AstroMMBench: A Benchmark for Evaluating Multimodal Large Language Models Capabilities in Astronomy

Anonymous ACL submission

Abstract

Astronomical image interpretation presents a significant challenge for applying multimodal large language models (MLLMs) to specialized scientific tasks. Existing benchmarks focus on general multimodal capabilities but fail to capture the complexity of astronomical data. To bridge this gap, we introduce **AstroMMBench**, the first comprehensive benchmark designed to evaluate MLLMs in astronomical image understanding. AstroMMBench comprises 621 multiple-choice questions across six astrophysical subfields, curated and reviewed by 15 domain experts for quality and relevance. We conducted an extensive evaluation of 25 diverse MLLMs, including 22 open-source and 3 closed-source models, using AstroMMBench. The results show that Ovis2-34B achieved the highest overall accuracy (70.5%), demonstrating leading capabilities even compared to strong closed-source models. Performance showed variations across the six astrophysical subfields, proving particularly challenging in domains like cosmology and high-energy astrophysics, while models performed relatively better in others, such as instrumentation and solar astrophysics. These findings underscore the vital role of domain-specific benchmarks like AstroMMBench in critically evaluating MLLM performance and guiding their targeted development for scientific applications. AstroMMBench provides a foundational resource and a dynamic tool to catalyze advancements at the intersection of AI and astronomy.

1 Introduction

Astronomy is a field that relies heavily on observation. The analysis and interpretation of telescope-collected image data is a crucial method for astronomers to understand the universe. The increasing volume and complexity of astronomical data, driven by advanced telescopic technologies, pose increasing challenges for efficient and accurate data interpretation. Consequently, the quest for more advanced image analysis technologies has consistently been a significant direction in astronomical research.

Recently, as large language models (LLMs) (Devlin, 2018; Brown et al., 2020; Zeng et al., 2022; Bai et al., 2023a; Grattafiori et al., 2024; DeepSeek-AI, 2024) and large visual models (LVMs) (Ramesh et al., 2021; Zhang et al., 2022; Kirillov et al., 2023; Shen et al., 2023; Zhai et al., 2023; Fini et al., 2024; Chen et al., 2024b) have been advancing, researchers have increasingly acknowledged the synergy effects that exist between these two types of models. This recognition has accelerated the formation and advancement of multimodal large language models (MLLMs) (Achiam et al., 2023; Wang et al., 2023; Yao et al., 2024; Tong et al., 2024; Abdin et al., 2024; Liu et al., 2024; GLM et al., 2024; Bai et al., 2025; Wu et al., 2024; Team, 2025; Team et al., 2025; Zhu et al., 2025; Dong et al., 2025). MLLMs combine the advanced natural language processing capabilities of LLMs with the visual comprehension strengths of LVMs, enabling them to possess both extensive world knowledge and advanced abilities in solving general visual tasks and complex reasoning (Huang et al., 2024a). This combination of capabilities allows MLLMs to perform deeper and more detailed analysis of text and images, showing significant potential and value across various domains, such as healthcare (Guo and Wan, 2024), autonomous driving (Cui et al., 2023), and art (Ko et al., 2022).

It is foreseeable that MLLMs, with their powerful visual perception and understanding capabilities, will have enormous potential to assist astronomers in analyzing astronomical observation images. However, evaluating the performance of MLLMs in astronomical image understanding remains challenging. Although there are many multimodal benchmarks (Yue et al., 2024; Chen et al., 2024a; Wang et al., 2024; Masry et al., 2022; Li et al., 2023; Huang et al., 2024b; Lu et al.,

2024a) available for evaluating the performance of MLLMs, they focus either on the models' comprehensive capabilities or on specific nonastronomical tasks. These benchmarks lack the domain specificity needed to assess a model's ability to handle tasks that require specialized knowledge of astrophysical processes.

To address this gap, we introduce **AstroMMBench**, the first benchmark specifically designed to evaluate the performance of MLLMs in astronomy. AstroMMBench includes 621 multiple-choice questions generated through an automated pipeline using images of papers on arxiv[1], which have been rigorously vetted by 15 domain experts. These questions span six major subfields, from Galactic Astrophysics to Cosmology, providing a comprehensive framework for assessing capabilities in the field of MLLMs astronomy.

We evaluated 25 diverse MLLMs, comprising 22 publicly available open-source and 3 powerful closed-source models, using the VLMEvalKit framework and found significant performance differences across models and different subfields of astronomy. The results indicate that Ovis2-34B (Lu et al., 2024b) performs particularly well in various astrophysical tasks, achieving an overall score of 70.53%. Notably, its performance surpassed that of leading closed-source models like ChatGPT-4o (Hurst et al., 2024) (69.07%) and Doubao-1.5-vision-pro (68.12%), demonstrating the strong competitiveness of open-source solutions in professional field tasks. These findings underscore the importance of domain-specific benchmarks for advancing MLLMs in scientific research. We hope that AstroMMBench will become a key tool at the intersection of astronomy and artificial intelligence, promoting the development of models with better astronomical image understanding capabilities.

2 Related Work

Evaluating the diverse capabilities of MLLMs necessitates comprehensive benchmarks. Existing general multimodal benchmarks (Yu et al., 2024; Chen et al., 2024a; Yue et al., 2024; Ying et al., 2024; Song et al., 2024; Li et al., 2024, 2023; Fu et al., 2024), primarily focus on everyday scenarios and common knowledge. They cover tasks such as image captioning, visual question answering (VQA), object perception, and complex reasoning across more than 20 skill dimensions. While these benchmarks are essential for measuring fundamental multimodal abilities and general world knowledge, they typically rely on common image types and scenarios, thus lacking the specialized content and nuanced understanding required for performance evaluation in fine-grained scientific domains.

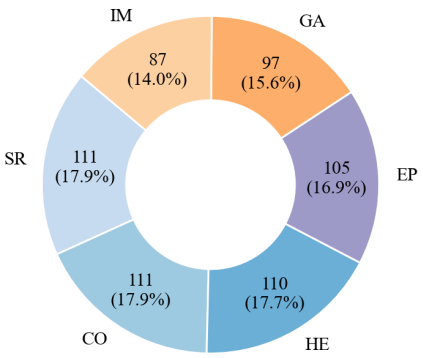


Figure 1: Distribution of questions across astronomy subfields in AstroMMBench.

To address the limitations of general-purpose benchmarks in terms of domain-specific knowledge coverage and task complexity, researchers have developed an increasing number of specialized evaluation suites. In the domain of mathematical and logical reasoning, benchmarks such as MathVista (Lu et al., 2024a), MathVerse (Zhang et al., 2024), and We-Math (Qiao et al., 2024) have been introduced to assess models' capabilities in understanding and solving visually presented mathematical problems. For chart and diagram understanding, datasets like ChartQA (Masry et al., 2022), ChartX (Xia et al., 2024), and CharXiv (Wang et al., 2024) focus on evaluating model performance in chart recognition and complex reasoning tasks. Significant progress has also been made in multimodal evaluation for the medical domain, where benchmarks such as MedXpertQA (Zuo et al., 2025) and MediConfusion (Sepehri et al., 2024) systematically examine model performance in medical image diagnosis, pathology recognition, and other critical clinical tasks. The emergence of these domain-specific benchmarks has significantly advanced the evaluation of MLLMs in complex, specialized scenarios, offering a standardized framework for fine-grained capability analysis and promoting their applicability in high-stakes, expert-driven contexts.

