

---

# BMMR: A Large-Scale Bilingual Multimodal Multi-Discipline Reasoning Dataset

---

Zhiheng Xi<sup>1\*†</sup> Guanyu Li<sup>1\*</sup> Yutao Fan<sup>2,3\*</sup> Honglin Guo<sup>1\*</sup>  
Yufang Liu<sup>4</sup> Xiaoran Fan<sup>1</sup> Jiaqi Liu<sup>1</sup> Jingchao Ding<sup>7</sup> Wangmeng Zuo<sup>3</sup>  
Zhenfei Yin<sup>5,6†</sup> Lei Bai<sup>2</sup> Tao Ji<sup>1</sup> Tao Gui<sup>1,8†</sup> Qi Zhang<sup>1</sup> Xuanjing Huang<sup>1</sup>

<sup>1</sup>Fudan University <sup>2</sup>Shanghai AI Laboratory <sup>3</sup>Harbin Institute of Technology  
<sup>4</sup>East China Normal University <sup>5</sup>Oxford <sup>6</sup>University of Sydney  
<sup>7</sup>Yimudata <sup>8</sup>Pengcheng Laboratory

## Abstract

In this paper, we introduce BMMR, a large-scale bilingual, multimodal, multi-disciplinary reasoning dataset for the community to develop and evaluate large multimodal models (LMMs). BMMR comprises 110k college-level questions spanning 300 UNESCO-defined subjects, spanning diverse formats—multiple-choice, fill-in-the-blank, and open-ended QA—and sourced from both print and digital media such as books, exams, and quizzes. All data are curated and filtered via a human-in-the-loop and scalable framework, and each instance is paired with a high-quality reasoning path. The dataset is organized into two parts: BMMR-Eval that comprises 20,458 high-quality instances to comprehensively assess LMMs’ knowledge and reasoning across multiple disciplines in both Chinese and English; and BMMR-Train that contains 88,991 instances to support further research and development, extending the current focus on mathematical reasoning to diverse disciplines and domains. In addition, we propose the process-based multi-discipline verifier (i.e., BMMR-Verifier) for accurate and fine-grained evaluation of reasoning paths. Extensive experiments on 24 models reveal that (i) even SOTA models (e.g., o3 and Gemini-2.5-Pro) leave substantial headroom on BMMR-Eval; (ii) reasoning models exhibit discipline bias and outperform LMMs only on specific subjects; (iii) open-source models still trail their proprietary counterparts; and (iv) fine-tuning on BMMR-Train narrows this gap. Additionally, we conduct reasoning-chain analyses using BMMR-Verifier and other in-depth studies, uncovering the challenges LMMs currently face in multidisciplinary reasoning. We will release the data, and we hope our work can offers insights and contributions to the community.

Project Site: <https://bmmr.pages.dev/>  
Code & Sources: <https://github.com/WooooDyy/BMMR/>

## 1 Introduction

Large multimodal models (LMMs) [1–3] and large reasoning models (LRMs) [4] have demonstrated extraordinary expertise and reasoning capabilities across a wide range of academic fields—such as mathematics, physics, and chemistry [5–7]. These models, represented by GPT-4o [8] and OpenAI-o1 [9], can process and reason over both textual and visual inputs, and have generated

---

\* Equal Contribution.

† Correspondence to: zhxi22@m.fudan.edu.cn, tgui@fudan.edu.cn, zhenfei.yin@sydney.edu.au

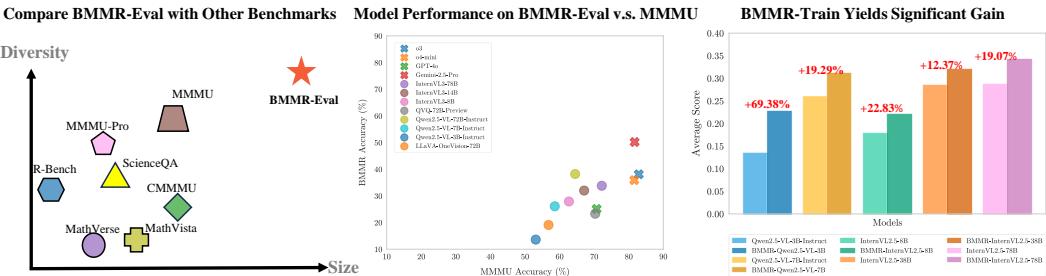
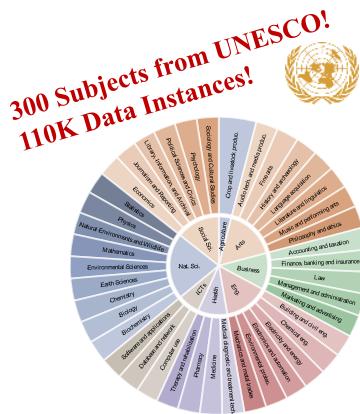


Figure 1: Overview of the BMMR dataset. It encompasses 110k instances across 300 subjects defined by UNESCO. We present two illustrative examples for visualization (top-middle and top-right). Furthermore, we compare our BMMR-Eval with other benchmarks regarding size and diversity (bottom-left). A comparison of model performance on BMMR-Eval versus MMMU is also included (bottom-middle), highlighting the challenging nature of our test set. Finally, we demonstrate that fine-tuning open-source models of various sizes (3B-78B) on our BMMR-Train yields significant performance enhancements (bottom-right).

significant interest in the AI community due to their potential to enable more general AI systems, i.e., AGI [10, 11].

However, with these advancements, comprehensively and accurately evaluating knowledge and reasoning capabilities of LMMs and LRMs across disciplines has become increasingly challenging. Existing benchmarks [12–14] struggle to strike a balance among subject diversity, problem complexity, reasoning depth, and language coverage, and have recently begun to exhibit performance saturation [15–18]. At the same time, the community lacks a multimodal, multidisciplinary training dataset—one that offers diverse questions and curated reasoning paths—to support research and development, especially within the open-source community [6, 19].

To bridge this gap, we introduce BMMR (Section 3): a large-scale bilingual, multimodal, multi-disciplinary reasoning dataset that contains 110k college-level high-quality instances, spanning 8 high-level disciplines and 300 sub-fields from UNESCO (The United Nations Educational, Scientific and Cultural Organization) [20], as illustrated in Figure 1. BMMR is organized into two parts: (1) BMMR-Eval, which comprises 20k instances with broad subject coverage and multiple difficulty levels for comprehensively assessing models' knowledge and reasoning across disciplines in both English and Chinese (see Table 1); and (2) BMMR-Train, which contains 89k instances to support further research and development, and extend the community's focus on mathematical reasoning to more diverse disciplines and domains.

We collect BMMR data from both digital and print sources—including books, exams, and quiz collections—and the dataset encompasses diverse formats such as multiple-choice, fill-in-the-blank, and open-ended QA. All instances are curated and filtered through a human-in-the-loop and scalable processing framework and paired with a high-quality reasoning path to ensure robustness and solidity. Every retained question in BMMR demands precise cross-modal comprehension, specialized domain knowledge, and advanced reasoning skills to solve [21–23].

To further enable accurate and fine-grained evaluation of models’ reasoning abilities across disciplines and to prevent models from simply recalling or guessing the correct answers [24–26], we also propose **BMMR-Verifier**—a process-based bilingual, multimodal, multidisciplinary verifier (Section 4).

Extensive experiments on 24 LMMs and LRM<sub>s</sub> (Section 5.2 and Section 5.3) reveal that: (1) Even SOTA models perform suboptimally—for instance, o3 and Gemini2.5-Pro only achieves 38.06 and 50.15, revealing substantial headroom; (2) Contrary to intuition, LRM<sub>s</sub> do not consistently outperform LMMs across all disciplines. Instead, they exhibit clear subject bias, excelling only in specific areas such as mathematical reasoning. This further validates BMMR’s emphasis on multi-discipline knowledge; (3) Open-source models still lag behind their proprietary counterparts, highlighting the academia-industry gap. (4) Fine-tuning on BMMR-Train narrows this gap—for example, the finetuned BMMR-InternVL2.5-78B achieves a 19.07% improvement in overall performance.

Additionally, using the developed BMMR-Verifier, we conduct a fine-grained analysis of reasoning processes (Section 5.4). We present the distribution of reasoning-step quality across different models and examine, at a granular level, their reasoning abilities in various disciplines. Furthermore, through error categorization, qualitative studies, and deeper analyses (Section 6), we highlight key challenges in multimodal reasoning—such as overthinking [27, 28] and hallucination [29, 30]—and hope these findings offer valuable insights for advancing the next-generation models.

