert et al., 2024; Wang et al., 2024d; Tu et al., 2025; Yasunaga et al., 2025) that assess overall response quality through preference ranking, and process-oriented variants (Luo et al., 2025a; Zheng et al., 2024; Song et al., 2025; Xu et al., 2025b; Chan et al., 2025b; Zhou et al., 2026) that evaluate stepwise reasoning fidelity. Concurrently, generative reward models employ LLM-as-a-judge frameworks (Pu et al., 2025; Tan et al., 2024; Ruan et al., 2025; Li et al., 2025a; Chan et al., 2025a) to produce either scalar assessments or pairwise comparisons through carefully designed prompting strategies. While these benchmarks have established robust evaluation methodologies, they remain fundamentally limited by their uni or bi-modal nature, failing to address the critical challenges of omni-modal preference learning that emerge in complex, real-world AI systems.

3 Omni-RewardBench

Our design philosophy for *Omni-RewardBench* is to construct a benchmark that can evaluate ORMs' capabilities to distinguish good responses from omni-modal inputs, where we define omni-modal

Algorithm 1 Variance-Controlled Borda Count Ranking

Require: $R = \{r_1, r_2, \ldots, r_n\}$: set of responses;
Require: $E = \{e_1, e_2, \ldots, e_m\}$: set of evaluators;
Require: $\{S^{(j)}(r_i)\}$: scores by evaluator e_j for response r_i;
Require: σ_{thresh}: variance threshold; $k_{\text{top}}, k_{\text{bottom}}$: number of top/bottom responses to select
Ensure: R_{top}: selected top responses; R_{bottom}: selected bottom responses

Step 1: Calculate Borda Count for each response
for each response $r_i \in R$ **do**
  $BC(r_i) \leftarrow 0$
  for each evaluator $e_j \in E$ **do**
    $rank_{ji} \leftarrow \text{GETRANK}(r_i, S^{(j)})$ {rank of r_i in evaluator j's ranking}
    $BC(r_i) \leftarrow BC(r_i) + (n - rank_{ji})$ {Borda count increment}
  end for
end for
Step 2: Calculate score variance for each response
for each response $r_i \in R$ **do**
  $scores_i \leftarrow [S^{(1)}(r_i), S^{(2)}(r_i), \ldots, S^{(m)}(r_i)]$
  $\sigma^2(r_i) \leftarrow \text{VARIANCE}(scores_i)$
end for
Step 3: Filter responses by variance threshold
$R_{\text{filtered}} \leftarrow \{r_i \in R : \sigma^2(r_i) \leq \sigma_{\text{thresh}}\}$
Step 4: Sort filtered responses by Borda Count
$R_{\text{sorted}} \leftarrow \text{SORTBYBORDACOUNT}(R_{\text{filtered}}, \text{descending} = \text{True})$
Step 5: Select top and bottom responses
if $|R_{\text{sorted}}| < k_{\text{top}} + k_{\text{bottom}}$ **then**
  return "Insufficient low-variance responses"
end if
$R_{\text{top}} \leftarrow R_{\text{sorted}}[1 : k_{\text{top}}]$
$R_{\text{bottom}} \leftarrow R_{\text{sorted}}[-k_{\text{bottom}} :]$
return $R_{\text{top}}, R_{\text{bottom}}$

Supporting Functions:

Function $\text{GETRANK}(response, evaluator\_scores)$:
  $sorted\_responses \leftarrow \text{SORTBYSCORE}(evaluator\_scores, \text{descending} = \text{True})$
  return $\text{POSITION}(response, sorted\_responses)$ {1-based ranking}

Function $\text{VARIANCE}(scores)$:
  $\mu \leftarrow \text{MEAN}(scores)$
  return $\frac{1}{|scores|} \sum_{score \in scores} (score - \mu)^2$

Function $\text{SORTBYBORDACOUNT}(responses, descending)$:
  return $\text{SORT}(responses, \text{key} = BC, \text{reverse} = descending)$

inputs that contain three modalities (textual, auditory, visual (image or video)) simultaneously. We achieve this by meticulously accumulating open source omni-modal inputs from the community and propose a carefully designed response generation and preference annotation pipeline as detailed in Section 3.2.

3.1 Task Definition

We evaluate reward models based on their ability to correctly predict preferences for response pairs given omni-modal inputs. Specifically, our evaluation adopts the LLM-as-a-judge paradigm, where the model is prompted to compare pairs of responses or directly rate a single response and finally determine the preferable one.

Given a response triplet $(question, r_{\text{chosen}}, r_{\text{rejected}})$, the pair-wise evaluation prompts the model to compare both responses simultaneously and determine the preferred one:

$$r_{\text{pair-wise}} = \text{ORM}_p(\text{question}, r_{\text{chosen}}, r_{\text{rejected}}). \quad (1)$$

In contrast, the direct evaluation method independently assesses each response in isolation, assigning scores separately as follows:

$$r_{\text{direct}} = \text{ORM}_d(\text{question}, r_{\text{chosen||rejected}}) \quad (2)$$

Then, the accuracy for each pair is calculated as:

$$\text{Accuracy} = \begin{cases} 1, & \text{if } r_{\text{pair-wise}} = r_{\text{chosen}} \succ r_{\text{rejected}}, \\ 1, & \text{or } r_{\text{direct}[chosen]} \succ r_{\text{direct}[rejected]}, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

3.2 Benchmark Curation Pipeline

In this section, we introduce the pipeline for constructing *Omni-RewardBench*, a meticulously curated omni-modal preference benchmark. Our methodology consists of four key stages designed to ensure both scalability and reliability in preference annotation across multiple modalities.

Stage 1: Omni-Modal Data Collection We systematically collect omni-modal questions from prominent open-source communities, focusing on prompts that simultaneously incorporate textual, auditory and visual inputs. Our data collection encompasses four major datasets: OmniBench (Li et al., 2024c), AV-Odyssey (Gong et al., 2024), WorldSense (Hong et al., 2025), and AVHBench (Sung-Bin et al., 2024). These datasets are selected to ensure comprehensive coverage of real-world multi-modal scenarios.

As shown in Figure 3, the collected data span 8 major categories and 22 subcategories, providing

Model Configuration	Pairwise Accuracy (%)				Direct Accuracy	
	Original	Shuffle	Avg	|Bias|	Score (%)	Rank Diff
Omni Model						
Ola-7B	74.76	74.18	74.47	0.78	75.78	+1
Video-Llama2	68.65	69.16	68.91	0.74	62.83	+3
Vita-1.5	65.16	67.05	66.10	2.86	68.43	+5
Uio2-Large	42.03	42.25	42.14	0.52	14.83	0
Qwen-Omni-7B	74.62	73.45	74.03	1.58	54.76	-3
Qwen-Omni-3B	69.31	72.14	70.72	4.00	42.47	-3
Baichuan-Omni	69.16	70.61	69.88	2.07	54.76	-1
MiniCPM-2.6-O	70.34	71.63	70.98	1.82	61.53	-1
Gemini-2.5-Flash	86.18	81.45	83.81	5.64	71.92	-1

Table 1: Performance of omni-modal models.

a broad domain coverage across different types of reasoning tasks. The categories encompass multimodal understanding, content analysis, art and culture, technology and science, among others.

Stage 2: Response Generation via Modality Bridging For each question, we generate diverse responses using a carefully designed modality bridging approach as introduced in Section 1. Given that only a limited number of current models support processing three modalities simultaneously, we leveraged powerful bi-modal models by employing a modality bridging strategy, which converts unsupported modalities into textual captions, thereby effectively enabling tri-modal processing. This approach allows us to utilize the strongest available models for each modality combination, thereby generating more diverse and higher-quality responses than would be possible with limited tri-modal systems alone.