Despite significant progress in MLLM evaluation across general and various specific domains, a

[1]https://arxiv.org/

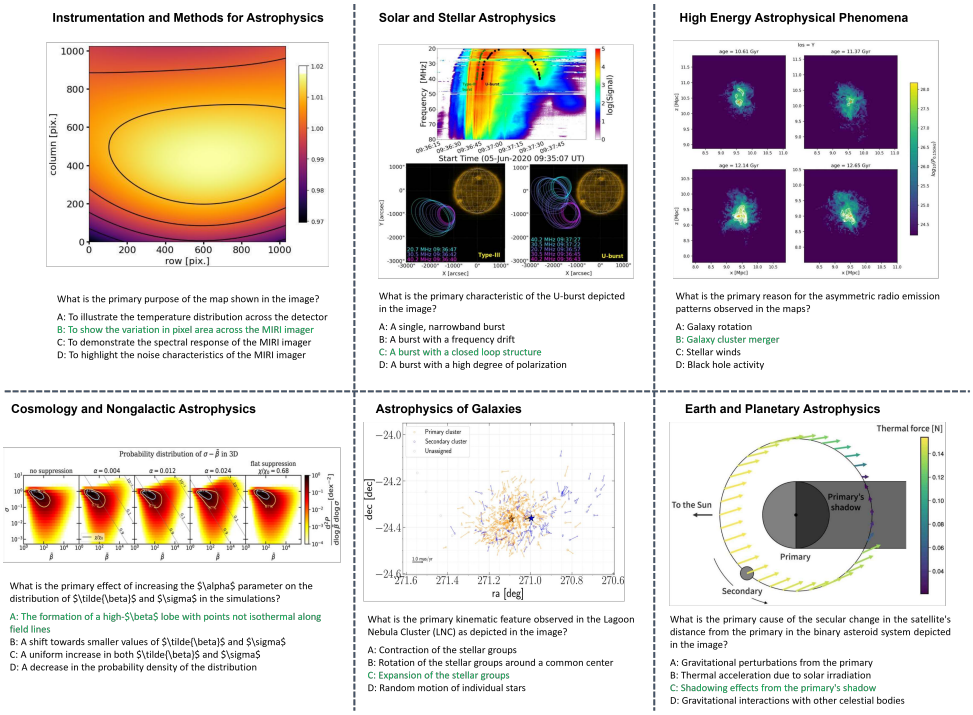


Figure 2: Examples of randomly selected questions in AstroMMBench.

dedicated multimodal evaluation benchmark specifically for astronomical images remains absent. Our work directly addresses this critical gap by introducing AstroMMBench.

3 AstroMMBench

3.1 Overview Of AstroMMBench

AstroMMBench is the first benchmark specifically designed to evaluate the performance of MLLMs in the domain of astronomical image interpretation. It comprises 621 multiple-choice questions, meticulously curated to cover six core subfields of astrophysics: Astrophysics of Galaxies (GA), Cosmology and Nongalactic Astrophysics (CO), Earth and Planetary Astrophysics (EP), High Energy Astrophysical Phenomena (HE), Instrumentation and Methods for Astrophysics (IM), and Solar and Stellar Astrophysics (SR). This structure ensures a broad yet deep assessment of MLLM performance across the discipline.

As illustrated in Figure 1, the questions are well-distributed across these subfields, with counts ranging from 87 to 111 per category, ensuring representative topical coverage. Each question, paired with an astronomical image, requires a model to select the correct answer from four options. Figure 2 showcases representative examples from AstroMMBench.

The construction of AstroMMBench involved a multi-stage process, detailed in the subsequent sections. This process began with the collection of image-text pairs from recent astrophysical literature (§3.2), followed by an automated pipeline for question generation (§3.3.1), and culminated in a rigorous, expert-led review phase to ensure the quality, relevance, and scientific accuracy of each question (§3.3.2).

3.2 Data Collection

Constructing a high-quality and domain-specific evaluation dataset that remains relevant in the rapidly evolving field of MLLMs presents challenges, particularly regarding data leakage. To address this, we need a data source that is both rich in domain-specific content and continuously updated. The arXiv repository perfectly fits this requirement, serving as a vast and dynamic archive of scientific preprints that reflect the very latest advancements across diverse subdisciplines of astrophysics. Its continuous nature allows for the potential generation of future benchmark versions utilizing data published after the training cutoff of new models, thereby mitigating the risk of data contamination.

For the initial construction of AstroMMBench, we focused exclusively on the "Astrophysics" (astro-ph) category on arXiv. We collected the TeX source files of 3,592 papers submitted between Jan-

uary 1, 2024, and July 31, 2024. From these collected papers, we extracted images along with their corresponding captions and contextual references found within the main body of the text, yielding an initial corpus of 19,299 image-text pairs.

This collection process, based on arXiv's constantly updating content, forms the foundation for a benchmark design that can be readily updated. While the specific timeframe of this initial dataset (collected in 2024) provides a snapshot of astrophysical research up to that point, the methodology allows for the creation of subsequent versions of AstroMMBench using newer arXiv data. This inherent flexibility is key to maintaining a high-quality benchmark that minimizes potential data leakage as MLLMs are continually trained on ever-larger and more recent datasets.

3.3 Automatic Pipeline

Manually creating high-quality exam questions, especially in specialized fields like astronomy, is not only time-consuming and laborious but may not be able to adapt to the ever-improving model development in a timely manner. With the rapid rise of MLLMs, it is possible to automatically generate high-quality questions from images with detailed text descriptions.

To efficiently construct a large-scale benchmark, we developed an automated pipeline for question generation and curation. Specifically, we employed LLaMA3.3-70B-Instruct and InternVL2.5-78B (Chen et al., 2024c) for question generation. This automation significantly reduced the manual effort in building a large-scale benchmark. Figure 3 shows the entire automated process, which is divided into two main stages: stage one is used to generate multiple-choice questions, and stage two filters the generated questions through multi-step review to ensure question quality.

3.3.1 Questions Autogeneration

The first stage of our pipeline focuses on automatically generating multiple-choice questions from the collected image-text pairs. Initially, we refine the textual data associated with each image to enhance its consistency and clarity. We observed that captions and contextual references extracted directly from research papers often contain:

- **Information Redundancy**: Descriptions that include details irrelevant to the specific image or residual LaTeX formatting;
- **Style Inconsistency**: Variations in writing styles across different authors, which impact the standardization of the input text.

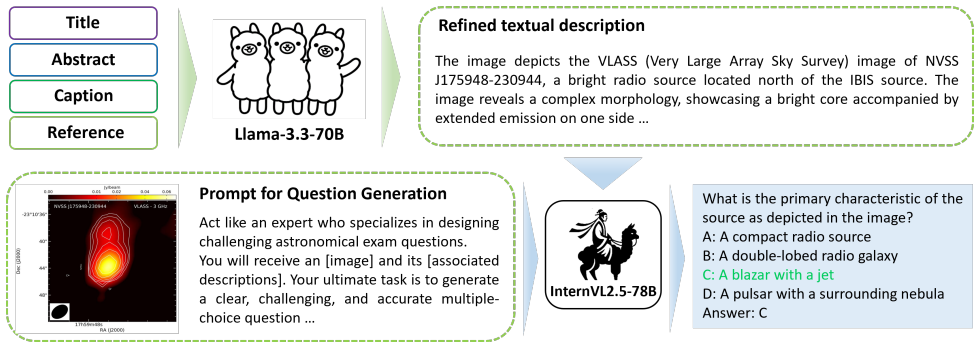
To address these challenges, we used the LLaMA3.3-70B-Instruct model to rewrite the textual data. This rewriting process was guided by a carefully designed prompt (see Appendix A.1) aimed at ensuring the accuracy and completeness of the content while effectively reducing redundancy and unifying the expression style. To provide essential background context, we supplied the LLaMA model with the paper's title and abstract, in addition to the image captions and contextual references.

Following the text refinement, the polished textual descriptions, paired with their corresponding astronomical images, were input into the InternVL2.5-78B model to generate the multiple-choice questions. To ensure the generated questions were clear, challenging, and scientifically accurate, we utilized an implicit thought chain prompt (see Appendix A.2). This prompt was designed to guide the model through a structured reasoning process, facilitating the generation of questions that effectively probe understanding of the image and its context, along with plausible answer options.