In summary, our main contributions are:

1. We introduce BMMR, a large-scale bilingual, multimodal, multidisciplinary reasoning dataset—comprising BMMR-Eval and BMMR-Train—to enable comprehensive evaluation and support research and development of multimodal foundation models.
2. We propose the multimodal, multidisciplinary, process-based **BMMR-Verifier** for accurate and fine-grained evaluation of the models’ reasoning capabilities.
3. We conduct extensive experiments and analysis on 24 open-source and proprietary LMMs and LRM<sub>s</sub>, and provide key findings and insights. We hope our work can contribute to the field and inspire future research.

## 2 Related Work

**Benchmarks for LMMs.** The evaluation of multimodal models’ intelligence remains a critical endeavor [31]. While fundamental benchmarks have been introduced to evaluate core visual understanding skills of LMMs, including visual classification [32], retrieval [33], grounding [34], and question-answering [35], they do not specifically focus on reasoning capabilities in multidisciplinary tasks. MMMU [5] notably pioneered multi-discipline understanding evaluation with its 11k problems spanning 30 subjects. However, such traditional multi-discipline benchmarks demonstrate insufficient logic reasoning demands, failing to challenge contemporary state-of-the-art LMMs such as Gemini 2.5 [36] and InternVL3 [2]. Recent research has shifted toward evaluating System-2 reasoning through advanced benchmarks requiring a significantly higher cognitive standard: MathVista [37] employs both multiple-choice and open-ended formats to probe mathematical reasoning, while MathVerse [38] systematically investigates modality-specific performance variations to isolate visual understanding impacts. Although these emerging benchmarks pose significant challenges for current LMMs [12, 39, 40], they still exhibit critical limitations in providing holistic assessments of reasoning abilities across multiple disciplines. In this work, we build the larger-scale BMMR-Eval that covers more diverse subjects (see Table 1).

**Multimodal reasoning datasets.** To advance the reasoning capabilities of LMMs, researchers have developed specialized multimodal training datasets [6, 43]. Current efforts include datasets targeting foundational visual reasoning tasks such as commonsense reasoning, embodied planning [44], and spatial reasoning [45–47]. For complex reasoning challenges, studies like LLaVA-CoT [48] and MAMmoTH-VL [49] generate structured reasoning paths across diverse visual reasoning domains, while ScienceQA [6] and MM-Eureka [50] offer multidisciplinary question-answer datasets with detailed chain-of-thought annotations. However, these resources remain constrained by their exclusive focus on K-12-level content, which limits their effectiveness in advancing state-of-the-art models that require higher-order reasoning. In this work, we address these limitations by constructing a new college-level multimodal dataset featuring cross-modal comprehension, specialized domain knowledge and advanced reasoning.

Table 1: Overall comparison between BMMR-Eval and other existing benchmarks. In the Source column, D means digital-based data sources, such as websites and existing datasets; P means print-based data sources, such as college textbooks and exams; R means repurposed data sources. The column Multiple Images implies the presence of questions that contains multiple images. In the Question Type column, MC means multiple-choice questions, FIB means fill-in-the-blank questions, ans OE means open-ended questions, TF means true-or-false questions. (t) in the Language column means “translated”. In the Difficulty column, C means college level, K means K-12 level, and H means high-school level. Information for R-Bench only cover its multimodal subset. For all datasets, we only report statistics on their test split.

	Source	#Item	#Discipline	Multiple Images	Reasoning Path	Question Type	Language	Difficulty
MMMU [5]	D, P	10.5k	6/30/183	Yes	Partial	MC, OE	EN.	C
MMMU-Pro [35]	D, P	1.7k	6/30/183	Yes	Partial	MC, OE	EN	C
CMMMU [41]	D, P	11k	6/30	Yes	No	MC, FIB, TF	ZH	C
MathVista [37]	D	6.1k	Math	No	Partial	MC, OE	EN	K, C
MathVerse [38]	D	3.9k	Math	No	Partial	MC, OE	EN	H, C
ScienceQA [6]	P	4.2k	3/26/127	No	Yes	MC	EN	K
R-Bench [42]	P	665	83	No	No	MC, TF	EN, ZH (t)	C
<b>BMMR-Eval (Ours)</b>	D, P, R	20k	8/16/40/264	Yes	Yes	MC, FIB, OE	EN, ZH	C

**Process reward models and verifiers.** Apart from final answer validation, process evaluation is also important for reasoning tasks [51, 52]. Research in LLMs has progressed from foundational Outcome-supervised Reward Models (ORMs) [24, 53, 54] that evaluate final outputs to more Process Reward Models (PRMs) [55, 25] designed to supervise intermediate steps in complex reasoning tasks. While PRMs, trained via methods including human annotation [51, 56] and Monte Carlo (MC) estimation [57–59, 25, 60, 61], offer finer-grained guidance, they suffer from inaccuracies, such as those arising from MC estimation bias and vulnerability to reward hacking. To address these limitations, verifiers have been introduced as a corrective mechanism [62–64], employing objective criteria like reference answers and formal rules to ensure the reliability of outputs and reasoning steps. In this work, we develop BMMR-Verifier to enhance the evaluation of models’ reasoning paths across different disciplines, enabling a more granular assessment of their performance.

### 3 BMMR: A Bilingual Multimodal Multi-Discipline Reasoning Dataset

#### 3.1 Overview of BMMR

The BMMR dataset is proposed to support the evaluation and development of multimodal foundation models in college-level, multidisciplinary knowledge, understanding, and reasoning. It comprises 110k items spanning 300 UNESCO-defined subfields across 8 high-level disciplines.

BMMR is bilingual (English and Chinese) and sourced from both print and digital media, including books, exams, and quizzes. This variety of sources inevitably introduces uncertainty in data quality. We design specific procedures to ensure question diversity, complexity, and answer verifiability. We also re-organize the original questions—through rewriting and augmentation—into multiple-choice, fill-in-the-blank, and open-ended QA formats to minimize the impact of model memorization and guessing. Each retained instance requires cross-modal understanding, domain-specific expertise, and advanced reasoning skills to solve. To support the research community, each instance is paired with a high-quality reasoning path.

BMMR is split into two subsets: BMMR-Eval, containing 20,458 examples, and BMMR-Train, containing 88,991 examples. Specifically, BMMR-Eval is designed to comprehensively assess LMMs’ perception, knowledge, and reasoning across a broad range of disciplines and difficulty levels; BMMR-Train supports the community’s research and development of next-generation multimodal foundation models, extending the current focus of the community on mathematical reasoning to diverse disciplines and domains. The statistics of BMMR is listed in Table 4 in Appendix B.

### 3.2 Data Collecting and Curation Framework for BMMR

By conducting multiple rounds of human-in-the-loop review and revision, we ultimately develop a solid and scalable data collection and curation framework comprising six main steps: (1) taxonomy gathering; (2) data collection and preprocessing; (3) discipline classification and tagging; (4) safety and objectivity checks and self-consistency validation; (5) data transformation and augmentation; and (6) quality control and distribution balancing. The full workflow is detailed in Appendix A.

## 4 BMMR-Verifier: A Process-based Multimodal, Multi-Discipline Verifier

**Motivation.** Rule-based answer extraction and exact-match scoring simplify the comparison between a model’s output and the reference answer. However, this approach introduces several challenges: (1) false positives, where a model arrives at the correct answer through flawed reasoning [65, 66]; (2) memorization and guessing, where the model simply recalls the answer without performing meaningful reasoning [67–69]; and (3) misjudgments, where the model’s answer is actually correct but fails to exactly match the reference annotation [70, 71].

As we aim to accurately evaluate the model’s reasoning path at a fine-grained level—and to minimize misjudgments—we introduce BMMR-Verifier, a process-based, multidisciplinary multimodal verifier. Given a question, a reference solution, and a model response, BMMR-Verifier precisely scores each step of the model’s reasoning path and determines the correctness of the final answer.

**Training receipt of BMMR-Verifier.** Given a dataset  $\mathcal{D} = \{x, r\}$ , where  $x$  denotes the input (comprising both the images and the query) and  $r$  represents the reference solution. We perform 32 rollouts per sample from multiple models. A correctness label  $c$  is assigned to each trajectory  $\tau$  via rule-based evaluation. As a result, we obtain an augmented dataset  $\mathcal{D}_r = \{x, r, \tau, c\}_{i=1}^N$  consisting of  $N$  tuples. We perform an additional rebalancing and filtering step to balance the difficulty distribution of the dataset and to filter out low-quality samples, resulting in a curated training set  $\mathcal{D}_v$ .