Specifically, we employ text as a unified medium to bridge modality gaps. For example, we convert audio inputs to textual captions for vision-language models and convert images/videos to descriptive captions for audio-language models. This strategy enable us to harness the capabilities of state-of-the-art bi-modal models while maintaining the omni-modal nature of our evaluation.

In this stage, we utilized the following eight different models to generate candidate responses: two Omni models (Qwen-Omni-3B, Qwen-Omni-7B); two Audio Language models (Qwen2-Audio-7B-Instruct, Qwen-Audio-Chat); and four Vision Language models (Qwen2.5-VL-3B, Qwen2.5-VL-72B, InternVL2.5-4B, InternVL2.5-78B).

Stage 3: Automated Preference Annotation with Variance-Controlled Borda Count Ranking Following response generation, we implement an automated preference annotation system designed to collect reliable preference scores. For each question and its corresponding responses, we generate multiple preference annotations using the same modality bridging methodology described above.

In this stage, we employed a different set of eight distinct evaluators to mitigate model specific biases. These include two Omni models (Qwen-Omni-7B, MiniCPM-o-2.6); two Audio Language models (Qwen2-Audio-7B-Instruct, Kimi-Audio-7B); and four Vision Language models (InternVL3-78B, Qwen2.5-72B, LLaVA-OneVision-Qwen2-72B-OV, Gemma-3-27B)

Therefore, the above two stages result in $8 * 8 = 64$ total preference scores per response.

Recognizing that annotation variance is a key indicator of reliability, we developed a variance-controlled Borda count ranking algorithm. The intuition behind this approach is that high variance in annotations reflects disagreement among annotators, signaling unreliable scoring. Additionally, simple averaging of absolute scores may be inadequate due to differing scoring scales across models. To address this, we propose a score-insensitive ranking aggregation method based on the Borda count. The full process is detailed in Algorithm 1 and Figure 2.

Model Configuration		Pairwise Accuracy (%)				Direct Accuracy	
Policy Model	**Caption Model**	**Original**	**Shuffle**	**Avg.**	**|Bias|**	**Score (%)**	**Rank Diff**
Modality Bridging (AL Caption for VL Model)							
InternVL-3-2B	Qwen2-Audio-Instruct	42.32	44.72	43.52	5.51	66.76	+10
InternVL-3-8B	Qwen2-Audio-Instruct	76.14	75.05	75.59	1.44	73.13	+7
InternVL-3-14B	Qwen2-Audio-Instruct	77.45	77.52	77.48	0.09	68.87	+1
InternVL-3-38B	Qwen2-Audio-Instruct	78.90	77.53	78.22	1.75	70.18	-1
InternVL-3-78B	Qwen2-Audio-Instruct	79.12	78.54	78.83	0.74	75.93	+5
Qwen-2.5-VL-3B	Qwen2-Audio-Instruct	64.29	65.81	65.05	2.34	37.52	-2
Qwen-2.5-VL-7B	Qwen2-Audio-Instruct	66.61	68.14	67.38	2.27	61.09	+1
Qwen-2.5-VL-32B	Qwen2-Audio-Instruct	75.85	75.20	75.53	0.86	63.34	-3
Qwen-2.5-VL-72B	Qwen2-Audio-Instruct	76.54	75.56	76.05	1.29	67.27	-1
DeepSeek-VL2	Qwen2-Audio-Instruct	74.25	74.54	74.40	0.39	62.18	-2
DeepSeek-VL2-Small	Qwen2-Audio-Instruct	69.74	70.54	70.14	1.14	54.98	-2
DeepSeek-VL2-Tiny	Qwen2-Audio-Instruct	39.85	40.07	39.96	0.55	0.65	0
GPT-4o	Qwen2-Audio-Instruct	82.54	81.16	81.85	1.69	75.34	0
Modality Bridging (VL Caption for AL Model)							
Qwen2-Audio-7B	QwenVL-2-2B	64.94	66.83	65.88	2.87	25.74	-1
Qwen-Audio-Chat	QwenVL-2-2B	61.81	61.81	61.81	0.00	22.47	-3
Kimi-Audio	QwenVL-2-2B	56.29	54.69	55.49	2.88	70.83	+4

Table 2: Performance of bi-modal models utilizing modality-bridging method. We provide more results on using different caption models in Appendix C.1.

Our approach aggregates preference annotations while accounting for variance by filtering out responses with variance exceeding a predefined threshold. For each triplet, the highest-scoring response is selected as the chosen response, and the lowest-scoring response as the rejected response.

Stage 4: Quality Assurance with Human Verification The next stage employs a rigorous human verification process to ensure annotation quality and to mitigate potential biases from automated scoring, such as model self-preference or length bias. During this stage, we recruit graduate researchers specializing in Computer Science who are highly proficient in the protocols of omni-modal evaluation.

Expert annotators are required to independently review each triplet and are instructed *not* to assume that the highest-scoring response is the "chosen" one. They evaluate all responses and their preference scores for correctness and reasoning quality, revising or discarding any ranking that conflicts with human judgment. Only samples that exhibit clear human agreement on the rankings and contain no obvious errors are retained. This verification stage provides a critical safeguard against noisy cases, ensuring that *Omni-RewardBench* remains a reliable and robust benchmark.

4 Experiments

We conduct extensive experiments to evaluate the performance of various models on the proposed *Omni-RewardBench*. As is also mentioned in Section 1, considering that only a limited number of current models support processing three modalities, our evaluation strategy encompasses not only omni-modal models but also bi-modal models by utilizing "modality bridging" method.

Since the LLM-as-a-judge paradigm may exhibit position bias (Zheng et al., 2023), where the ordering of responses influences the model's preference decision, we evaluate each model under two distinct conditions: *Original* and *Shuffled*. Specifically, in the original condition, responses are presented in one order, whereas in the shuffled condition, the order is reversed. By measuring performance under these two presentation orders, we aim to gauge the robustness and consistency of ORM's judgments. To quantify the potential position bias, we also calculate the bias magnitude as follows:

$$|Bias| = \frac{|\text{Original} - \text{Shuffled}|}{(\text{Original} + \text{Shuffled})/2} \times 100. \quad (4)$$

Moreover, to provide a comprehensive assessment beyond pairwise comparisons, we also per-