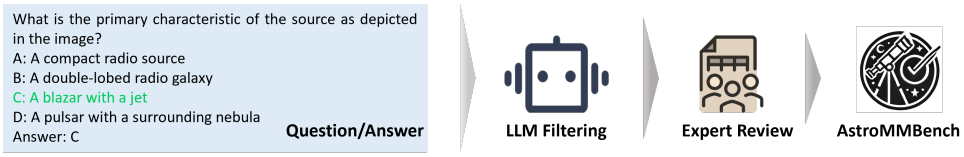
3.3.2 Questions Review

After generating preliminary questions, we introduced a multi-stage review process to ensure the quality and academic rigor of the generated questions. This is the core step of the second stage, which aims to ensure that the final retained questions are highly accurate and challenging through multi-model evaluation and expert review.

First, to ensure that the generated questions can effectively assess the respondents' ability to analyze astronomical images, we used five LLMs for an initial filtering step, including InternLM2.5-7B-Chat (Cai et al., 2024), LLaMA-3.1-8B-Instruct, Yi-1.5-34B-Chat (Young et al., 2024), Qwen2.5-32B (Team, 2024), and InternLM2-20B-Chat (Cai et al., 2024). Each model answered each question five times, and a question was considered correctly answered by a model if at least three out of five responses were correct. To eliminate questions that could be answered primarily through linguistic reasoning without requiring visual understanding, we discarded questions that were correctly answered by at least two models. As a result of this filtering process, the initial pool of 19,299 generated questions was reduced to 9,677, retaining those more



(a) Multiple Choice Questions Autogeneration



(b) Question/Answer Review

Figure 3: Automated pipeline for multiple-choice question generation and review. The pipeline is divided into two stages. (a) The initial stage involves the autogeneration of multiple-choice questions. Llama-3.3-70B-Instruct refines textual descriptions associated with astronomical images, while InternVL2.5-78B generates corresponding questions. (b) The second stage is the review process, where the generated questions undergo filtering by large language models (LLMs) and expert evaluation to ensure the quality, correctness, and relevance of both the questions and answers before their inclusion in the final benchmark.

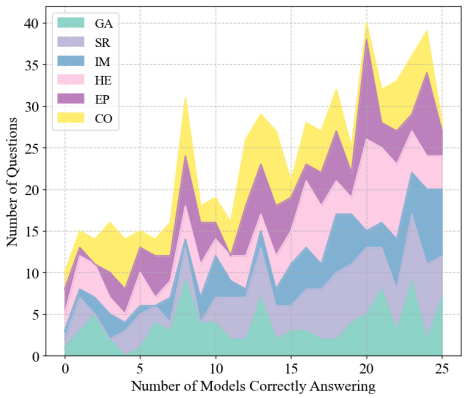


Figure 4: Distribution of question difficulty in AstroMMBench, based on the number of evaluated models that correctly answered each question. The x-axis indicates the "Number of Models Correctly Answering" a question (0-25), and the y-axis shows the count of questions at each correctness level, broken down by subfield.

likely to require visual input for accurate interpretation.

To ensure the accuracy, relevance, and rigor of the dataset, a panel of 15 astronomy experts—each holding at least a master's degree in astronomy or a related field—conducted a thorough review of 1,800 randomly selected questions from the remaining 9,677 questions. Each question was independently evaluated by an expert within the corresponding subfield. Based on criteria including image-question alignment, contextual completeness, answer accuracy and uniqueness, and the necessity of domain-specific knowledge, a total of 621 high-quality questions were retained. Furthermore, to mitigate potential biases in model responses and ensure fair evaluation, the correct answer options for these 621 questions were reassigned to be uniformly distributed across options A, B, C, and D.

3.4 Difficulty Distribution

To further characterize AstroMMBench as an evaluation benchmark, we analyzed the distribution of question difficulty. The difficulty of each question in AstroMMBench, as revealed by the evaluation results presented in Section 4, is characterized by the number of the 25 evaluated models that were able to correctly answer it. Figure 4 illustrates this difficulty distribution based on the performance of the evaluated models. The x-axis represents the number of models that correctly answered a given question, ranging from 0 for the most challenging questions (answered correctly by no models) to 25 for the easiest (answered correctly by all models). The y-axis shows the number of questions at each level of correctness, with stacked areas representing the contribution of each astrophysical subfield.

The overall trend in the figure indicates that the number of questions gradually decreases as their difficulty increases. This suggests that the benchmark deliberately avoids an overrepresentation of extremely difficult questions that current models struggle to solve. Most questions fall within the medium difficulty range, which is effective for differentiating model capabilities. Additionally, the stacked area chart shows that questions from all subfields contribute to the overall difficulty spectrum, although their proportions vary across difficulty levels. This distribution demonstrates that AstroMMBench offers a challenging yet balanced evaluation framework.

4 Experiments

4.1 Baselines

MLLMs To evaluate the performance of current MLLMs in the domain of astronomy, we selected a diverse set of 25 models, comprising 22 publicly available open-source models and 3 carefully selected powerful closed-source models. The complete list of evaluated models, along with their overall and subfield-specific performance on AstroMMBench, is presented in Table 1.

Evaluation For the evaluation process, we utilized VLMEvalKit (Duan et al., 2024), a widely used open-source evaluation framework specifically designed for MLLMs, which provides standardized protocols and metrics. As AstroMMBench is composed exclusively of multiple-choice questions with a single correct answer, the primary evaluation metric used is accuracy (proportion of correctly answered questions). A model's response is considered correct only if the extracted answer option precisely matches the predefined correct answer. To accurately extract the chosen answer option from the potentially verbose text outputs of the evaluated MLLMs, we employed DeepSeek-V3 (DeepSeek-AI, 2024) to parse model responses and identify the intended answer option (A, B, C, or D), thereby mitigating issues arising from simple pattern matching in free-form generation. All experiments were conducted on hardware equipped with eight NVIDIA A100 GPUs.

4.2 Main Results on AstroMMBench

4.2.1 Overall Performance

Table 1 summarizes the performance of 25 MLLMs evaluated on AstroMMBench, sorted by overall accuracy. The OpenCompass scores are drawn from the OpenCompass multimodal model leaderboard[2], which reflects model capabilities across general-purpose tasks. The results demonstrate substantial variation in performance across models. Among them, the open-source Ovis2-34B model achieved the highest overall accuracy (70.53%), outperforming all other models on this benchmark. It is followed by ChatGPT-4o (69.07%) and Doubao-1.5-Vision-Pro (68.12%), highlighting the competitiveness of state-of-the-art commercial MLLMs. Remarkably, Ovis2-34B's leading performance over these proprietary models underscores the rapid advancement and potential of open-source MLLMs for domain-specific tasks.

The scores for other models span a wide range, with Gemma3-4B and InternVL3-1B achieving the lowest overall accuracies of 42.51% and 45.73%, respectively. Although all models outperform the 25% accuracy expected from random guessing on a four-choice multiple-choice task, their overall performance remains limited. This highlights the difficulty of the AstroMMBench benchmark and reveals significant room for improvement in current MLLMs' ability to process astronomical images.

4.2.2 Relationship with General Capabilities

As illustrated in Figure 5, there is a clear positive correlation between the models' general performance (OpenCompass score) and their astronomical domain performance (AstroMMBench overall score). The red dashed line in the figure represents the linear regression fit between the two, revealing a linear correlation. To quantify the strength of this linear relationship, we calculated the Pearson correlation coefficient, obtaining $(r = 0.82)$, which demonstrates a significant positive correlation. The calculation method for the Pearson correlation coefficient (r) is provided in equation 1. This suggests that models performing well on general tasks also tend to excel in astrophysical tasks, validating the robustness and scientific soundness of AstroMMBench.

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}} \quad (1)$$

However, this correlation is not without exceptions. For example, Ovis2-34B outperforms models with higher general scores like ChatGPT-4o, Qwen2.5-VL-72B, and InternVL3-38B on AstroMMBench. This anomaly suggests that while

[2] https://rank.opencompass.org.cn/leaderboard-multimodal/?m=REALTIME

Table 1: Performance of 3 closed-source and 22 open-source models on the AstroMMBench dataset across subdomains of astrophysics. The best-performing model in each category is in bold.