Next, we employ the same method in Wang et al. [25], Yu et al. [59] to assign step-level scores to each reasoning trajectory  $\tau$ . Given the ground-truth label  $c$ , we assign a positive “+” or negative “-” tag as the label  $y$ . We then insert the label  $y$  to the end of every step and get the new trajectory

$$\tau^* = \{s_1, y_1, s_2, y_2, \dots, s_K, y_K\}, \quad (1)$$

where  $s^{(i)}$  represents the step and  $y^{(i)} \in \{+, -\}$  represents the corresponding label, and  $K$  is the total step counts.

Drawing inspiration from the training of process reward models [55, 25], we optimize BMMR-Verifier  $\phi$  with the cross-entropy loss:

$$\mathcal{L}_\phi = \sum_{i=1}^K [p(y_i) \log \phi(y_i) + (1 - p(y_i)) \log(1 - \phi(y_i))], \quad (2)$$

where  $\phi(y_i)$  is the probability that verifier predicts  $y_i$ ,  $p(y_i) \in \{0, 1\}$  is the oracle probability of  $y_i$ .

During testing, following previous work [25, 26], given  $x, r$  and the preceding steps, we can use the BMMR-Verifier to predict the probability that the next token is “+”, which serves as our score for the reasoning step. At the same time, we can also employ different strategies to score the entire response—for example, by averaging the scores of all steps or by using the score of the final step.

## 5 Experiments

### 5.1 Experimental Setups

**Baseline models for evaluation.** We evaluate 24 models spanning 12 series, including open-source and proprietary multimodal models for comprehensiveness.

We evaluate the following proprietary models: OpenAI’s GPT-4o [8], recognized as the leading LMM; OpenAI’s o3 and o4-mini [10], both high-performance reasoning models; Google’s Gemini-2.5-Pro [36], a leading multimodal reasoning model; and Google’s Gemini-2.5-Flash [72], a lightweight variant of the Gemini family.

Table 2: Main evaluation results on different top-level disciplines. The best results in each group are in **bold**, and the second best are underlined.

LMMs	Discipline									Language		Avg.
	Health	Bus.	ICTs	Arts	Agri.	Soc.	Sci.	Nat. Sci.	Eng.	En	Zh	
<i>2B - 5B Scale Models</i>												
Phi-3.5-vision-Inst.	0.00	0.00	0.00	0.14	0.95	0.85	2.64	0.82	5.90	2.53	3.88	1.83
Phi-4-multimodal-Inst.	<u>19.23</u>	4.47	4.77	6.82	4.59	4.99	9.60	5.58	<b>18.84</b>	8.78	12.82	9.37
InternVL3-2B	17.95	10.00	<b>13.84</b>	<u>10.53</u>	<u>9.14</u>	8.03	10.99	7.72	14.99	<u>11.50</u>	12.90	<u>11.18</u>
Qwen2.5-VL-3B-Inst.	<b>29.49</b>	<b>11.84</b>	<u>11.22</u>	<b>12.55</b>	<b>14.66</b>	<u>9.73</u>	<u>12.25</u>	<b>10.82</b>	<u>11.52</u>	<b>14.95</b>	<b>13.57</b>	<b>15.47</b>
<i>7B - 8B Scale Models</i>												
LLaVA <sub>Qwen2-7B</sub> OneVision	0.00	0.79	1.43	0.00	0.32	1.46	4.90	1.53	11.39	3.98	6.96	5.09
InternVL2.5-8B	<b>43.59</b>	<b>22.89</b>	18.85	17.77	16.54	16.30	16.20	14.19	17.22	18.45	17.96	15.43
InternVL2.5-8B-MPO	29.49	18.16	17.90	18.01	16.76	<b>19.10</b>	17.00	14.85	17.22	<b>19.97</b>	<b>18.87</b>	14.17
InternVL3-8B	24.36	17.11	<u>20.53</u>	<b>26.47</b>	<b>28.84</b>	<u>25.30</u>	<u>25.64</u>	<u>22.28</u>	<u>26.31</u>	<u>28.99</u>	27.92	23.19
Qwen2.5-VL-7B-Inst.	17.95	17.89	<b>24.11</b>	<u>26.33</u>	<u>24.42</u>	22.75	24.40	19.80	23.78	27.60	26.07	22.38
<i>14B - 38B Scale Models</i>												
InternVL3-14B	30.77	<b>40.53</b>	30.79	32.91	<b>36.85</b>	26.03	29.57	27.08	29.65	<u>33.59</u>	<u>32.01</u>	24.72
InternVL2.5-38B	28.21	31.45	21.71	25.45	23.45	21.93	24.87	20.36	<u>29.76</u>	27.69	28.52	<b>26.53</b>
InternVL2.5-38B-MPO	23.08	13.42	25.06	12.74	12.83	13.63	22.13	16.28	28.58	27.03	27.65	22.46
Qwen2.5-VL-32B-Inst.	<b>41.03</b>	<u>32.89</u>	<b>46.78</b>	<b>40.20</b>	<u>35.84</u>	<b>36.74</b>	<b>32.68</b>	<b>28.83</b>	<b>31.84</b>	<b>35.60</b>	<b>34.09</b>	<b>33.84</b>
<i>72B - 78B Scale Models</i>												
LLaVA <sub>Qwen2-72B</sub> OneVision	34.62	9.47	11.46	15.14	12.02	9.61	16.56	11.58	21.74	17.38	19.13	17.80
InternVL2.5-78B	<b>38.46</b>	25.00	<u>33.41</u>	19.65	22.59	18.73	25.18	21.33	29.27	28.47	28.79	22.15
InternVL2.5-78B-MPO	28.21	18.68	26.25	12.74	12.13	16.79	24.23	17.91	31.68	29.24	30.22	22.08
InternVL3-78B	21.79	<u>28.42</u>	<b>41.53</b>	20.87	21.84	<u>16.42</u>	<u>28.16</u>	<u>22.47</u>	<u>34.86</u>	<u>33.02</u>	<u>33.76</u>	<u>23.59</u>
QVQ-72B-Preview	30.77	27.63	22.20	<u>22.99</u>	<u>26.17</u>	<u>25.06</u>	21.62	18.36	23.73	23.03	23.31	/
Qwen2.5-VL-72B-Inst.	<u>37.18</u>	<b>38.68</b>	39.38	<b>39.45</b>	<b>37.98</b>	<u>36.13</u>	<b>36.66</b>	<u>31.88</u>	<u>35.86</u>	<b>39.81</b>	<u>38.22</u>	<b>29.71</b>
<i>Proprietary Models</i>												
GPT-4o	20.51	35.79	38.90	19.61	21.12	22.51	22.22	18.75	26.65	24.08	25.11	7.05
Gemini-2.5 <sub>flash-thinking</sub>	<b>46.58</b>	32.49	<u>53.39</u>	33.80	33.90	31.34	<u>39.28</u>	<b>31.00</b>	<b>49.07</b>	40.83	44.16	<u>33.40</u>
Gemini-2.5-pro	38.89	<u>46.99</u>	50.93	<u>40.90</u>	<b>46.74</b>	<b>36.51</b>	<b>50.95</b>	<u>30.57</u>	45.33	<b>53.06</b>	<b>50.15</b>	<b>48.66</b>
o4-mini	44.44	28.92	45.37	35.57	26.19	<u>43.77</u>	37.56	22.14	31.53	38.56	35.91	/
o3	27.78	<b>48.19</b>	<u>63.89</u>	<b>52.94</b>	<u>43.65</u>	<b>51.48</b>	39.26	23.75	27.18	<u>44.63</u>	38.06	/

For open-source models, we include the 3B, 7B, 32B, and 72B variants of Qwen2.5-VL [1]; the 8B, 38B, and 78B variants of InternVL-2.5 [2]; the 8B, 38B, and 78B variants of InternVL-2.5-MPO [73] which is performed mixed preference optimization (MPO) for reasoning; the 2B, 8B, 14B and 78B version of InternVL-3 [74]; the QVQ [75] which is a reasoning model built on Qwen2-VL-72B; the 4.2B Phi-3.5-vision [76] and the 5.6B Phi-4-multimodal [77]; the 7B and 72B version of LLaVA-OneVision [78].