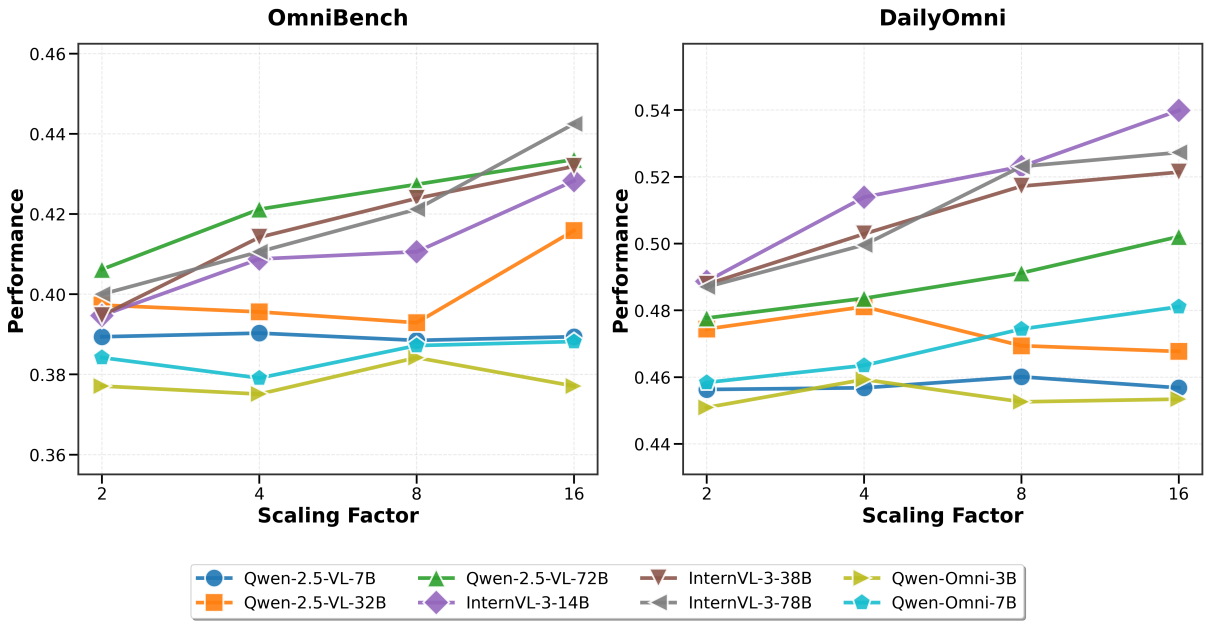

Figure 4: Best-of-N experiments on two downstream tasks: OmniBench (IID) and DailyOmni (OOD).

form direct evaluations, where models explicitly rate a single responses and we use this score to determine which solution is chosen. Additionally, within each evaluation group, we calculate the rank difference (*Rank Diff*) which quantifies the difference in a model's rank obtained from direct evaluation compared to its rank from the pairwise evaluation to measure how model rankings vary under different evaluation conditions, offering further insights into model stability across evaluation metrics.

4.1 Performance of Omni-Modal Models

Table 1 summarizes the performance of omni-modal models. Among these models, Gemini-2.5-Flash achieved the highest average pairwise accuracy (83.8%), substantially outperforming other omni-modal systems. Open-source models, such as Qwen-Omni-7B and MiniCPM-2.6-O, displayed moderate performance with average accuracy of around 74.0% and 71.0%, respectively, indicating room for improvement in current open-source omni-modal reward modeling capabilities. Notably, Gemini-2.5-Flash also exhibited a relatively higher sensitivity to response ordering (bias magnitude of 5.64), suggesting potential instability in preference prediction depending on the input sequence order.

4.2 Performance of Bi-Modal Models via Modality Bridging

We further examine model performance using our proposed modality bridging approach. Table 2 presents these results. Remarkably, GPT-4o combined with Qwen2-Audio-Instruct as the captioning model achieves superior performance, attaining an average accuracy of 82.5%. Moreover, InternVL-3 models (particularly InternVL-3-78B) consistently demonstrate strong results, achieving average accuracies around 79.1%. Smaller-sized VL models like InternVL-3-2B and Qwen-2.5-VL-3B show comparatively lower performance, highlighting clear scaling benefits of larger model sizes for robust omni-modal reasoning.

In the modality-bridging configuration employing VL captioning for AL models, models show moderate results, with Kimi-Audio achieving the highest accuracy (around 56.3%) when combined with QwenVL-2-2B captions, suggesting potential limitations or complexities in effectively distinguishing high-quality solutions using an audio-language model.

Additionally, when pairing the same policy models with stronger captioning models, such as replacing Qwen2-Audio-Instruct with the more powerful Gemini-2.5-Pro (See Appendix C.1), we observe consistent improvements in accuracy. This highlights the critical role of caption quality in the effectiveness of modality bridging.

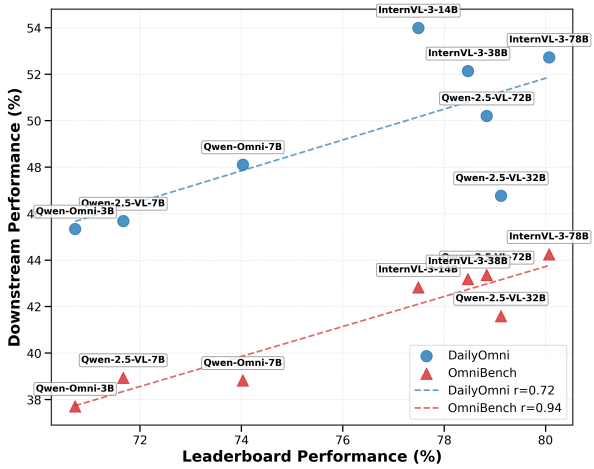

Figure 5: Correlation between leaderboard performance and downstream task performance (Best-of-16). Dashed lines represent linear regression fits.

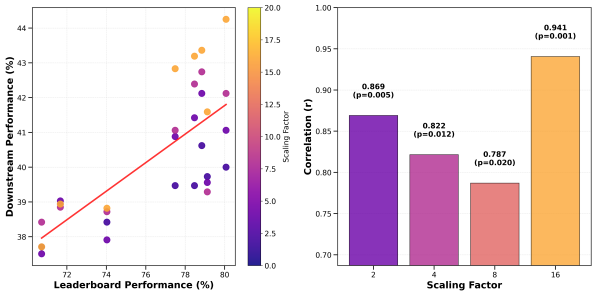

Figure 6: Correlation between leaderboard and downstream task performance (OmniBench) across different scaling factors (Best-of-N). Left: Scatter plot with regression line; color denotes scaling factor. Right: Pearson's correlation (r) under different scaling factors.

4.3 Position Bias & Rank Diff Analysis

Beyond overall model accuracy, the stability and robustness of model performance are also critical considerations. We quantify this sensitivity through the metric |Bias| and *Rank Diff*.

A smaller bias magnitude (|Bias|) indicates that the model's preference judgments are less influenced by response ordering, demonstrating higher decision consistency. For instance, among the omni-modal models, Gemini-2.5-Flash, despite achieving the highest overall performance (83.81% average accuracy), exhibited the largest position bias (5.64), indicating potential susceptibility to response ordering. In contrast, Qwen-Omni-7B maintained a relatively lower position bias (1.58), highlighting a more stable decision-making behavior across different input orders.

Furthermore, to comprehensively assess model consistency beyond pairwise comparisons, we calculate the Rank Difference (*Rank Diff*) between rankings obtained through direct evaluation and pairwise evaluation. Positive *Rank Diff* values indicate that models perform relatively better under direct evaluation, while negative values suggest the opposite. For example, in modality bridging group, InternVL-3-2B exhibited a notably large positive *Rank Diff* (+10 with Qwen2-Audio-Instruct and +12 with Gemini-2.5-Pro), demonstrating substantially improved relative performance under direct evaluation conditions. Conversely, larger vision-language models, such as Qwen-2.5-VL-32B, displays negative rank differences (-3 and -8 for Qwen2-Audio-Instruct and Gemini-2.5-Pro captioning, respectively), implying these models' direct evaluation accuracy are less competitive compared to their performance on pairwise scenarios.

In general, this analysis of position bias and rank differences provides critical insights into model reliability, robustness, and evaluation consistency, highlighting that achieving high overall accuracy does not necessarily imply robustness to positional effects or consistency across evaluation paradigms.

4.4 Scaling Performance with Best-of-N & Correlation Analysis

To validate the practical utility of reward models beyond static evaluation, we investigate their integration into inference-time scaling via the Best-of-N sampling paradigm. In this setting, a policy model generates multiple candidate responses, and a reward model selects the one with the highest score. This framework directly assesses how reward models enhance downstream task performance. In our experiments, we employ MiniCPM-2.6-O as the policy model and representative reward models from different *Omni-RewardBench* performance tiers, evaluating scaling factors $n \in \{2, 4, 8, 16\}$.