Model	Overall (621)	GA (97)	CO (111)	EP (105)	HE (110)	IM (87)	SR (111)	OpenCompass
Closed-source Models								
ChatGPT-4o (Hurst et al., 2024)	69.07	69.07	**67.57**	67.62	70.91	68.97	71.17	**47.49**
Doubao-1.5-vision-pro	68.12	**70.10**	**67.57**	64.76	**72.73**	67.82	65.77	–
QwenVLMax (Bai et al., 2023b)	66.83	58.76	58.56	69.52	65.45	78.16	72.07	–
Open-source Models								
Ovis2-34B (Lu et al., 2024b)	**70.53**	68.04	**67.57**	68.57	**72.73**	78.16	69.37	42.82
InternVL3-38B (Zhu et al., 2025)	67.63	68.04	53.15	61.90	70.91	**80.46**	**73.87**	45.55
Qwen2.5-VL-72B (Bai et al., 2025)	67.47	59.79	57.66	**72.38**	69.09	74.71	72.07	48.25
Ovis2-16B (Lu et al., 2024b)	67.31	63.92	63.06	61.90	70.91	75.86	69.37	39.60
Qwen2.5-VL-32B (Bai et al., 2025)	64.25	57.73	60.36	60.00	68.18	68.97	70.27	–
InternVL3-78B (Zhu et al., 2025)	64.25	63.92	51.35	60.95	65.45	73.56	72.07	45.96
InternVL3-14B (Zhu et al., 2025)	63.77	61.86	53.15	62.86	61.82	73.56	71.17	40.72
SAIL-VL-1.6-8B (Dong et al., 2025)	62.32	56.70	58.56	63.81	62.73	68.97	63.96	37.92
InternVL3-8B (Zhu et al., 2025)	61.03	59.79	52.25	61.90	60.91	66.67	65.77	37.40
InternVL3-9B (Zhu et al., 2025)	60.55	60.82	49.55	55.24	62.73	65.52	70.27	–
DeepSeek_VL2 (Wu et al., 2024)	59.90	57.73	53.15	61.90	61.82	68.97	57.66	38.70
Qwen2.5-VL-3B (Bai et al., 2025)	58.94	52.58	59.46	56.19	58.18	65.52	62.16	38.33
MiniCPM-o-2.6 (Yao et al., 2024)	57.97	54.64	51.35	50.48	62.73	67.82	62.16	34.67
Qwen2.5-VL-7B (Bai et al., 2025)	57.33	52.58	56.76	53.33	54.55	67.82	60.36	43.21
Kimi-VL-A3B-Instruct (Team et al., 2025)	56.68	50.52	54.95	54.29	55.45	68.97	57.66	37.00
LLaVA_Onevision_72B (Li et al., 2024)	55.39	52.58	54.95	53.33	50.91	63.22	58.56	39.05
Gemma3-12B (Team, 2025)	52.82	49.48	51.35	60.95	45.45	50.57	58.56	34.15
InternVL3-2B (Zhu et al., 2025)	51.69	53.61	47.75	46.67	50.00	55.17	57.66	30.96
Kimi-VL-A3B-Thinking (Team et al., 2025)	50.08	50.52	45.05	43.81	47.27	57.47	57.66	–
GLM-4v-9B (GLM et al., 2024)	49.76	46.39	39.64	48.57	52.73	59.77	53.15	37.85
InternVL3-1B (Zhu et al., 2025)	45.73	47.42	38.74	39.05	42.73	57.47	51.35	24.39
Gemma3-4B (Team, 2025)	42.51	42.27	34.23	39.05	41.82	48.28	50.45	32.21

general multimodal capabilities can predict success in specialized fields, some models may face difficulties when confronted with domain-specific challenges in astrophysics. It also underscores the unique challenges posed by AstroMMBench, where models must handle domain-specific questions that may not be adequately captured by general-purpose multimodal benchmarks.

4.2.3 Analysis by Subfield

AstroMMBench encompasses six major subfields of astrophysics. A detailed analysis of model performance within these distinct domains allows for a deeper understanding of their strengths and potential limitations when tackling different types of astronomical tasks. Figure 6 presents a radar chart offering a visual overview of the performance profiles for selected representative models across the six subfields. Examples of these models' responses within each subfield and varying difficulty questions are provided in Appendix B.

Our analysis indicates that performance disparities across different subfields reflect the varying capabilities required by the questions in each category. Specifically, questions in the IM and SR subfields primarily demand skills related to interpreting standard astronomical plots (e.g., time series, relationships between physical quantities) and recognizing common astronomical objects or instrument components. These tasks may align well with the graph understanding and object recognition capabilities models acquire during general domain training, thus resulting in generally higher scores in these subfields. Conversely, questions in the CO and HE subfields typically require a deeper understanding of abstract theoretical concepts, interpretation of highly specialized or unconventional visualizations (e.g., statistical maps of cosmic structures, signatures of particle interactions), and complex multi-step reasoning based on fragmented information. These capabilities may be less developed or consistently present in current general-purpose MLLMs. The GA and EP subfields, covering a wide range of question types from galaxy morphological classification to interpreting planetary atmospheric data or orbital dynamics plots, require a mix of these abilities and exhibit intermediate difficulty.

The radar chart in Figure 6 visualizes the performance profiles of various models across astrophysical subfields. Top performers like Ovis2-34B and ChatGPT-4o display balanced, consistent polygons,

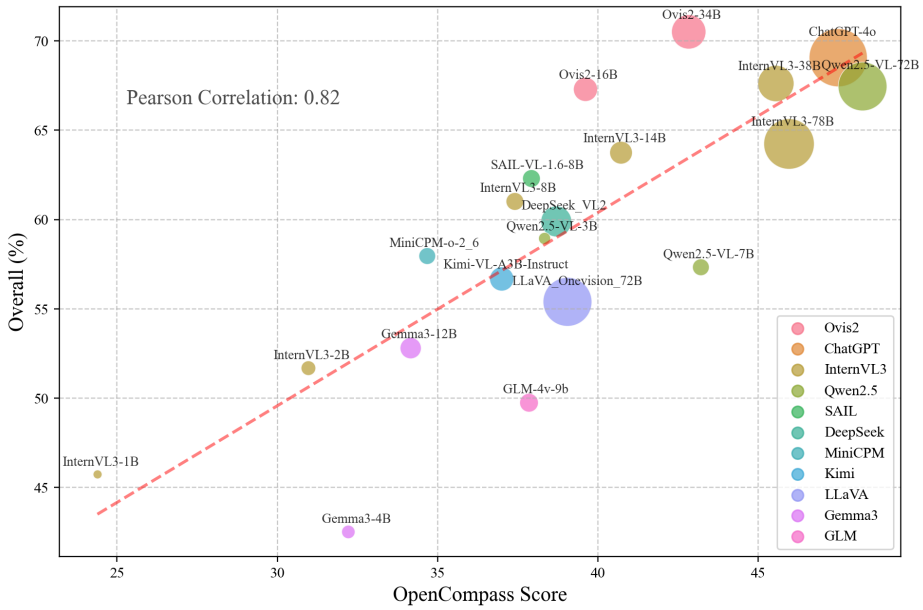


Figure 5: Relationship between general multimodal performance (OpenCompass score) and specialized astronomical image interpretation performance (AstroMMBench overall accuracy) for 22 MLLMs. Point size represents model scale (parameter count)

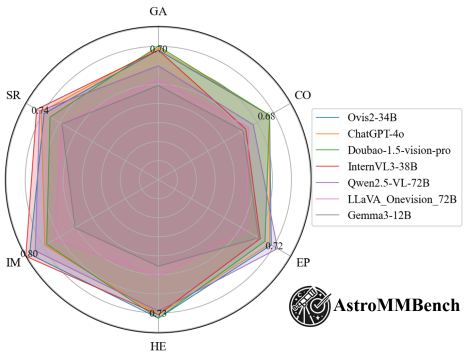


Figure 6: Comparison of model performance across six astrophysical subfields in AstroMMBench.

underscoring their robustness and versatility in astronomy. In contrast, InternVL3-38B, while achieving leading scores in the IM and SR fields, shows a notable decrease in performance in the CO field. This suggests that its ability to interpret standard astronomical plots might be stronger than its capacity to handle the more abstract concepts and specialized imagery common in cosmology. Other models also showcase their specific characteristics, such as Doubao-1.5-vision-pro's prominent strength in the HE field and Qwen2.5-VL-72B's leading performance in the EP field. These variations highlight that different models may possess specific proficiencies aligned with particular astrophysical domains, likely stemming from differences in their training or architecture.