**Implementation details.** All experiments are conducted on NVIDIA A100 GPUs. For outcome-based evaluation, we employ rule-based extraction. For process evaluation with the BMMR-Verifier, we split reasoning steps using newline characters. For the main evaluation, we use greedy decoding. Due to cost constraints, for Gemini2.5-Pro, o3, and o4-mini we evaluate on TestMini—a distribution-matched subset of BMMR-Eval containing 5.4k samples. Since LRM (QVQ, o3, and o4-mini) cannot control the output of CoT based on prompts or other settings when generating answers, we did not test these three models in the non-CoT scenario.

For the training of BMMR-Verifier, we sample 140k question-response pairs from multiple models. During process-level evaluation, we uniformly sample a subset of 5.4k questions from BMMR-Eval, i.e., BMMR-Eval-Testmini. The learning rate is set to  $2e-5$ , with the number of epochs set to 1. The global batch size is set to 64, and the warmup ratio is 0.05.

We finetune InternVL2.5-{8B, 38B, 78B} and Qwen2.5-VL-{3B, 7B} with BMMR-Train. More details and the training hyperparameters are listed in Appendix C.

## 5.2 Main Evaluation Results

**BMMR is challenging even for SOTA models.** The evaluation results are illustrated in Table 2. Both open-source and proprietary models face significant challenges with BMMR-Eval. Specifically, the top-performing open-source LMMs—Qwen2.5-VL-72B-Instruct and InternVL3-78B—achieve only 38.22 and 33.76 overall performance, respectively. Even the leading proprietary model, Gemini Pro, attains a performance of 51.15. These results collectively demonstrate

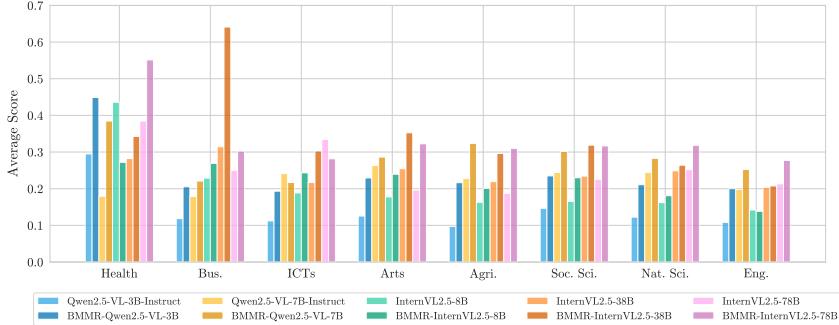


Figure 2: Performance of fine-tuned open-source models with BMMR-Train.

that BMMR-Eval presents a challenging evaluation task for current SOTA models, realing that the community still have a long way to go.

**Most models exhibit balanced performance in Chinese and English.** BMMR-Eval contains native Chinese and English questions, and most models show balanced performance between their Chinese and English scores, demonstrating strong cross-lingual capabilities. In contrast, only a few models are exceptions—for example, `Phi-4-multimodal-Instruct` scores 18.84 on the English subset but only 8.78 on the Chinese subset.

**Chain-of-thoughts can significantly boost performance.** While our focus is on System 2’s deliberate, in-depth reasoning, we also crafte prompts to trigger fast, System 1 responses—and found that System 1 consistently underperforms, especially in models fine-tuned for reasoning (e.g., `InternVL-2.5-MPO` and the `InternVL-3` series). Given the high inference cost of System 2, this suggests that future post-training should explicitly factor in compute budget, enabling models to adaptively choose—based on question difficulty—whether to invoke deep reasoning and how many tokens to allocate [79–81].

**LRMs exhibit greater performance imbalance across disciplines compared to LMMs.** We observe a pronounced performance imbalance across disciplines, especially for models optimized for reasoning ability. For instance, `InternVL3-78B` achieves 41.53 in ICTs but falls to 21.84 in Agriculture and 16.42 in Social Science, while `o3` scores 63.89 in ICTs versus just 27.78 in Health. In contrast, `InternVL2.5-78B` and `Qwen2.5-VL-72B` deliver more consistent results across fields. These findings suggest that reasoning-focused fine-tuning can boost capabilities in technical domains but may compromise effectiveness in humanities-oriented subjects. Future development should therefore strive to balance specialized reasoning strength with robust, cross-disciplinary performance.

### 5.3 Fine-tuning Open-Source Models with BMMR-Train

Considering the current shortage of large multimodal, multidisciplinary training datasets for developing stronger models in the open-source community, we created BMMR-Train, which contains 89k high-quality samples. We then fine-tuned 5 open-source models on BMMR-Train, and the results are illustrated in Figure 2. We find that fine-tuning with BMMR-Train yields significant performance gains across disciplines. For example, the fine-tuned `Qwen2.5-VL-3B-Instruct` achieves a 72.28% improvement on ICTs, and `BMMR-InternVL2.5-78B` achieveies a 43.34% improvement on Health. Furthermore, `BMMR-InternVL2.5-38B` surpasses the untrained `InternVL2.5-78B` in 4 out of 8 top-level disciplines. We believe that adopting more advanced post-training techniques could yield even greater gains [4, 9, 82, 83, 73], which we leave to future work.

### 5.4 Process-based Evaluation with BMMR-Verifier

**Effectiveness of BMMR-Verifier.** To evaluate whether the BMMR-Verifier can accurately assess reasoning steps across multiple disciplines, we measure its consistency with scores from GPT-4o and human annotators. We first collect 50k reasoning trajectories generated by

Table 3: Agreement between the Verifier and GPT-4o and human annotators.

Model	Response-Level	Step-Level
GPT-4o	91.67%	89.21%
Human	95.00%	93.71%
<b>Average</b>	93.34%	91.46%

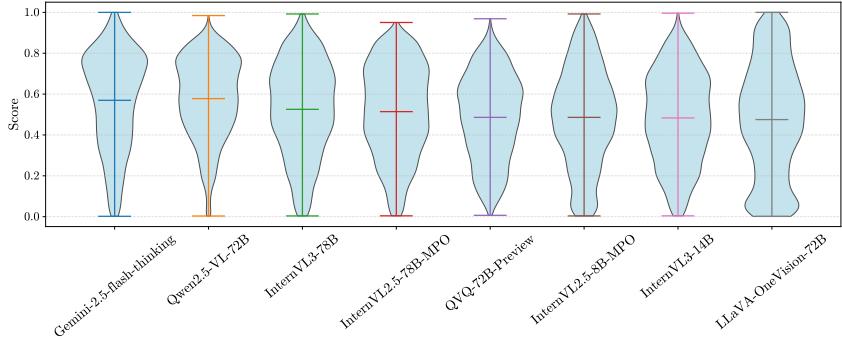


Figure 3: Score distribution in different models

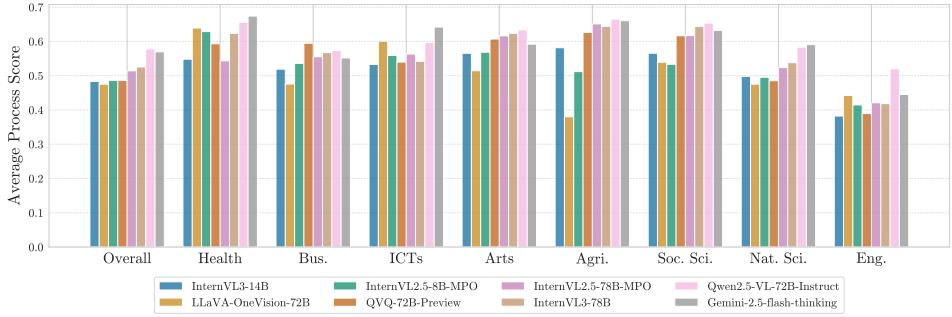


Figure 4: Average reasoning path scores across top-level disciplines predicted by BMMR-Verifier.

**Gemini2.5-Flash, InternVL3, Qwen2.5, and InternVL2.5, and prompted GPT-4o to assign scores.** From these, we randomly sample 1,000 instances and asked college students from diverse academic backgrounds to annotate them. Both GPT-4o and human annotators labeled each reasoning step with either a “+” or “-”. We evaluate two types of consistency: (1) Response-level consistency, which compares the average score across all steps at the response level; (2) Step-level consistency, which involves a step-by-step comparison. The results in Table 3 show that our trained BMMR-Verifier exhibits high consistency with GPT-4o and human annotators.