We conduct experiments on two evaluation sets: *OmniBench* (IID) and *DailyOmni* (OOD). As shown in Figure 4, inference-time scaling consistently improves performance across most reward models, with gains increasing alongside larger scaling factors. This indicates that sampling more candidates raises the likelihood of optimal generation, and that effective reward models are crucial for accurate selection.

To examine whether *Omni-RewardBench* leaderboard scores reliably predict downstream utility, we compute the Pearson correlation between leaderboard performance and actual task outcomes (Figure 5). Strong correlations are observed for both

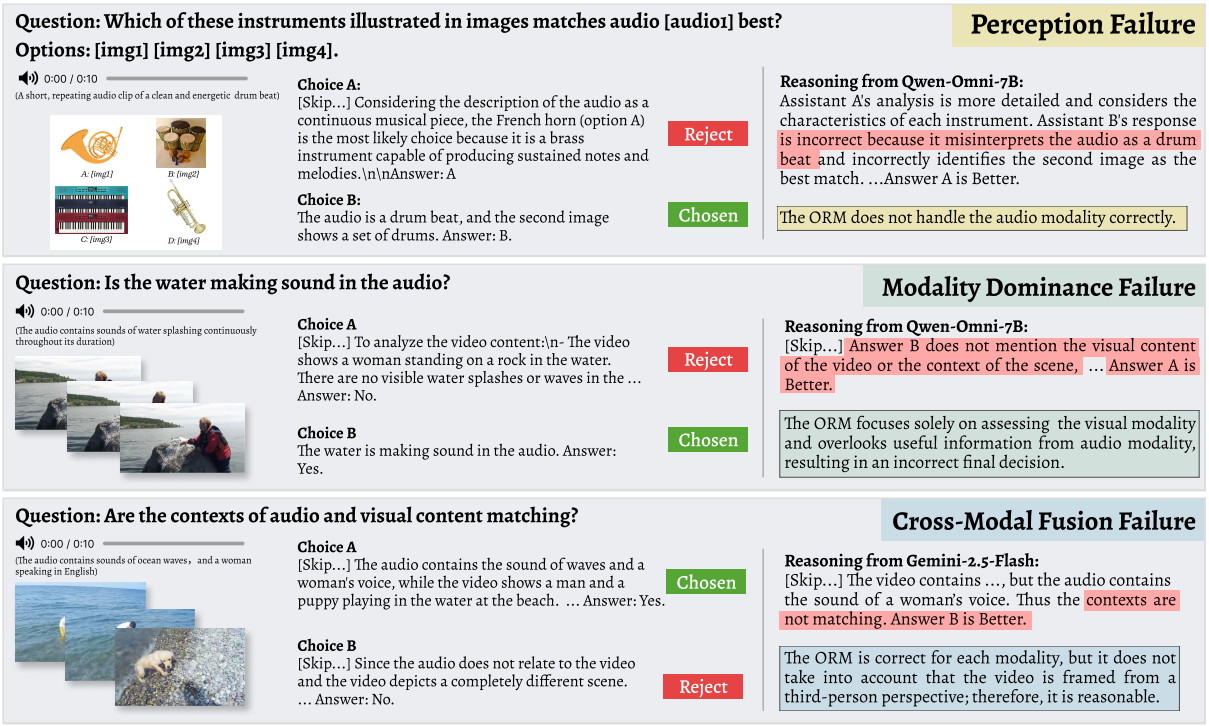

Figure 7: Failure mode of ORMs on *Omni-RewardBench*.

IID ($r = 0.94$) and OOD ($r = 0.72$) settings, confirming the benchmark's predictive reliability. Furthermore, Figure 6 shows that higher-scoring reward models exhibit superior discriminative capability and that this correlation remains stable across scaling factors. Overall, these findings validate the robustness, generalization, and practical utility of *Omni-RewardBench* as a reliable guide for inference-time scaling.

5 Case Study on Failure Mode of ORMs

To provide a more granular understanding of model weaknesses, we conducted a qualitative analysis of common failure modes that ORMs adjudicate. This investigation reveals several recurring error archetypes, as illustrated in Figure 7. Specifically, **Perception Failure** highlight fundamental failures in correctly interpreting sensory input, such as when a model mischaracterizes a continuous musical piece from a brass instrument as a "drum beat," leading to an incorrect preference judgment based on flawed perceptual grounding. A second category, **Modality Dominance Failure**, arises when a model assigns undue weight to information from one modality while neglecting salient, contradictory evidence from another, such as relying on a textual audio caption that asserts the absence of water sounds despite clear visual evidence of splashing. Finally, a more subtle but critical issue is **Cross-Modal Fusion Failure**, where the model may accurately perceive and describe the content of individual modalities but subsequently fails to logically synthesize this information, resulting in a final judgment that contradicts its own step-by-step analysis. By identifying and categorizing these distinct failure modes, our work provides crucial insights to the community and advocates for future optimization efforts to target these specific challenges in integrated perception and reasoning.

6 Conclusion

This work presents *Omni-RewardBench*, the first comprehensive benchmark for evaluating reward models across omni-modal contexts that integrate text, audio, and visual inputs. Experiments on over 20 state-of-the-art models highlight persistent limitations in omni-modal reasoning, while demonstrating the modality-bridging strategy as a practical interim solution. Strong correlations between *Omni-RewardBench* scores and downstream performance confirm its reliability as a predictor of real-world capability and a valuable foundation for advancing future omni-modality reward modeling research.

7 Ethical considerations

This study involved human annotators for the verification of automatically generated annotations. Annotators were recruited from a pool of qualified experts with backgrounds in multimodal reasoning and were compensated at fair rates consistent with academic research standards. The full annotation workflow, including detailed instructions and quality-control procedures, is described in Section 3.2. All annotators provided informed consent after being briefed on the task objectives, expected workload, and data usage policies. The dataset construction relied exclusively on publicly available or appropriately licensed resources, ensuring full compliance with licensing requirements. The data collection and human verification protocol were reviewed and approved by the institutional Ethics Review Board (ERB). No personally identifiable or sensitive data were collected during the study.

8 Limitations

Although *Omni-RewardBench* provides the first large-scale benchmark for evaluating omni-modality reward models (ORMs), it still faces several limitations.

First, the current benchmark scale, while diverse across eight categories and twenty-two subcategories, remains relatively small compared with existing unimodal or bimodal benchmarks. Expanding both the dataset size and coverage of real-world omni-modal interactions will further enhance statistical robustness and generalization.

Second, while this work introduces a comprehensive evaluation framework and detailed case studies that identify representative ORM failure patterns, it does not propose a training or optimization method to mitigate these issues. This choice is primarily due to the current lack of large-scale, high-quality omni-modal preference data required for effective reward model training, and because such algorithmic design lies beyond the intended scope of this benchmark-oriented study.

Finally, the annotator pool involved in the human verification stage is relatively narrow in terms of demographics. Our verification team primarily consisted of graduate researchers specializing in Computer Science who are highly proficient in omni-modal evaluation protocols. While this expertise ensures high technical rigor in assessing complex reasoning traces, it may inadvertently introduce a demographic or professional bias into the preference judgments. Consequently, expanding the annotator pool to include a more diverse, cross-disciplinary, and globally representative demographic remains an important direction for future work.

Acknowledgments

This work is funded in part by the HKUST Start-up Fund (R9911), Theme-based Research Scheme grant (T45-205/21-N), the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, and the research funding under HKUST-DXM AI for Finance Joint Laboratory (DXM25EG01).