5 Conclusion

In this paper, we introduce AstroMMBench, the first benchmark tailored to assess MLLMs in astronomy. It features 621 multiple-choice questions spanning six key astrophysics subfields, automatically generated and expert-reviewed for accuracy and relevance.

Using the VLMEvalKit framework, we evaluated 25 MLLMs and observed significant performance differences. The open-source Ovis2-34B outperformed top closed-source models like ChatGPT-4o and Doubao-1.5-vision-pro with a 70.53% score, emphasizing the promise of open models in scientific domains. We found a strong positive correlation between general MLLM performance and AstroMMBench scores, yet exceptions demonstrate the critical need for domain-specific evaluation to truly assess specialized proficiency. Furthermore, performance varied across astrophysical subfields, with domains like Cosmology and Nongalactic Astrophysics and High Energy Astrophysics proving generally more challenging than Instrumentation and Methods for Astrophysics and Solar and Stellar Astrophysics, reflecting the diverse demands on MLLM capabilities. We hope that AstroMMBench can become a continuously evolving platform to support the evaluation and promotion of the next generation of MLLMs in astronomy.

Limitations

Our study provides a benchmark and framework for the performance of multimodal large language models in the astronomy domain, but we acknowledge that there are several limitations to our study that may require further exploration:

Limited benchmark size and task diversity The current benchmark size of 621 questions, while substantial for a first benchmark of this nature, is relatively limited compared to the vastness and complexity of astronomical phenomena and tasks. Furthermore, the task format is currently restricted to multiple-choice Visual Question Answering. Incorporating more diverse question types, such as open-ended questions, multi-step reasoning tasks, or predictive analysis challenges, would provide a more comprehensive assessment of MLLMs' advanced capabilities needed for complex scientific analysis beyond direct VQA.

Challenges in automated question generation and curation Although an automated pipeline is employed for initial question generation from scientific literature, the quality of the generated questions can be inconsistent. This often results in a proportion of low-quality or irrelevant questions that do not adequately test specialized astronomical knowledge. Consequently, ensuring the scientific rigor and quality of the final benchmark set heavily relies on a costly and time-consuming manual expert review process. This reliance on manual curation limits the scalability and efficiency of expanding the benchmark size and providing frequent updates with new data, presenting a key challenge for maintaining a dynamic benchmark.

In our future work, we will focus on addressing these shortcomings to overall improve the quality and scalability of AstroMMBench while ensuring that it becomes a comprehensive and evolving benchmark for evaluating MLLMs in the specialized field of astronomy.

Ethics Statement

Copyright and License Regarding the data used in AstroMMBench, all images and associated textual content are sourced from publicly available preprints on arXiv. We ensure compliance with copyright regulations by strictly adhering to the terms of use for arXiv data, which permits the re-use and distribution of content under specific licenses (typically Creative Commons licenses as specified by the authors). We maintain adherence to the established legal and ethical standards for using publicly available scientific literature.

We are committed to making AstroMMBench openly accessible to the research community to facilitate further research and evaluation of MLLMs in astronomy. AstroMMBench will be released under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

References

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *ArXiv*, abs/2309.16609.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tongxi Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. 2023. A survey on multimodal large language models for autonomous driving. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 958–979.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. 2025. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. 2024. Multimodal autoregressive pre-training of large vision encoders.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

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

Yuhang Guo and Zhiyu Wan. 2024. Performance evaluation of multimodal large language models (llava and gpt-4-based chatgpt) in medical image classification tasks. *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 541–543.

Dawei Huang, Chuan Yan, Qing Li, and Xiaojiang Peng. 2024a. From large language models to large multimodal models: A literature review. *Applied Sciences*.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. 2024b. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003.

Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2022. Large-scale text-to-image generation models for visual artists' creative works. *Proceedings of the 28th International Conference on Intelligent User Interfaces*.

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

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023. Seed-bench-2: Benchmarking multimodal large language models. *ArXiv*, abs/2311.17092.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024b. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ArXiv*, abs/2203.10244.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092.

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. 2024. Mediconfusion: Can you trust your ai radiologist? probing the reliability of multimodal medical foundation models. *arXiv preprint arXiv:2409.15477*.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. 2023. Aligning and prompting everything all at once for universal visual perception. *ArXiv*, abs/2312.02153.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*.

Gemma Team. 2025. Gemma 3.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *ArXiv*, abs/2406.18521.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, et al. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR.

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

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, P. Zhang, Yuxiao Dong, and Jie Tang. 2022. Glm-130b: An open bilingual pre-trained model. *ArXiv*, abs/2210.02414.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun-Juan Zhu, Lionel Ming shuan Ni, and Heung yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *ArXiv*, abs/2203.03605.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Prompts

A.1 Prompt for Rewriting Descriptions

The following Promt is used in LLaMA3.3-70B-Instruct model to rewrite the context text description of the image.

SYSTEM PROMPT:
Act like an expert with extensive experience writing in the field of astrophysics.

Objective:
You will be provided with the `[CAPTION]`, and `[CONTEXT]` of an image mentioned in the paper. Meanwhile, the `[TITLE]` and `[ABSTRACT]` of the paper are also provided as background information to you. Your task is to generate a concise, precise, and scholarly description of the image, reflecting its content and relevance within the scientific discourse of the paper. Your answer will serve senior scholars, please describe it in a formal and scholarly manner.

Detailed Instructions:

1. **Content Analysis:**
   - Carefully review `[CAPTION]`, and `[CONTEXT]` of the image, determining target image and ensuring a thorough understanding of its significance and details.
   - Examine the `[TITLE]` and `[ABSTRACT]` of the paper and use those background informations if necessary.
2. **Formatting and Content Restrictions:**
   - Ensure all LaTeX formats are deleted except for mathematical formulas.
   - Ignore any content related to unknown objects in the paper, such as other formulas, images or sections, and do not summarize them.
   - If the content you are given is not related to the target image, ignore it and do not summarize it.
   - When you refer to the target image, use expressions such as "The image" or "The figure", instead of "Figure `\ref{?}`" or "Figure ?".
3. **Writing the Description:**
   - Formulate a comprehensive and scholarly description of the image using the gathered information.

Output Format:
{
"description": "The description you generated here"
}

USER PROMPT:
Please give your description based on the following informations:
`[CAPTION]`: {caption}
`[CONTEXT]`: {context}
`[TITLE]`: {title}
`[ABSTRACT]`: {abstract}

A.2 Prompt for Question Generation

The following Promt is used in the InternVL2.5-78B model to generate high-quality multi-modal multiple-choice questions in the field of astronomy.

Act like a domain expert in astronomy education, with extensive experience in designing high-level exam questions that assess advanced conceptual and analytical skills.

Objective:
You will receive an image and its associated descriptions. Your task is to generate a multiple-choice question (include one correct option and three plausible but incorrect options) at a professional level that tests the respondent's ability to analyze images and apply comprehensive astronomical knowledge.