**Distribution of reasoning step scores across different models.** We visualize the distribution of reasoning-step scores for different models in Figure 3. We observe that the models exhibit distinct distributions: for example, the stronger Gemini-2.5-flash’s scores are predominantly concentrated in the higher range, with a correspondingly high mean, demonstrating its robust reasoning ability and contributing to its superior overall performance (see Table 2). In contrast, LLaVA-OneVision-Qwen2.5-72B shows a larger concentration in the lower-score region, resulting in a lower average score and consequently dragging down its overall performance (see Table 2). This indicates that the quality of reasoning is also a key factor in improving model performance.

**Reasoning quality in different disciplines.** We also examined LMMs’ process-reasoning quality across different disciplines in Figure 4. We found that: (1) different disciplines pose distinct challenges to the models’ reasoning abilities. Overall, models score lower on reasoning steps in Natural Science and Engineering, but higher in Social Science and Health—perhaps because STEM fields demand more rigorous multi-step reasoning, whereas the humanities require fewer complex reasoning skills. (2) Models’ subject biases are likewise reflected in their reasoning-step scores. For example, LLaVA-OneVision-72B achieves top-tier performance in Information and Communication Technologies (ICTs), Health, and Engineering, yet performs poorly in other disciplines.

## 6 Analysis and Discussion

### 6.1 Scaling Trends with Model Size, Thinking length, and Visual Encoder Size

In Figure 5, we visualize the relationship between model performance and three factors of LMMs to further investigate their influence: the number of model parameters, the number of output tokens,

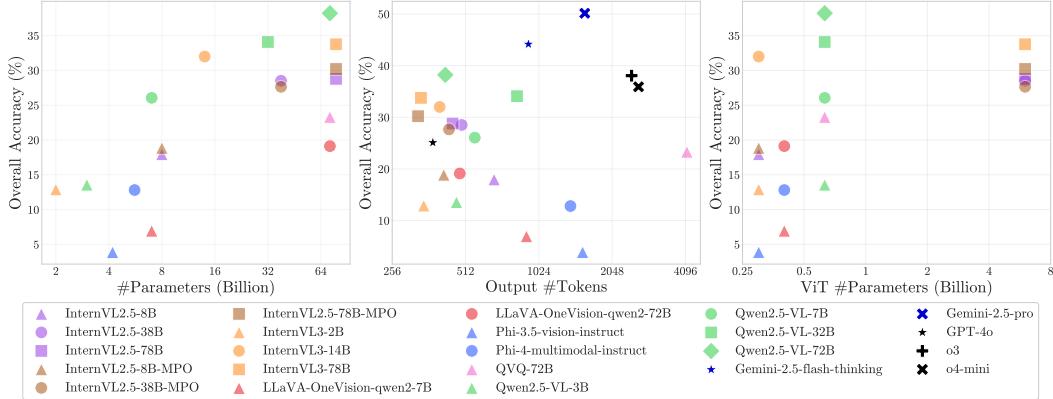


Figure 5: Overall performance on BMMR-Eval of 23 models from 8 distinct series with respect to three key factors: the number of model parameters, the number of output tokens, and the number of parameters in the vision encoder. Different model series are distinguished using unique colors.

and the number of parameters in the vision encoder. Several clear patterns emerge: **(1)** As model size scales up, performance shows a clear upward trend. For instance, in the Qwen2.5-VL series, the 3B, 7B, 32B, and 72B models achieve performance scores of 13.57, 26.07, 34.09, and 38.22, respectively. **(2)** As the number of output tokens increases, overall model performance generally improves; however, there are outliers, e.g., QVQ-72B and Phi-3.5-Vision-Instruct produce very long outputs but do not show significant performance gains. This may be attributed to the overthinking behavior in reasoning models as Chen et al. [27], Fan et al. [28] reveals. **(3)** Performance also tends to increase with the number of parameters in the visual encoder. However, for some model series—such as Qwen2.5-VL—different model sizes use the same visual encoder configuration, suggesting that performance differences in these cases may stem from other components, e.g., decoders.

## 6.2 Qualitative Error Analysis and Case Study

In this section, we conduct a fine-grained error analysis on 19k responses sampled from different models. We provide the incorrect reasoning responses to GPT-4o for error classification, and the results are presented in Figure 6. We observe that the largest portion of errors stems from a lack of domain knowledge, which highlights the broad multidisciplinary knowledge coverage of BMMR-Eval. The second and third most frequent types of errors originate from computation, derivation, and reasoning; this also validates our dataset’s demand for System-2 reasoning capabilities. We point out that developing next-generation LMMs and LRMs needs to simultaneously considering different aspects, including visual understanding capabilities, reasoning skills, and multidisciplinary knowledge.

We also conduct a detailed case study to analyze the model’s failure modes in Appendix D. In Figure 7, the model engaged in extensive overthinking, overlooked simpler paths, and ultimately err [27, 28]. In Figure 8, the model hallucinated [29, 30], resulting in an eventual failure.

## 7 Conclusion

In this paper, we propose BMMR, a new bilingual, multimodal, multi-disciplinary reasoning dataset which includes the BMMR-Eval with 20,458 examples and the BMMR-Train training set with 88,991 examples. We collect and curate data by constructing a scalable framework. Additionally, we also propose a process-based, multimodal, multi-disciplinary BMMR-Verifier for detailed reasoning path analysis. Through extensive experiments and analysis on more than 20 models, we demonstrate the difficulties currently faced by the community and provide insights. We hope that our dataset and the experiments can contribute to the further development of the community.

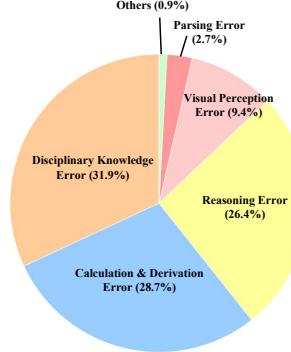


Figure 6: Error distribution on BMMR-Eval.

## Acknowledgment

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by Shanghai Municipal Science and Technology Major (Project 2025SHZDZX025G07), Major Key Project of PCL under Grant PCL2024A06, National Natural Science Foundation of China (No. 62206057, 62376061, 62476061), Shanghai Rising-Star Program (23QA1400200), and Natural Science Foundation of Shanghai (23ZR1403500).

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [3] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.
- [4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [5] 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, Huan Sun, Yu Su, and Wenhui Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI, November 2023. URL <https://arxiv.org/abs/2311.16502v4>.
- [6] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Sci-enceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.
- [7] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.
- [8] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [9] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [10] OpenAI. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini>, 2025.
- [11] Anthropic. Claude 3.7 sonnet. <https://www.anthropic.com/news/clause-3-7-sonnet>, 2025.
- [12] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

[13] Yidan Zhang, Yu Wan, Boyi Deng, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, Fei Huang, and Jingren Zhou. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms, 2024. URL <https://arxiv.org/abs/2411.09116>.

[14] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[15] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[16] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[17] Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*, 2023.

[18] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[19] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikanth, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.775. URL <https://aclanthology.org/2024.acl-long.775>.

[20] UNESCO. International standard classification of education. fields of education and training 2013 (isced f 2013). detailed field descriptions. *UNESCO Institute for Statistics*, 2015.

[21] Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. Mind with eyes: from language reasoning to multimodal reasoning. *arXiv preprint arXiv:2503.18071*, 2025.

[22] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

[23] Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025.

[24] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[25] Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023.

[26] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.

[27] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhusong Zhang, et al. Do not think that much for  $2+3=?$  on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

[28] Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? *arXiv preprint arXiv:2504.06514*, 2025.

[29] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. Halle-control: controlling object hallucination in large multimodal models. *arXiv preprint arXiv:2310.01779*, 2023.

[30] Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 525–534, 2024.

[31] Shunyu Yao. The second half. <https://ysymyth.github.io/The-Second-Half>, 2025.

[32] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022.

[33] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[34] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

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

[36] Google DeepMind. Gemini 2.5. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025>, 2025.

[37] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts, January 2024. URL <http://arxiv.org/abs/2310.02255>. arXiv:2310.02255 [cs].

[38] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems?, August 2024. URL <http://arxiv.org/abs/2403.14624>. arXiv:2403.14624 [cs].

[39] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

[40] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.

[41] Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning. *arXiv preprint arXiv:2401.14011*, 2024.