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Anthropic. 2025. Claude4. https://www.anthropic.com.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

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

Chi-Min Chan, Chunpu Xu, Jiaming Ji, Zhen Ye, Pengcheng Wen, Chunyang Jiang, Yaodong Yang, Wei Xue, Sirui Han, and Yike Guo. 2025a. J1: Exploring simple test-time scaling for llm-as-a-judge. *arXiv preprint arXiv:2505.11875*.

Chi-Min Chan, Chunpu Xu, Junqi Zhu, Jiaming Ji, Donghai Hong, Pengcheng Wen, Chunyang Jiang, Zhen Ye, Yaodong Yang, Wei Xue, and 1 others. 2025b. Boosting policy and process reward models with monte carlo tree search in open-domain qa. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7433–7451.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, and 1 others. 2025. Emova: Empowering language models to see, hear and speak with vivid emotions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5455–5466.

Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, and 1 others. 2024a. Omnixr: Evaluating omni-modality language models on reasoning across modalities. *arXiv preprint arXiv:2410.12219*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

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*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025a. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025b. Mm-ifengine: Towards multimodal instruction following. *arXiv preprint arXiv:2504.07957*.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, and 1 others. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and 1 others. 2024. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1-2):116–134.

Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, and 1 others. 2024. Align anything: Training all-modality models to follow instructions with language feedback. *arXiv preprint arXiv:2412.15838*.

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, and 1 others. 2025. Wavreward: Spoken dialogue models with generalist reward evaluators. *arXiv preprint arXiv:2505.09558*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, and 1 others. 2025a. Vl-rewardbench: A challenging benchmark for vision-language generative reward models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24657–24668.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, and 1 others. 2025b. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, and 1 others. 2024c. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv e-prints*, pages arXiv–2502.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025a. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, and 1 others. 2025b. Openomni: Advancing open-source omnimodal large language models with progressive multimodal alignment and real-time self-aware emotional speech synthesis. *arXiv preprint arXiv:2501.04561*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Abhirama Subramanyam Penamakuri, Kiran Chhatre, and Akshat Jain. 2025. Audiopedia: Audio qa with knowledge. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, and 1 others. 2025. Judge anything: Mllm as a judge across any modality. *arXiv preprint arXiv:2503.17489*.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. 2024. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*.

Jiacheng Ruan, Wenzhen Yuan, Xian Gao, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Ting Liu, and Yuzhuo Fu. 2025. Vlrmbench: A comprehensive and challenging benchmark for vision-language reward models. *arXiv preprint arXiv:2503.07478*.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

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

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Haoqin Tu, Weitao Feng, Hardy Chen, Hui Liu, Xianfeng Tang, and Cihang Xie. 2025. Vilbench: A suite for vision-language process reward modeling. *arXiv preprint arXiv:2503.20271*.

Matthew Turk. 2014. Multimodal interaction: A review. *Pattern recognition letters*, 36:189–195.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024a. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

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. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024d. Helpsteer2-preference: Complementing ratings with preferences. *arXiv preprint arXiv:2410.01257*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv preprint arXiv:2408.01337*.

Xiaobao Wu. 2025. Sailing ai by the stars: A survey of learning from rewards in post-training and test-time scaling of large language models. *arXiv preprint arXiv:2505.02686*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Zhaopan Xu, Pengfei Zhou, Jiaxin Ai, Wangbo Zhao, Kai Wang, Xiaojiang Peng, Wenqi Shao, Hongxun Yao, and Kaipeng Zhang. 2025b. Mpbench: A comprehensive multimodal reasoning benchmark for process errors identification. *arXiv preprint arXiv:2503.12505*.

Cui Yakun, Peng Qi, Fushuo Huo, Hang Du, Weijie Shi, Juntao Dai, Zhenghao Zhu, Sirui Han, and Yike Guo. 2025. Perception, understanding and reasoning, a multimodal benchmark for video fake news detection. *arXiv preprint arXiv:2510.24816*.

Cui Yakun, Yanting Zhang, Zhu Lei, Jian Xie, Zhizhuo Kou, Hang Du, Zhenghao Zhu, and Sirui Han. 2026. Mmfctub: Multi-modal financial credit table understanding benchmark. *arXiv preprint arXiv:2601.04643*.

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*.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Shaolei Zhang, Shoutao Guo, Qingkai Fang, Yan Zhou, and Yang Feng. 2025. Stream-omni: Simultaneous multimodal interactions with large language-vision-speech model. *arXiv preprint arXiv:2506.13642*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yujin Zhou, Pengcheng Wen, Jiale Chen, Boqin Yin, Han Zhu, Jiaming Ji, Juntao Dai, Chi-Min Chan, and Sirui Han. 2026. What, whether and how? unveiling process reward models for thinking with images reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29071–29079.

A Models Evaluated in *Omni-RewardBench*

To provide a comprehensive evaluation of various models on *Omni-RewardBench*, we select a diverse set of models. As mentioned in Section 1, due to the lack of omni-models that natively support full multimodal processing, we also introduce modality bridging to evaluate bi-modal models.

Specifically, for omni-models, we evaluate both proprietary and open-source models, including Ola-7B (Liu et al., 2025), Video-LlamA2 (Cheng et al., 2024), Vita-1.5 (Fu et al., 2025), Uio2-Large (Lu et al., 2024), Gemini-2.5 Flash (Comanici et al., 2025), Qwen-Omni (Xu et al., 2025a), Baichuan-Omni (Li et al., 2025b), and MiniCPM-2.6-O (Yao et al., 2024). For modality-bridging evaluation, we assess vision-language models such as InternVL (Chen et al., 2024b), Qwen-2.5-VL (Bai et al., 2025), DeepSeek-VL (Wu et al., 2024), and GPT-4o (Hurst et al., 2024), while using Qwen2-Audio-Instruct (Chu et al., 2024) and Gemini-1.5 Pro (Team et al., 2023) as audio caption models. For audio-language models, we evaluate Qwen2-Audio-7B (Chu et al., 2024), Qwen-Audio-Chat (Chu et al., 2023), and Kimi-Audio (Ding et al., 2025a), similarly employing Qwen-VL-2-2B (Wang et al., 2024c) and GPT-4o (Hurst et al., 2024) as visual caption models. By default, we use temperature $= 0.0$ during evaluation. All experiments are run on a NVIDIA H800 cluster.

Figure 8: Two examples in *Omni-RewardBench*.

B Details of Curating *Omni-RewardBench*

B.1 Description of Source Dataset

Our benchmark primarily sources the dataset from the following benchmark. We introduce the details of each dataset as follows.

OmniBench (Li et al., 2024c) is a benchmark designed to evaluate OLMs capable of reasoning and solving across visual, audio, and textual inputs simultaneously. It addresses a critical gap in existing benchmarks by requiring integrated understanding of all three modalities through 1,142 rigorously annotated question-answer pairs. These samples span diverse task types, from low-level perception (e.g., object identification) to high-level reasoning (e.g., plot inference), and include three audio categories: speech, sound events, and music.

AV-Odyssey (Gong et al., 2024) is a comprehensive dataset of 4,555 carefully curated multimodal questions (combining text, images/videos, and audio) spanning 26 tasks across diverse domains (e.g., music, daily life). The benchmark employs a multiple-choice format to ensure objective evaluation, eliminating reliance on human or LLM-assisted scoring. Key findings reveal that even state-of-the-art models perform near-random accuracy in integrating audio-visual cues, with errors rooted in flawed audio perception rather than reasoning complexity.

AVHBench (Sung-Bin et al., 2024) is a comprehensive benchmark designed to evaluate cross-modal hallucinations in audio-visual LLMs, addressing a critical gap in multimodal AI research. It introduces 5,816 question-answer pairs and 1,238 audio-visual captions across four tasks (Audio-driven Video Hallucination, Video-driven Audio Hallucination, Audio-visual Matching, and Audio-visual Captioning) to systematically assess models' ability to integrate and reason about combined audio-visual inputs without fabricating non-existent sensory information (e.g., "hearing" sounds in silent videos).