Detailed Instructions:

1. **Image and Description Analysis:**
   - View the `[IMAGE]` provided thoroughly, noting any important subjects, features, and text, etc.
   - Read the `[IMAGE DESCRIPTIONS]` carefully to determine the relationship between the description and the image, and consider the astronomical knowledge involved.
2. **Question Design:**
   - Create a question that requires image analysis, astronomical knowledge, and in-depth analysis to solve, ensuring it does not provide hints.
3. **Create Answer Choices:**
   - Determine an answer to the question as the correct option.
   - Develop three plausible but incorrect options.
4. **Explanation of the Correct Answer:**
   - Provide a detailed explanation for why the correct answer is accurate.
   - Optionally, briefly state why each incorrect option is misleading or incorrect.

Output Format:
{
"question": "Your image-based astronomical question here",
"options": {
"A": "Option A content",
"B": "Option B content",
"C": "Option C content",
"D": "Option D content"
},
"answer": "Correct option letter"
"explanation": "Brief justification for the correct answer." }

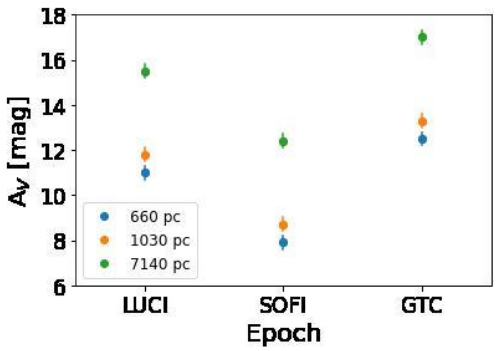
Input:
Please generate the question based on the following:
`[IMAGE DESCRIPTIONS]:{image_descriptions}`

B Model Evaluation Examples

This section presents 18 example questions on random sampling across varying levels of difficulty, with three questions selected from each subdomain.

B.1 Solar and Stellar Astrophysics (SR)

Correct responses: 24/25 models



Question: Which epoch shows the highest visual extinction (A_V) for Gaial8cjb at distance of 7140 pc?
Option:

(A) LUCI

(B) SOFI

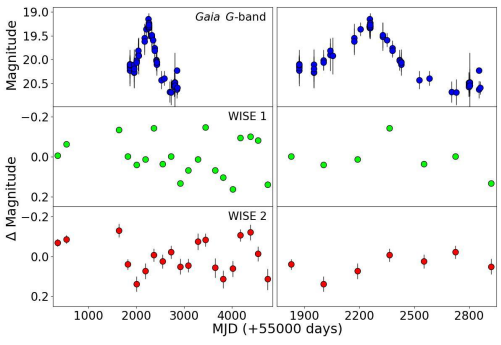
(C) GTC

(D) None of the above

Answer: C
Ovis2-34B: C
ChatGPT-4o: C
Doubao-1.5-vision-pro: C
InternVL3-38B: C
Qwen2.5-VL-72B: C
LLaVA_Onevision_72B: B
Gemma3-12B: C

Figure B1: Case 1 of AstroMMBench in SR subdomain.

Correct responses: 17/25 models



Question: Which of the following best describes the periodicity of the light curve for NGC300-59 in the Gaia G-band?
Option:

(A) Approximately 500 days

(B) Approximately 1000 days

(C) Approximately 2000 days

(D) Approximately 4000 days

Answer: B

Ovis2-34B: C

ChatGPT-4o: B

Doubao-1.5-vision-pro: C

InternVL3-38B: B

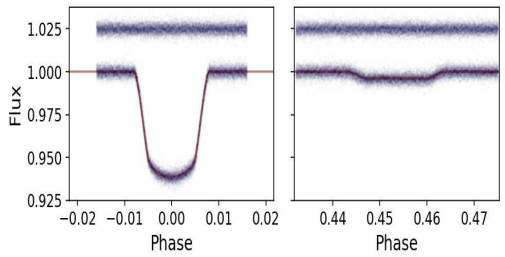
Qwen2.5-VL-72B: C

LLaVA_Onevision_72B: B

Gemma3-12B: B

Figure B2: Case 2 of AstroMMBench in SR subdomain.

Correct responses: 4/25 models



Question: What is the primary feature observed in the light curve of the EBLM J0608-59 system?
Option:

(A) A single transit event

(B) A double eclipse event

(C) A single eclipse event

(D) A continuous out-of-eclipse variation

Answer: B

Ovis2-34B: A

ChatGPT-4o: C

Doubao-1.5-vision-pro: C

InternVL3-38B: B

Qwen2.5-VL-72B: C

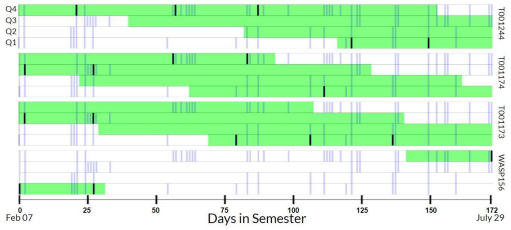
LLaVA_Onevision_72B: A

Gemma3-12B: A

Figure B3: Case 3 of AstroMMBench in SR subdomain.

B.2 Instrumentation and Methods for Astrophysics (IM)

Correct responses: 24/25 models



Question: Which of the following targets has the most limited visibility in the 2023A California Planet Search(CPs)simulation?
Option:

(A) Target 100244

(B) Target 100174

(C) Target 100173

(D) Target WASP159

Answer: D

Ovis2-34B: D

ChatGPT-4o: D

Doubao-1.5-vision-pro: D

InternVL3-38B: D

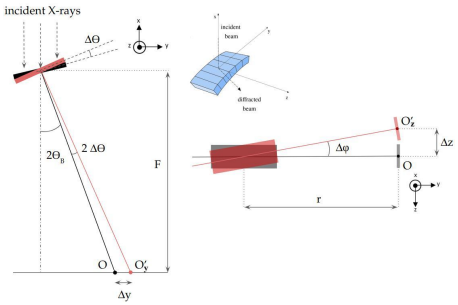
Qwen2.5-VL-72B: D

LLaVA_Onevision_72B: D

Gemma3-12B: D

Figure B4: Case 4 of AstroMMBench in IM subdomain.

Correct responses: 16/25 models



Question: What is the primary consequence of a misalignment in the Bragg angle $\theta(\Delta\theta)$ on the diffracted X-ray beam in the Laue lens?
Option:

(A) A shift in the diffracted signal along the z-axis

(B) A shift in the diffracted signal along the y-axis

(C) A change in the intensity of the diffracted beam

(D) A change in the polarization of the diffracted beam

Answer: B

Ovis2-34B: B

ChatGPT-4o: B

Doubao-1.5-vision-pro: B

InternVL3-38B: B

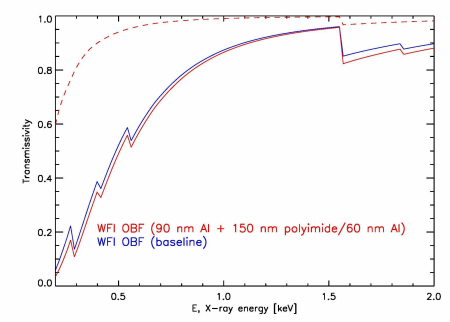
Qwen2.5-VL-72B: B

LLaVA_Onevision_72B: A

Gemma3-12B: A

Figure B5: Case 5 of AstroMMBench in IM subdomain.

Correct responses: 1/25 models



Question: Which of the following statements is true regarding the transmissivity of the WFI OBF configurations shown in the image?
Option:

(A) The baseline WFI OBF has higher transmissivity across all energy bands compared to the 90 nm Al + 150 nm polyimide/60 nm Al configuration.

(B) The 90 nm Al + 150 nm polyimide/60 nm Al configuration has higher transmissivity below 0.5 keV compared to the baseline WFI OBF.

(C) The transmissivity ratio between the two configurations is constant across the entire energy range.

(D) The 90 nm Al + 150 nm polyimide/60 nm Al configuration has higher transmissivity above 1.5 keV compared to the baseline WFI OBF.