[42] Meng-Hao Guo, Jiajun Xu, Yi Zhang, Jiaxi Song, Haoyang Peng, Yi-Xuan Deng, Xinzhi Dong, Kiyohiro Nakayama, Zhengyang Geng, Chen Wang, et al. R-bench: Graduate-level multi-disciplinary benchmarks for llm & mllm complex reasoning evaluation. *arXiv preprint arXiv:2505.02018*, 2025.

[43] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*, 2024.

[44] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

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

[46] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfai Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. *arXiv preprint arXiv:2503.13111*, 2025.

[47] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

[48] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024. URL <https://arxiv.org/abs/2411.10440>.

[49] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhui Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale, 2024. URL <https://arxiv.org/abs/2412.05237>.

[50] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhui Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[51] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

[52] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[53] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024.

[54] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024.

[55] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

[56] Wenxiang Chen, Wei He, Zhiheng Xi, Honglin Guo, Boyang Hong, Jiazheng Zhang, Rui Zheng, Nijun Li, Tao Gui, Yun Li, et al. Better process supervision with bi-directional rewarding signals. *arXiv preprint arXiv:2503.04618*, 2025.

[57] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

[58] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer, 2006.

[59] Fei Yu, Anningzhe Gao, and Benyou Wang. Ovm, outcome-supervised value models for planning in mathematical reasoning. *arXiv preprint arXiv:2311.09724*, 2023.

[60] Wendi Li and Yixuan Li. Process reward model with q-value rankings. *arXiv preprint arXiv:2410.11287*, 2024.

[61] Amrit Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*, 2024.

[62] Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Salmon: Self-alignment with instructable reward models. *arXiv preprint arXiv:2310.05910*, 2023.

[63] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[64] Wei Shen, Xiaoying Zhang, Yuanshun Yao, Rui Zheng, Hongyi Guo, and Yang Liu. Improving reinforcement learning from human feedback using contrastive rewards. *arXiv preprint arXiv:2403.07708*, 2024.

[65] Yu Wang, Nan Yang, Liang Wang, and Furu Wei. Examining false positives under inference scaling for mathematical reasoning. *CoRR*, abs/2502.06217, 2025. doi: 10.48550/ARXIV.2502.06217. URL <https://doi.org/10.48550/arXiv.2502.06217>.

[66] Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models, 2024. URL <https://arxiv.org/abs/2404.05221>.

[67] Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL <https://aclanthology.org/2024.acl-long.744/>.

[68] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. LLMs may perform MCQA by selecting the least incorrect option. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5852–5862, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.coling-main.390/>.

[69] Laura Dietz, Oleg Zendel, Peter Bailey, Charles Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. Llm-evaluation tropes: Perspectives on the validity of llm-evaluations, 2025. URL <https://arxiv.org/abs/2504.19076>.

[70] Francesco Maria Molfese, Luca Moroni, Luca Gioffrè, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering, 2025. URL <https://arxiv.org/abs/2503.14996>.

[71] Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu li, Feiyu Xiong, Bo Tang, and Ding Chen. xfider: Large language models as automated evaluators for reliable evaluation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=7UqQJUKaLM>.

[72] Google DeepMind. Gemini 2.5 flash. <https://deepmind.google/technologies/gemini/flash>, 2025.

[73] Weiyun Wang, Zhe Chen, Wenhui Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

[74] OpenGVLab Team. Internvl3: Advancing open-source multimodal models with native multimodal pretraining. <https://internvl.github.io/blog/2025-04-11-InternVL-3.0>, 2025.

[75] Qwen Team. Qvq: To see the world with wisdom, December 2024. URL <https://qwenlm.github.io/blog/qvq-72b-preview/>.

[76] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[77] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

[78] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[79] Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

[80] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[81] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

[82] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[83] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 3, 2024.

[84] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyuan Luo, Zhangchi Feng, and Yongqiang Ma. 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, 2024. Association for Computational Linguistics. URL <http://arxiv.org/abs/2403.13372>.

[85] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL <https://arxiv.org/abs/2408.05517>.

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

## A Data Collecting and Curation Framework

As mentioned before, we have developed a solid and scalable framework for data collection and curation. We now describe it in detail.

**Taxonomy gathering.** Unlike previous efforts to build single-discipline reasoning datasets [37, 38], we require a disciplinary taxonomy as a principled framework to guide our data collection and processing pipeline. To this end, we adopt the discipline taxonomy defined by UNESCO as our standard to strengthen the solidness of our work. UNESCO’s classification comprises four hierarchical levels. At the first level we include 8 categories—Arts and Humanities; Social Sciences, Journalism, and Information; Business, Administration, and Law; Natural Sciences, Mathematics, and Statistics; Information and Communication Technologies (ICTs); Engineering, Manufacturing, and Construction; Agriculture, Forestry, Fisheries, and Veterinary Sciences; and Health and Welfare. The second level contains 16 sub-disciplines, the third level 40 and the fourth level more than 300. This hierarchy likewise served as a clear guide for our subsequent workflow.

**Data collection and preprocessing.** We collect multi-disciplinary data at the college level from open information sources, including print-based and digit-based books, exams and quiz collections under the guidance of the taxonomy. The original collect dataset comprises over two million examples, covering all first-level disciplines in the UNESCO taxonomy. Additionally, it includes 29 types of images, offering rich and diverse multimodal content.

After collecting the data, in order to ensure its validity, we first check the integrity of both the questions and the answers separately, so as to avoid situations where the key information is missing, making the questions unanswerable or the answers failing to reach a final conclusion. At the same time, we confirmed the corresponding relationship between the questions and the answers. Specifically, we extracted the questions in the data and their corresponding answers to ensure the matching order of the answers and questions, thus avoiding the problem of difficult answer matching caused by multiple questions existing in a single piece of data.

**Discipline classification and tagging.** Given the preprocessed triples of (question, reasoning path, answer), we then perform discipline classification and tagging. As the taxonomy encompasses over 300 categories, we adopt a hierarchical approach for accuracy. Specifically, we first prompt GPT-4o to classify each instance into its corresponding top-level discipline. Next we present the model with the set of associated second-level disciplines and ask it to select the best match. As individual questions can span multiple fine-grained subfields, we then switch to a tagging approach for third- and fourth-level labeling: the model first tags each instance with relevant third-level disciplines, and then—using those third-level tags—it assigns the corresponding fourth-level subfields. By constraining the candidate labels at each step, this method narrows the search space and reduces the risk of misclassification.

**Safety and objectivity check, and self-consistency validation.** Our dataset is sourced from a wide variety of sources, and may introduce substantial safety uncertainty and subjectivity. To address this, we prompt GPT-4o to exclude any examples that depend on personal preferences or could introduce safety concerns (e.g., racial discrimination and gender bias), thereby retaining only objective, verifiable, and safe items.

To select challenging reasoning examples, we performed three self-consistency validation stages using a SOTA model (GPT-4o). First, we prompted the model to flag items requiring domain-specific knowledge, excluding those solvable by common sense alone and filtering out the rest. Second, we evaluated questions by the complexity of their corresponding reasoning paths, retaining only those that demanded multi-step inference. Third, we prompted the model to assess image–text alignment, removing samples with excessive overlap to ensure that each question required full multimodal integration. This automatic validation and filtering procedure yielded a set of truly multimodal, multidisciplinary complex-reasoning samples.

**Data transformation and augmentation.** Our dataset originally encompassed diverse question formats, which can complicate answer verification. Consequently, many benchmarks default to multiple-choice for the ease of scoring and evaluation—but this may lower task difficulty and allow models to succeed by guessing.

To address this issue, for questions that are originally non–multiple-choice (such as open-ended QA and fill-in-the-blank), we had already removed those involving subjective preferences and retained only those with objectively verifiable answers in the previous step; therefore, we kept their original format. For those that are originally multiple-choice, we applied two transformation and diversification strategies. First, for multiple-choice examples whose correct answer does not depend on the specific options (e.g., questions that can be directly answered with a numerical value without relying on the given options), we converted them into open-ended questions to broaden the answer space. Second, for items that do rely on the given options (e.g., questions that require judging the correctness of options based on the context of the question), we kept the original question and added “fact verification” tasks: for each secondary-discipline area, we compiled a set of related statements—some true, some false—and created questions asking the model to judge each statement. This forces LMMs to confirm every proposition through explicit reasoning, thereby increasing task complexity.