WorldSense (Hong et al., 2025) features 1,662 synchronized audio-visual videos spanning 8 domains (e.g., driving, music) and 3,172 high-quality multiple-choice QA pairs requiring synergistic perception of all modalities for correct answers. The benchmark emphasizes omni-modality collaboration, where models must jointly analyze audio-visual-textual cues to solve tasks ranging from basic recognition to complex reasoning (e.g., inferring causal relationships).

B.2 Full Taxonomy

Table 4 presents the detailed data composition of Omni-RewardBench, which comprises 1,395 samples across five major categories. The dataset is dominated by multimodal understanding tasks (603

Model	Model Size	Vision Encoder	Audio Encoder	Base LLM
GPT-4o	-	-	-	-
Gemini-2.5-flash	-	-	-	-
Qwen-Omni-3B	3B	ViT	Whisper-large-v3	Qwen2.5 LLM
Qwen-Omni-7B	7B	ViT	Whisper-large-v3	Qwen2.5 LLM
Baichuan-Omni	7B	NaViT	Whisper-Large	Qwen2.5 LLM
MiniCPM-2.6-O	7B	SigLip-400M	Whisper-medium-300M	Qwen2.5 LLM
Ola-7B	7B	OryxViT	Whisper-large-v3	Qwen2.5 LLM
Video-Llama2	7B	ViT-L/14	BEATs	Qwen2 LLM
Vita-1.5	7B	InternViT-300M-448px	Transformer-based Encoder	Qwen2 LLM
Uio2-Large	1.1B	ViT	AST	BPE
Qwen2.5-VL-3B-Instruct	3B	ViT-bigG	-	Qwen2.5 LLM
Qwen2.5-VL-7B-Instruct	7B	ViT-bigG	-	Qwen2.5 LLM
Qwen2.5-VL-32B-Instruct	32B	ViT-H/14	-	Qwen2.5 LLM
Qwen2.5-VL-72B-Instruct	72B	ViT-H/14	-	Qwen2.5 LLM
DeepSeek-VL2	27B	SigLIP-SO400M	-	DeepSeekMOE LLM
DeepSeek-VL2-Small	7B	SigLIP-SO400M	-	DeepSeekMOE LLM
DeepSeek-VL2-Tiny	3B	SigLIP-SO400M	-	DeepSeekMOE LLM
InternVL3-2B	2B	InternViT-300M-448px-V2_5	-	Qwen2.5-1.5B
InternVL3-8B	8B	InternViT-300M-448px-V2_5	-	Qwen2.5-7B
InternVL3-14B	14B	InternViT-300M-448px-V2_5	-	Qwen2.5-14B
InternVL3-38B	38B	InternViT-6B-448px-V2_5	-	Qwen2.5-32B
InternVL3-78B	78B	InternViT-6B-448px-V2_5	-	Qwen2.5-72B
Qwen2-Audio-7B	7B	-	Whisper-large-v3	Qwen2 LLM
Qwen-Audio-Chat	8B	-	Whisper-large-v2	Qwen LLM
Kimi-Audio	7B	-	Whisper	Transformer Decoder

Table 3: Evaluated models in *Omni-RewardBench*.

samples, 43.2%), with Cross-Modal Hallucination being the largest sub-category at 350 samples, followed by Audio-Visual Correspondence with 209 samples. Content Analysis represents the second largest category with 326 samples (23.4%), encompassing diverse analytical tasks ranging from narrative analysis to spatial analysis. The Arts & Culture category contributes 349 samples (25.0%), with Audio Arts being particularly prominent at 321 samples. Entertainment & Media and Others categories provide smaller but important contributions, ensuring comprehensive coverage across different domains. Figure 9 illustrates the distribution of samples across source datasets, showing that AVHBench contributes the largest portion at 557 samples (40.51%), followed by AV-Odyssey with 417 samples (30.33%) and OmniBench with 281 samples (20.43%). WorldSense provides the

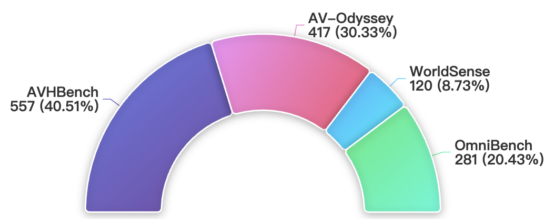

Figure 9: Distribution of samples across source datasets.

remaining 120 samples (8.73%), demonstrating a well-balanced integration from multiple established benchmarks to create a comprehensive evaluation framework.

B.3 Statistics of *Omni-RewardBench*

This figure 10 presents three key statistical analyses of *Omni-RewardBench* that validate the dataset's quality and balance. The Distribution of Question Lengths (left panel) demonstrates the diversity of questions in the benchmark, with lengths ranging from short queries around 10 tokens to more complex questions extending beyond 80 tokens, ensuring comprehensive evaluation across various question complexities. The Distribution of Response Lengths (middle panel) reveals that chosen and rejected responses exhibit similar length distributions, with both categories spanning from concise answers around 25 tokens to detailed responses exceeding 300 tokens. Importantly, the overlapping distributions with nearly identical kernel density estimation (KDE) curves indicate that response selection is not biased toward length, ensuring that quality judgments are based on content rather than verbosity. The Distribution of Scores (right panel) provides compelling evidence of the dataset's discriminative power, showing a clear separation be-

Category	Sub-Category	Number
Multimodal Understanding	Cross-Modal Hallucination	350
	Audio-Visual Correspondence	209
	Multimodal Description	14
	Complexity Analysis	30
Content Analysis	Spatial Analysis	5
	Temporal Analysis	45
	Action Recognition	71
	Narrative Analysis	114
	Object Recognition	52
	Scene Understanding	32
	Text Recognition	3
	Quantitative Analysis	4
Arts & Culture	Performance Arts	13
	Audio Arts	321
	Visual Arts	15
Entertainment & Media	Music	15
	Film & TV	12
Others	Sports	14
	Gaming	10
	Home & Living	24
	Biology & Medicine	17
	Human Interaction	5
Total		1,395

Table 4: Data Statistics of *Omni-RewardBench*

tween chosen and rejected responses. Chosen responses cluster around higher scores (4-5 range) with a sharp, concentrated distribution, while rejected responses predominantly fall in the lower score range (1-2), with minimal overlap between the two distributions. This distinct pattern confirms that the chosen responses are genuinely of higher quality than their rejected counterparts, validating the effectiveness of the annotation process and the benchmark's ability to distinguish between different response qualities.

B.4 Existing Assets Licenses

Omni-RewardBench is released under the **CC BY-NC 4.0** License.

The incorporated benchmarks operate under different licensing terms: OmniBench (Li et al., 2024c) is distributed under the *CC BY-NC 4.0* License. AV-Odyssey (Gong et al., 2024) uses a proprietary license and requires contacting the authors for usage permissions. WorldSense (Hong et al., 2025) is available under the *CC BY 4.0* License. AVHBench (Sung-Bin et al., 2024) does not specify explicit licensing terms. Users should ensure compliance with the applicable license terms for their intended use case.

C More Experimental Results

C.1 Ablation on Using Different Caption Model

To further examine the impact of caption quality on the modality-bridging framework, we conduct an ablation study by replacing the original caption model (Qwen2-Audio-Instruct) with a stronger model (Gemini-2.5-Pro). As shown in Table 9, this substitution leads to consistent improvements in both pairwise and direct accuracy across VL and AL models. These results underscore the critical role of caption quality in bridging heterogeneous modalities effectively. High-quality captions provide more faithful and semantically rich intermediate representations, allowing the policy model to better align multimodal information and mitigate information loss during modality transfer.