Answer: A

Ovis2-34B: D

ChatGPT-4o: B

Doubao-1.5-vision-pro: D

InternVL3-38B: B

Qwen2.5-VL-72B: B

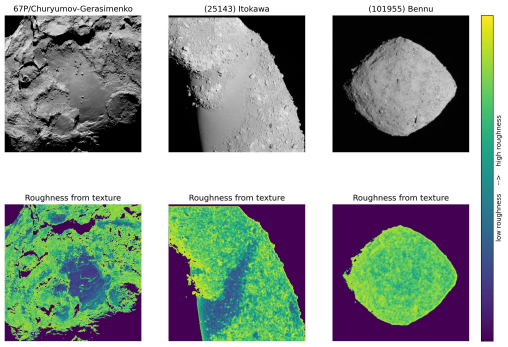
LLaVA_Onevision_72B: D

Gemma3-12B: B

Figure B6: Case 6 of AstroMMBench in IM subdomain.

B.3 Earth and Planetary Astrophysics (EP)

Correct responses: 24/25 models



Question: Which small body in the solar system exhibits the most uniform surface roughness according to the entropy of information measured in the image?
Option:

(A) 67P/Churyumov-Gerasimenko

(B) (25143) Itokawa

(C) (101955) Bennu

(D) None of the above

Answer: D

Ovis2-34B: D

ChatGPT-4o: D

Doubao-1.5-vision-pro: D

InternVL3-38B: D

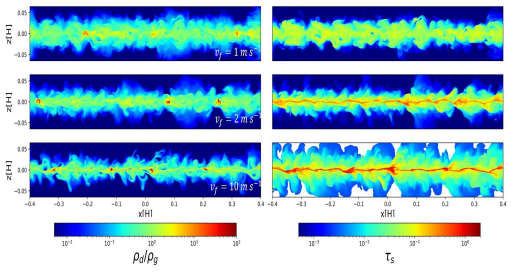
Qwen2.5-VL-72B: D

LLaVA_Onevision_72B: D

Gemma3-12B: D

Figure B7: Case 7 of AstroMMBench in EP subdomain.

Correct responses: 14/25 models



Question: What is the primary effect of increasing the fragmentation velocity v_f on the dust density distribution in the simulation?
Option:

(A) Formation of smaller dust clumps

(B) Decreased mass-averaged stopping time τ_s

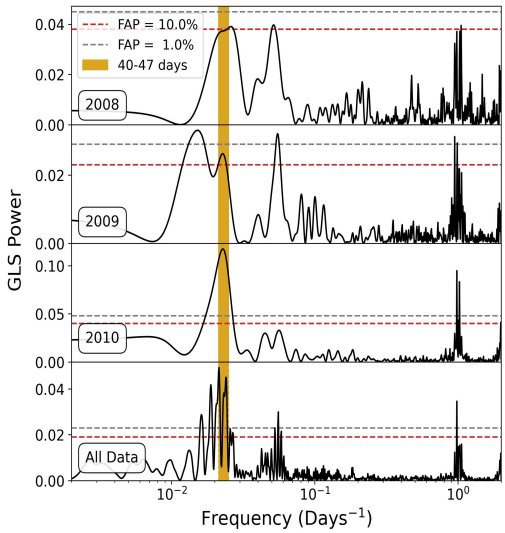
(C) Uniform distribution of dust

(D) Formation of larger dust clumps

Answer: D
Ovis2-34B: D
ChatGPT-4o: D
Doubao-1.5-vision-pro: D
InternVL3-38B: A
Qwen2.5-VL-72B: D
LLaVA_Onevision_72B: D
Gemma3-12B: A

Figure B8: Case 8 of AstroMMBench in EP subdomain.

Correct responses: 9/25 models



Question: What is the most likely cause of the strong signal at approximately 40 days in the GLS periodogram of TOI-1450A?
Option:

(A) Planetary transit

(B) Stellar rotation

(C) Binary star system

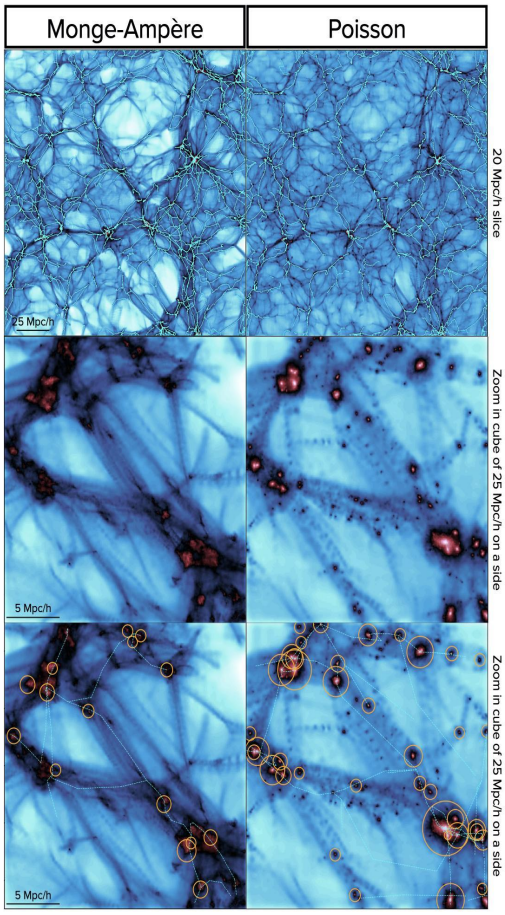
(D) Instrumental artifact

Answer: B
Ovis2-34B: A
ChatGPT-4o: A
Doubao-1.5-vision-pro: A
InternVL3-38B: A
Qwen2.5-VL-72B: B
LLaVA_Onevision_72B: A
Gemma3-12B: B

Figure B9: Case 9 of AstroMMBench in EP subdomain.

B.4 Cosmology and Nongalactic Astrophysics (CO)

Correct responses: 22/25 models



Question: Which gravity theory, as depicted in the image, exhibits a more complex and abundant network of cosmic filaments?
Option:

(A) Poisson (ΛCDM)

(B) Monge-Ampère

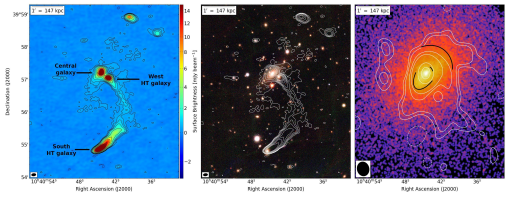
(C) Both exhibit similar complexity

(D) Neither, the complexity is indistinguishable

Answer: B
Ovis2-34B: B
ChatGPT-4o: B
Doubao-1.5-vision-pro: B
InternVL3-38B: B
Qwen2.5-VL-72B: B
LLaVA_Onevision_72B: A
Gemma3-12B: B

Figure B10: Case 10 of AstroMMBench in CO subdomain.

Correct responses: 18/25 models



Question: What is the primary feature indicated by the elongated X-ray emission in the right panel of the image?
Option:

(A) A single-armed spiral galaxy

(B) A region of high star formation activity

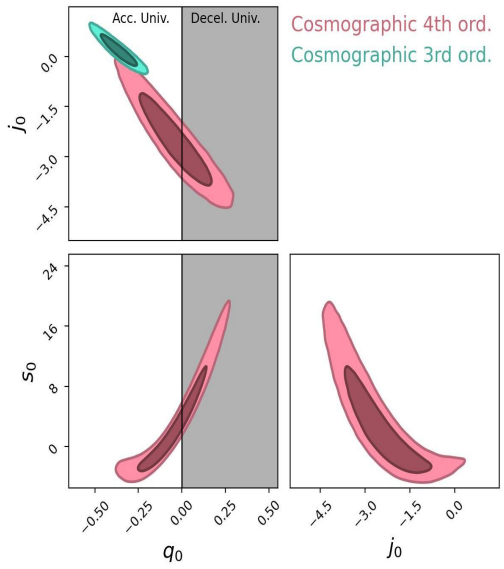
(C) A supermassive black hole

(D) A cold front in the galaxy cluster

Answer: D
Ovis2-34B: D
ChatGPT-4o: D
Doubao-1.5-vision-pro: D
InternVL3-38B: D
Qwen2.5-VL-72B: D
LLaVA_Onevision_72B: D
Gemma3-12B: D

Figure B11: Case 11 of AstroMMBench in CO subdomain.