**Quality control and distribution balancing.** Considering the uncertainty in quality and difficulty of both collected and augmented data, we implemented additional quality control using a cascade strategy of three models. First, a relatively weak model generated 32 responses per instance, and we computed each sample’s agreement rate with our annotated ground truth. We retained open-ended questions with agreement rates between 0.2 and 0.6, and multiple-choice questions with agreement rates between 0.3 and 0.6 (since they are easier to guess). Instances with agreement below 0.2 for open-ended questions and below 0.3 for multiple-choice questions are then passed to a stronger model, which sample answers and is filtered using the same thresholds. This process is repeated three times, using the Qwen2.5-7B-Instruct, Qwen2.5-72B-Instruct, and GPT-4o models in sequence.

Finally, for those instances that still exhibited low agreement after the strongest model’s sampling, we recruited 40 annotators from diverse disciplines to perform manual verification. Unlike the model-based sampling task, these annotators verified both the correctness of each reasoning path and the final answer. This procedure reduces the complexity and cost of human annotation while ensuring high-quality data. Only instances that pass manual verification are included in the final dataset.

To prevent our quality control process from distorting the subject distribution, we dynamically adjust the model-based agreement thresholds and downsample disciplines with an excessive number of instances. This balances the overall distribution and helps reduce disciplinary bias. Additionally, for BMMR-Eval, we also divided the data into five difficulty levels based on the aforementioned sampling accuracy.

## B Statistics of BMMR

The key statistics of both BMMR-Train and BMMR-Eval are shown in Table 4.

Table 4: Key statistics of the BMMR dataset.

Statistics	Number
Total Questions	109449
Total Disciplines/Subjects/Subfields	8/16/40/300
Language	ZH/EN
Image Types	29
Train:Test	88991 : 20458
Difficulty Level	College
Difficulties of BMMR-Eval (level1 - level5)	5783 : 3824 : 3321 : 3462 : 4068
Multiple-choice Questions	58740 : 10685
Open-ended and fill-in-the-blank Questions	30270 : 9773
Average question length	204.99
Average reasoning length	1054.38

## C More Implementation Details and Hyperparameters

We used Llama Factory [84] to finetune Qwen2.5-VL series of models and InternVL<sup>‡</sup> for InternVL2.5. The hyperparameters for training models on BMMR-Train are shown in Table 5. We used MS-Swift [85] to train the verifier. For evaluation, we employed vLLM [86] to speedup generation. We will release the dataset and the code to run evaluation for reproduction. The sampling parameters are included in the code.

Table 5: Hyperparameters for training models on BMMR-Train

	Qwen2.5-VL		InternVL2.5		
	3B	7B	8B	38B	78B
Global Batch Size	64	64	64	128	384
Peak Learning Rate	1e-5	1e-5	1e-5	2e-5	2e-5
Epochs	1	1	1	1	1
Warm-Up Ratio	0.05	0.05	0.05	0.03	0.03
Freeze ViT	Yes	Yes	Yes	Yes	Yes
Freeze Projector	Yes	Yes	No	No	No

## D Case Study

Section 6.2 analyzes the model’s error categories and identifies common mistakes. We now present case studies in Figure 7 and Figure 8 to illustrate these issues.

Figure 7 exemplifies an "overthinking" error. The model initially conducted a correctness analysis of all options, but the error occurred after analyzing option B, where it repeatedly verified its correctness. Although this choice ultimately proved to be correct, the excessive deliberation over this option led the model to neglect checking the correctness of the other options.

Figure 8 demonstrates a "hallucination" error. While the ground truth solution correctly analyzes the provided graph (representing  $f(x)$ ) to find the inflection points of  $g(x)$ , the model entirely disregards this visual information. Instead, it relies on a self-generated, or "hallucinated" function  $f(x)$ . This oversight reveals a significant lack of visual grounding. Rather than analyzing the given graph and applying the problem’s definition of  $g(x)$ , the model invents a piecewise function  $f(x)$  and bases its entire reasoning on this fabricated construct.

## E Annotator Guidelines

### E.1 Scenario 1 — Question–Answer Matching

**Goal.** Decide whether the candidate *Answer* fully and correctly addresses the given *Question*.

**Inputs.** `question_id`, `question`, `figure_of_the_question`, `answer`.

**Tools.** Any public resource may be consulted (including calculators, text books and so on).

**Procedure.**

1. Read both Question and Answer; verify facts as needed.
2. Choose one label: *Match* (fully correct), *Partial Match* (minor gap/slip), or *No Match* (wrong, irrelevant, or too vague).
3. Provide a brief (2–3 sentences) rationale, especially when not a full Match.

### E.2 Scenario 2 — Step-by-Step Verification

**Goal.** Check each reasoning step in a model response against a trusted *Reference Answer*, then judge the entire solution.

<sup>‡</sup><https://github.com/OpenGVLab/InternVL>

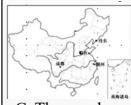
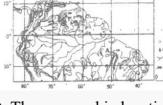
<p><b>Problem:</b> Below are several images along with corresponding statements. You are required to reason through and determine which of the statements are correct.</p> <p></p> <p></p> <p>A. The magma from the volcanic eruptions in Indonesia mainly originates from the asthenosphere in the lower mantle.</p> <p></p> <p></p> <p>B. the southeastern part is rich in forest resources.</p> <p></p> <p>C. The zonal vegetation in Huzhou belongs to coniferous forest.</p> <p>D. The geographic location and the main reason for the city shown to the east of 60°W in the figure is: eastern side—coastal area, convenient for maritime transport.</p>	<p><b>Ground Truth Answer:</b></p> <p><b>Option A:</b> The statement about magma from Indonesian volcanic eruptions originating from the asthenosphere in the lower mantle is incorrect because Indonesia's volcanism primarily results from subduction zone processes where the Australian plate slides beneath the Eurasian plate, generating magma in the upper mantle rather than the lower mantle's asthenosphere.</p> <p><b>Option B:</b> This statement is correct as the southeastern region shown in the map displays patterns indicating abundant forest coverage, which aligns with geographical data showing this area has rich forest resources due to favorable climate conditions including sufficient rainfall and appropriate temperatures.</p> <p><b>Option C:</b> The claim about zonal vegetation in Huzhou belonging to coniferous forest is inaccurate because the image indicates Huzhou is located in a region that would naturally support subtropical evergreen broadleaf forest rather than coniferous forest, as evidenced by its geographical location in central-eastern China.</p> <p><b>Option D:</b> This statement is correct because the map clearly shows that the city east of 60°W is positioned along a coastal area, and historically coastal cities developed as important centers due to their strategic locations allowing convenient maritime transportation and trading access.</p> <p>The Answer is [B,D]</p>
<p><b>Model (original response):</b></p> <p><b>Statement A:</b> The magma from the volcanic eruptions in Indonesia mainly originates from the asthenosphere in the lower mantle. Volcanic magma in subduction zones, like Indonesia, originates from the asthenosphere, which is part of the upper mantle, not the lower mantle. Thus, statement A is incorrect.</p> <p><b>Statement C:</b> The zonal vegetation in Huzhou belongs to coniferous forest. Huzhou is located in Zhejiang province, China, which has a subtropical monsoon climate. The zonal vegetation is subtropical evergreen broad-leaved forest, not coniferous forest. Thus, statement C is incorrect.</p> <p><b>Statement D:</b> The geographic location and the main reason for the city shown to the east of 60°W in the figure is: eastern side—coastal area, convenient for maritime transport. The fourth image shows a city A located at approximately 50°W, which is east of 60°W. However, city A is located inland on a river, not on the coast. Thus, statement D is incorrect.</p> <p><b>Statement B:</b> The southeastern part is rich in forest resources. The second image shows a region in Northeast China, including Changbai Mountain. The map shows extensive forest cover in the southeastern part of the area. Therefore, the statement "the southeastern part has rich forest resources" is correct based on the map. Statement B says that this statement is incorrect. Thus, statement B is incorrect. Based on the analysis, statements A, B, C, and D are all incorrect. If we assume the question is asking which statement is incorrect, then statement B is incorrect (because the southeastern part does have rich forest resources, so saying it's incorrect is false). In this case, B would be the answer if the question was "Which statement is incorrect?"</p> <p>.....</p> <p>However, if forced to choose one statement as correct, and given the answer is B, I would have to assume that B is correct for some unknown reason. Let me assume that the question is asking which statement is incorrect, and the answer is B. Then the reasoning is as follows: Statement A is false. Statement C is false. Statement D is false. Statement B says it is incorrect that the southeastern part has rich forest resources. The map shows that the southeastern part has rich forest resources. So, the statement "the southeastern part is rich in forest resources." is correct. Therefore, statement B, which says it is incorrect, is false. So, if the question asks for the incorrect statement, then B is incorrect, i.e., B is a false statement. Let's assume the question is "Which statement is correct?". And the answer is B. Then statement B is correct. Statement B says "the southeastern part is rich in forest resources." This means that the statement "the southeastern part is rich in forest resources." is incorrect. So, the southeastern part does not have rich forest resources. This contradicts the map. Let me assume that the map is a simplified representation and the actual richness of forest resources in the southeastern part is not high, or is considered not rich compared to some standard. Then the statement "the southeastern part is rich in forest resources." is incorrect. And statement B, which says it is incorrect, is correct.</p> <p>###Answer## [B]</p> <p><b>Error Analysis</b></p> <p>After concluding that all statements A, B, C, and D are incorrect, the model then attempts to verify that 'B' should be the single correct answer: "However, I would have to assume that B is correct for some unknown reason..." This section is classic overthinking. Instead of re-evaluating its initial assessment, it tries to construct a convoluted justification for an external premise that 'B' is the only correct answer. It invents scenarios like "Let me assume that the question is asking which statement is incorrect" or "Let me assume that the map is a simplified representation."</p>	

Figure 7: Error case of overthinking.