C.2 Modality Absence Ablation Study

To validate the necessity of omni-modal capabilities in our benchmark, we conduct an ablation study by systematically removing individual modalities from the models and evaluating their performance on vision-only and audio-only tasks, as shown in Table 10.

Vision-Only Performance Degradation: When models are restricted to visual inputs only (audio modality disabled), we observe significant performance drops across all model categories. For instance, Qwen-Omni-7B experiences a decrease from 70.72% to 69.42% in pairwise accuracy, while Gemini-2.5-Flash shows a more substantial drop from 83.18% to 80.69%.

Audio-Only Performance Analysis: The audio-only evaluation reveals even more pronounced limitations. Omni models, which are designed to handle multiple modalities, show substantial performance drops when restricted to audio inputs alone. Qwen-Omni-3B's pairwise accuracy falls from 70.72% to 60.13%, and Qwen-Omni-7B drops from 74.03% to 66.33%. Notably, vision-language models cannot process audio-only inputs, just as audio-language models are incompatible with vision-only scenarios, so the corresponding results are not shown in the table.

Counterintuitive Bias Reduction Under Modal Constraints: Interestingly, we observe that position bias often *decreases* when modalities are removed. For example, Qwen-Omni-3B shows reduced bias in vision-only mode (2.84% vs. 4.00% in the full omni-modal setting), and Gemini-2.5-

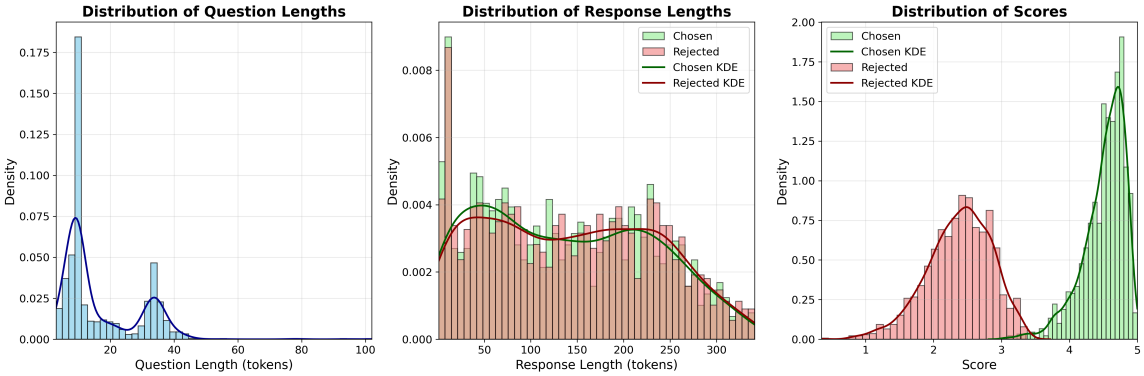

Figure 10: Statistics of *Omni-RewardBench*.

```
Please answer the following question based on the provided image and audio.

Question:{question}
Options:{options}

Please provide your step-by-step reasoning and answer with the option's letter
    from the given choices.
```

Table 5: Prompt for response generation.

Flash exhibits dramatically lower bias (1.08% vs. 5.64%). This counterintuitive pattern suggests that while multimodal information enhances overall performance, it may also introduce complexities in the evaluation process that can manifest as increased positional sensitivity.

This bias pattern reveals a fascinating paradox: models become more positionally sensitive when provided with richer multimodal information, yet their overall performance improves substantially. This suggests that multimodal processing introduces sophisticated reasoning processes that, while beneficial for task performance, create more nuanced dependencies on presentation order and context. The reduced bias under modal constraints may reflect simpler, more deterministic decision-making processes when information is limited.

Direct Evaluation Consistency Patterns: The direct evaluation scores show varied patterns under modal constraints, with some models maintaining consistency while others show degradation. This heterogeneous response indicates that different architectures handle modal information integration differently, with some being more robust to information loss than others.

Multimodal Necessity Validation: These results provide compelling evidence for the importance of our omni-modal benchmark design. The consistent performance degradation when any single modality is removed demonstrates that optimal performance requires the integration of both textual, audio and visual information.

To sum up, the ablation study conclusively shows that no single modality is sufficient for achieving peak performance, thereby justifying the omni-modal nature of our benchmark and highlighting the critical importance of developing truly omni-modal AI systems.

```
Please act as an impartial judge and evaluate the quality of the responses
    provided by two AI assistants to the user question displayed below. You
    should choose the assistant that follows the user's instructions and
    answers the user's question better. Begin your evaluation by comparing the
     two responses and provide a short explanation. Avoid any position biases
    and ensure that the order in which the responses were presented does not
    influence your decision. Do not allow the length of the responses to
    influence your evaluation. Do not favor certain names of the assistants.
    Be as objective as possible. Please first analysis both of the answer step
     by step, directly point out the position of error and output why it is an
     error in detail when finding error in analysis. If the question is open-
    ended, directly point out why the rejected answer is worse than the chosen
     one. After providing your explanation, output your final verdict by
    strictly following this format: '[[A]]' if assistant A is better, '[[B]]'
    if assistant B is better.

[User Question]
{question}
Here is the audio content: {caption} / Here is the visual content: {caption}{
{options_text}
{{The Start of Assistant A's Answer}}
{first_answer}
{{The End of Assistant A's Answer}}
{{The Start of Assistant B's Answer}}
{second_answer}
{{The End of Assistant B's Answer}}
```

Table 6: Prompt for pair-wise evaluation.

```
Please evaluate the following answer to the question based on the provided omni
    information and ground truth answer.

Question: {question}

Here is the audio content: {audio content} / Here is the visual content: {visual
    content}

Ground Truth Answer: {ground_truth}

Answer: {answer}

Evaluation Process:
1. First, compare the answer with the ground truth answer. Identify:
   - Key similarities and alignments
   - Important differences or omissions
   - Any incorrect information in the answer

2. Then evaluate the answer on the following criteria (score each on a scale of
    1-5):

   Integration: Effectiveness in combining information from all input modalities
   - 1: Ignores multiple modal inputs
   - 2: Acknowledges but poorly integrates different modalities
   - 3: Adequately combines information across modalities
   - 4: Effectively synthesizes information from all modalities
   - 5: Masterfully weaves together all modal information with perfect balance

   Relevance: Directness in addressing the question
   - 1: Significantly deviates from the question
   - 2: Tangentially related to the question
   - 3: Addresses the question but includes irrelevant content
   - 4: Directly addresses the question with minimal deviation
   - 5: Precisely focused on all aspects of the question

   Completeness: Coverage of all question aspects considering multimodal inputs
   - 1: Misses most critical elements
   - 2: Covers only basic elements while missing important details
   - 3: Addresses main points with some gaps
   - 4: Covers most aspects thoroughly
   - 5: Comprehensively addresses all aspects of the question

   Clarity: Structure and understandability of the answer
   - 1: Difficult to follow or understand
   - 2: Somewhat unclear with confusing elements
   - 3: Generally clear but could be better organized
   - 4: Well-structured and easy to understand
   - 5: Exceptionally clear, concise, and well-organized

   Consistency: Logical coherence and absence of contradictions
   - 1: Contains multiple contradictions
   - 2: Has noticeable inconsistencies
   - 3: Generally consistent with minor issues
   - 4: Maintains consistency throughout with rare exceptions
   - 5: Perfectly consistent across all points

3. Provide an overall score (1-5) that considers:
   - How well the answer matches the ground truth
   - Performance across all five evaluation criteria
   - Overall quality and usefulness of the response

Please format your response as follows:
Integration: [score]
Relevance: [score]
Completeness: [score]
Clarity: [score]
Consistency: [score]
Overall Score: [score]
```