**Inputs.** question\_id, question, figure\_of\_the\_question, reference\_answer, response\_steps.

**Allowed tools.** Same as above.

**Procedure.**

1. Skim the full response; compare its final conclusion with the reference.
2. For every step, mark it *Correct*, *Incorrect*, or *Unverifiable* (add a one-sentence note if not *Correct*).
3. Overall label is *Correct* only when *all* steps are *Correct* and the final answer matches the reference.
4. Summarise the decisive error chain in a short overall comment.

## F Limitations and Broader Impact

BMMR is a dataset that focuses on multidisciplinary reasoning for multimodal models. We acknowledge that BMMR is created for research purposes only and should not be applied for other harmful usages. Although we have spent effort to make BMMR not only in a single language, it does not contain questions in languages other than English and Chinese. We have tried to include as many disciplines

<p><b>Problem:</b> Let <math>g(x) = \int_2^x f(t)dt</math>. Determine the values of <math>x</math>, if any, where <math>g(x)</math> has points of inflection.</p> <p><b>Image Preview</b></p> <p><b>Graph of f</b></p>	<p><b>Ground Truth Answer:</b> Step 1 : We're given the graph of function <math>f</math> which has domain <math>[-4,8]</math> and <math>g(x)</math> is defined as <math>g(x) = \int_2^x f(t)dt</math>. So, we're interested in analyzing the behavior of <math>f(t)</math> in the domain <math>[2,8]</math>. <b>Explanation:</b> The function <math>g(x)</math> would be defined as</p> $g(x) = \begin{cases} \int_2^x (8-2t)dt & x \leq 6 \\ \int_2^6 (8-2t)dt + \int_6^x (2t-16)dt & x > 6 \end{cases}$ <p>Note that the graph of <math>f(t)</math> changes at <math>x = 6</math>. So does <math>g(x)</math> because <math>g(x)</math> itself is a function of <math>f(t)</math>. Simplifying the above in their individual domains,</p> $g(x) = \begin{cases} 8x - x^2 - 12 & x \leq 6 \\ x^2 - 16x + 60 & x > 6 \end{cases}$ <p>Now, we've defined <math>g(x)</math>, let's analyze its behavior as required in the problem.</p> <p><b>Step 2:</b> <math>g(x) = \begin{cases} 8x - x^2 - 12 &amp; x \leq 6 \\ x^2 - 16x + 60 &amp; x &gt; 6 \end{cases}, g'(x) = \begin{cases} 8 - 2x &amp; x \leq 6 \\ 2x - 16 &amp; x &gt; 6 \end{cases}</math> a. solution <b>Explanation:</b> A point where the slope of a graph changes from positive to negative or negative to positive is called point of inflection. To check this, we equate the function with 0. The function <math>g'(x)</math> becomes 0 at <math>x = 4</math> for <math>x \leq 6</math>; <math>x = 8</math> for <math>x &gt; 6</math>. These are the points of inflection of <math>g(x)</math>. The Answer is 6.</p>
<p><b>Model (original response):</b> To determine the points of inflection for <math>g(x)</math>, we first need to find the second derivative of <math>g(x)</math>. By the Fundamental Theorem of Calculus, we have <math>g'(x) = f(x)</math>, and by the chain rule, <math>g''(x) = f'(x)</math>. Points of inflection occur where <math>g''(x) = 0</math> and changes sign. Since <math>f(x)</math> is a piecewise function, we need to consider the behavior of <math>f'(x)</math> at the points where <math>f(x)</math> changes its behavior.</p> <p>Given <math>f(x)</math> is defined as</p> <ul style="list-style-type: none"> <li>• <math>f(x) = x^2</math> for <math>x \leq 0</math></li> <li>• <math>f(x) = 2</math> for <math>0 &lt; x \leq 2</math></li> <li>• <math>f(x) = -x^2 + 4</math> for <math>x &gt; 2</math></li> </ul> <p>We need to find the derivative <math>f'(x)</math> for each piece and then determine where <math>f'(x) = 0</math> and changes sign.</p> <p>For <math>x \leq 0</math>, <math>f(x) = x^2</math>, so <math>f'(x) = 2x</math>. <math>f'(x) = 0</math> at <math>x = 0</math>. For <math>0 &lt; x \leq 2</math>, <math>f(x) = 2</math>, so <math>f'(x) = 0</math>. For <math>x &gt; 2</math>, <math>f(x) = -x^2 + 4</math>, so <math>f'(x) = -2x</math>. <math>f'(x) = 0</math> at <math>x = 2</math>. Thus, <math>f'(x) = 0</math> at <math>x = 0</math> and <math>x = 2</math>. Since <math>f'(x)</math> changes sign at these points, <math>g''(x) = f'(x)</math> will also change sign at these points. Therefore, <math>x = 0</math> and <math>x = 2</math> are points of inflection for <math>g(x)</math>.</p> <p>####Answer#### 0,2</p>	<p><b>Error Analysis</b> The problem clearly includes a graph of function <math>f(x)</math>, but the model disregarded this visual information entirely. Instead of analyzing the given graph to determine <math>f(x)</math>, the model made up a completely different piecewise function. Based on this imaginary function, the model erroneously concluded that <math>x = 0</math> and <math>x = 2</math> are inflection points, which contradicts what can be observed in the graph and the proper mathematical analysis. All in all, the model bypassed the actual visual information and created a solution based on an imaginary function.</p>

Figure 8: Error case of hallucination.

as possible, while the dataset still does not cover all the subjects in the International Standard Classification of Education released by the UNESCO [20].

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We have highlighted our contributions in the abstract and introduction. We have supported our claims in the related work, experiments and analysis parts.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We have included a Limitations section in the Appendix F.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Please refer to Section 5.1 and Appendix C. We have introduced the detailed process of the construction of the dataset and we will release it. We have given the generation config for test result reproduction. We have given the training hyperparameters for both the verifier and the BMMR-trained models.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: We will release the BMMR dataset for both evaluation and training. As mentioned in Appendix C, codes such as the inference engine and training frameworks we used to perform the experiments are publicly available and we will provide the code to run the evaluation, including how to extract the answers.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We have given the generation config for evaluation and hyperparameters for training. Please refer to Section 5.1 and Appendix 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: Due to the high cost of the API calls to the proprietary models and the limitation of computing budgets, we cannot perform the experiments many times for error bars. To make the results more reliable, we evaluate them at the temperature of 0 to minimize randomness.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: **[Yes]**

Justification: We have indicated the type of GPUs in our paper in Section 5.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: **[Yes]**

Justification: We have confirmed that this work conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: **[Yes]**

Justification: We have discussed the broader impacts in the appendix F.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: We do not release models. The datasets are only from academic sources, such as textbooks and lectures, and we have introduced the collecting and curating processes in Appendix A. The intended usage of the dataset is specified in Appendix F.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The models are used under their corresponding licenses, including The MIT License, Apache-2.0, OpenAI Terms of Use, Gemini Terms of Use, and Meta Llama License. We have cited the works that created these assets. The data source of BMMR is the public domain.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: **[Yes]**

Justification: We have created the BMMR dataset and have given comprehensive introductions on how it is created and curated. The BMMR is distributed under CC-BY-4.0 license.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: **[NA]**

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: **[NA]**

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.