Table 7: Prompt for preference annotation.

```
Please evaluate the following answer to the question based on the provided
    multimodal information.

Question: {question}

Here is the audio content: {audio content} / Here is the visual content: {visual
    content}

Answer: {answer}

Evaluation Process:
1. First, analyze the answer in relation to the question. Consider:
   - Accuracy and correctness of the information provided
   - Completeness of the response
   - Relevance to the specific question asked

2. Then evaluate the answer on the following criteria (score each on a scale of
    1-5):

   Integration: Effectiveness in combining information from all input modalities
   - 1: Ignores multiple modal inputs
   - 2: Acknowledges but poorly integrates different modalities
   - 3: Adequately combines information across modalities
   - 4: Effectively synthesizes information from all modalities
   - 5: Masterfully weaves together all modal information with perfect balance

   Relevance: Directness in addressing the question
   - 1: Significantly deviates from the question
   - 2: Tangentially related to the question
   - 3: Addresses the question but includes irrelevant content
   - 4: Directly addresses the question with minimal deviation
   - 5: Precisely focused on all aspects of the question

   Completeness: Coverage of all question aspects considering multimodal inputs
   - 1: Misses most critical elements
   - 2: Covers only basic elements while missing important details
   - 3: Addresses main points with some gaps
   - 4: Covers most aspects thoroughly
   - 5: Comprehensively addresses all aspects of the question

   Clarity: Structure and understandability of the answer
   - 1: Difficult to follow or understand
   - 2: Somewhat unclear with confusing elements
   - 3: Generally clear but could be better organized
   - 4: Well-structured and easy to understand
   - 5: Exceptionally clear, concise, and well-organized

   Consistency: Logical coherence and absence of contradictions
   - 1: Contains multiple contradictions
   - 2: Has noticeable inconsistencies
   - 3: Generally consistent with minor issues
   - 4: Maintains consistency throughout with rare exceptions
   - 5: Perfectly consistent across all points

3. Provide an overall score (1-5) that considers:
   - Performance across all five evaluation criteria
   - Overall quality and usefulness of the response
   - How well the answer addresses the question

Please must format your response as follows:
Integration: [score]
Relevance: [score]
Completeness: [score]
Clarity: [score]
Consistency: [score]
Overall Score: [score]

Please must give your overall score in the end.
```

Table 8: Prompt for direct evaluation.

Model Configuration		Pairwise Accuracy (%)				Direct Accuracy	
Policy Model	**Caption Model**	**Original**	**Shuffle**	**Avg.**	**|Bias|**	**Score (%)**	**Rank Diff**
Modality Bridging (AL Caption for VL Model)							
InternVL-3-2B	Gemini-2.5-Pro	45.23	48.58	46.91	7.14	68.65	+12
InternVL-3-8B	Gemini-2.5-Pro	77.38	78.47	77.92	1.40	74.03	+5
InternVL-3-14B	Gemini-2.5-Pro	77.38	77.60	77.49	0.28	65.38	-6
InternVL-3-38B	Gemini-2.5-Pro	78.11	78.83	78.47	0.92	71.05	-1
InternVL-3-78B	Gemini-2.5-Pro	80.44	79.71	80.07	0.91	73.96	-2
Qwen-2.5-VL-3B	Gemini-2.5-Pro	64.00	63.71	63.86	0.45	35.27	-2
Qwen-2.5-VL-7B	Gemini-2.5-Pro	70.47	72.87	71.67	3.35	66.40	+3
Qwen-2.5-VL-32B	Gemini-2.5-Pro	79.27	78.98	79.12	0.37	68.51	-8
Qwen-2.5-VL-72B	Gemini-2.5-Pro	80.15	77.52	78.84	3.34	71.56	-2
DeepSeek-VL2	Gemini-2.5-Pro	74.47	76.51	75.49	2.70	60.07	-5
DeepSeek-VL2-Small	Gemini-2.5-Pro	72.72	73.30	73.01	0.79	50.91	-5
DeepSeek-VL2-Tiny	Gemini-2.5-Pro	38.03	40.15	39.09	5.42	0.65	0
GPT-4o	Gemini-2.5-Pro	83.13	83.05	83.09	0.10	75.34	-2
Modality Bridging (VL Caption for AL Model)							
Qwen2-Audio-7B	GPT-4o	66.11	68.00	67.06	2.82	25.24	-3
Qwen-Audio-Chat	GPT-4o	62.25	60.22	61.23	3.32	25.09	-1
Kimi-Audio	GPT-4o	57.09	56.65	56.87	0.77	73.74	+4

Table 9: Ablation results of using different caption model on the performance of bi-modal models.

Model	Vision-Only Performance					Audio-Only Performance				
	Pairwise Accuracy (%)			|Bias|	Direct	Pairwise Accuracy (%)			|Bias|	Direct
	Original	Shuffled	Avg.	(%)	(%)	Original	Shuffled	Avg.	(%)	(%)
Omni Models (Both Video and Audio Capabilities)										
Qwen-Omni-3B	68.43	70.40	69.42	2.84	42.82	60.87	60.13	60.50	1.22	41.71
Qwen-Omni-7B	71.70	71.41	71.56	0.41	50.39	65.23	67.42	66.33	3.30	50.83
Baichuan-Omni	66.90	67.63	67.27	1.09	51.31	67.63	65.89	66.76	2.61	50.79
MiniCPM-2.6-O	70.03	69.74	69.88	0.41	59.36	65.16	69.16	67.16	5.96	56.72
Gemini-2.5-Flash	81.13	80.26	80.69	1.08	70.12	78.75	78.44	78.59	0.39	69.41
Vision-Language Models										
InternVL-3-2B	56.80	51.12	53.96	10.53	65.09	–	–	–	–	–
InternVL-3-8B	65.38	60.43	62.91	7.87	73.81	–	–	–	–	–
InternVL-3-14B	73.23	74.61	73.92	1.87	70.24	–	–	–	–	–
InternVL-3-38B	73.14	75.16	74.15	2.72	70.58	–	–	–	–	–
InternVL-3-78B	74.45	78.25	76.35	4.98	73.22	–	–	–	–	–
Qwen-2.5-VL-3B	55.37	54.89	55.13	0.87	32.36	–	–	–	–	–
Qwen-2.5-VL-7B	57.81	60.21	59.01	4.07	46.18	–	–	–	–	–
Qwen-2.5-VL-32B	66.03	69.16	67.59	4.63	61.74	–	–	–	–	–
Qwen-2.5-VL-72B	74.32	74.33	74.32	0.01	67.85	–	–	–	–	–
DeepSeek-VL2	68.23	65.89	67.06	3.49	63.56	–	–	–	–	–
DeepSeek-VL2-Small	67.78	65.05	66.41	4.11	57.60	–	–	–	–	–
DeepSeek-VL2-Tiny	39.01	41.52	40.27	6.23	0.13	–	–	–	–	–
GPT-4o	79.16	79.49	79.32	0.42	72.33	–	–	–	–	–
Audio-Language Models										
Qwen2-Audio-7B	–	–	–	–	–	54.61	54.25	54.43	0.66	24.94
Qwen-Audio-Chat	–	–	–	–	–	61.04	61.09	61.06	0.08	20.17
Kimi-Audio	–	–	–	–	–	50.34	49.23	49.78	2.23	57.18

Table 10: Ablation Study; **Vision-Only:** The model receives only visual inputs (audio modality is disabled).**Audio-Only:** The model processes solely audio inputs (visual modality is disabled).