Correct responses: 4/25 models



Question: Which cosmographic model, as depicted in the image, provides stronger evidence for an accelerating universe?
Option:

(A) Cosmographic 3rd order model

(B) Cosmographic 4th order model

(C) Both models equally

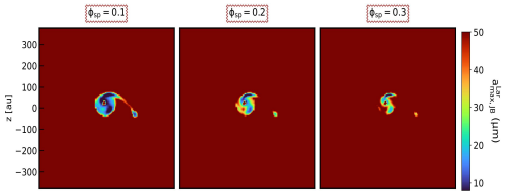
(D) Neither model

Answer: A
Ovis2-34B: A
ChatGPT-4o: A
Doubao-1.5-vision-pro: B
InternVL3-38B: B
Qwen2.5-VL-72B: B
LLaVA_Onevision_72B: B
Gemma3-12B: B

Figure B12: Case 12 of AstroMMBench in CO subdomain.

B.5 Astrophysics of Galaxies (GA)

Correct responses: 23/25 models



Question: What is the primary effect of increasing the volume filling factor of iron clusters inside dust grains (ϕ_{sp}) on the distribution of dust grain sizes within 400 au of the disk midplane?
Option:

(A) Decreased magnetic alignment of very large grains (VLGs)

(B) Increased internal alignment of dust grains

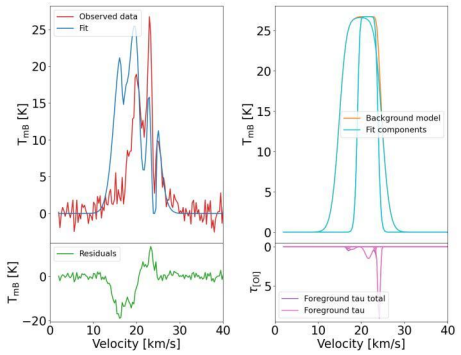
(C) Enhanced Magnetic Alignment by Radiative Torques (MRAT) alignment for micron-sized grains

(D) Reduced polarization degree within the disk scale

Answer: C
Ovis2-34B: C
ChatGPT-4o: C
Doubao-1.5-vision-pro: C
InternVL3-38B: C
Qwen2.5-VL-72B: C
LLaVA_Onevision_72B: C
Gemma3-12B: B

Figure B13: Case 13 of AstroMMBench in GA subdomain.

Correct responses: 15/25 models



Question: What is the primary reason for the significant residuals in the observed data fit shown in the image?
Option:

(A) Insufficient data points

(B) Incorrect background model

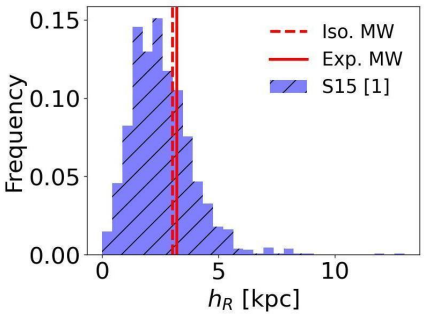
(C) Fixed foreground parameters

(D) Instrumental error

Answer: C
Ovis2-34B: C
ChatGPT-4o: C
Doubao-1.5-vision-pro: C
InternVL3-38B: C
Qwen2.5-VL-72B: C
LLaVA_Onevision_72B: C
Gemma3-12B: B

Figure B14: Case 14 of AstroMMBench in GA subdomain.

Correct responses: 1/25 models



Question: What is the primary purpose of the histogram in the image?
Option:

(A) To compare the scale length of the Milky Way with other galaxies

(B) To determine the frequency of galaxies with specific scale lengths

(C) To illustrate the distribution of galaxy types in the sample

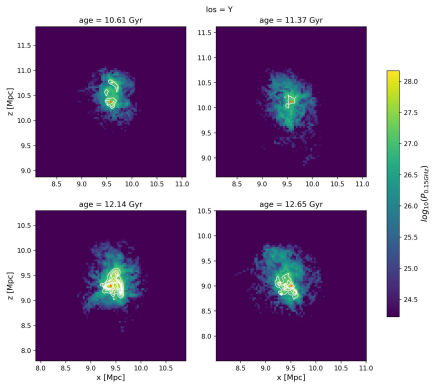
(D) To show the relationship between scale length and galaxy mass

Answer: A
Ovis2-34B: B
ChatGPT-4o: B
Doubao-1.5-vision-pro: B
InternVL3-38B: B
Qwen2.5-VL-72B: B
LLaVA_Onevision_72B: B
Gemma3-12B: A

Figure B15: Case 15 of AstroMMBench in GA subdomain.

B.6 High Energy Astrophysical Phenomena (HE)

Correct responses: 22/25 models



Question: What is the primary reason for the asymmetric radio emission patterns observed in the maps?
Option:

(A) Galaxy rotation

(B) Galaxy cluster merger

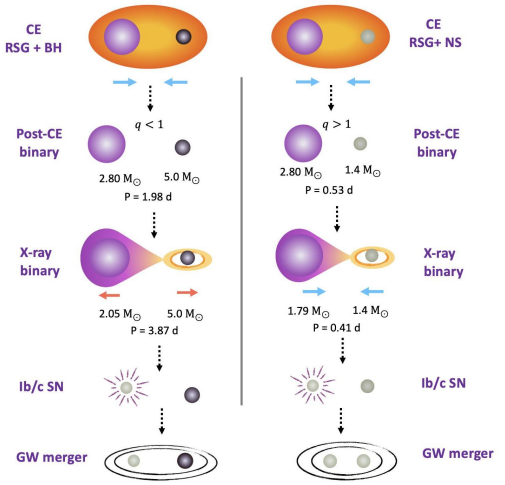
(C) Stellar winds

(D) Black hole activity

Answer: B
Ovis2-34B: B
ChatGPT-4o: B
Doubao-1.5-vision-pro: B
InternVL3-38B: B
Qwen2.5-VL-72B: B
LLaVA_Onevision_72B: B
Gemma3-12B: D

Figure B16: Case 16 of AstroMMBench in HE subdomain.

Correct responses: 9/25 models



Question: What is the final orbital period of the post-CE binary with a black hole (BH) companion after the mass transfer (MT) phase?
Option:

(A) 0.41 d

(B) 1.98 d

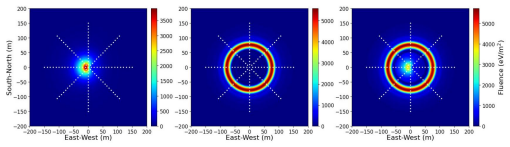
(C) 0.53 d

(D) 3.87 d

Answer: D
Ovis2-34B: D
ChatGPT-4o: D
Doubao-1.5-vision-pro: D
InternVL3-38B: D
Qwen2.5-VL-72B: B
LLaVA_Onevision_72B: A
Gemma3-12B: B

Figure B17: Case 17 of AstroMMBench in HE subdomain.

Correct responses: 1/25 models



Question: What is the primary reason for the distinct ring-like pattern observed in the middle panel of the image?
Option:

(A) Refraction of radio waves through the ice medium

(B) Reflection of radio waves off the ice surface

(C) Scattering of radio waves by ice crystals

(D) Interference between in-air and in-ice radio waves

Answer: A
Ovis2-34B: D
ChatGPT-4o: A
Doubao-1.5-vision-pro: D
InternVL3-38B: D
Qwen2.5-VL-72B: D
LLaVA_Onevision_72B: B
Gemma3-12B: D

Figure B18: Case 18 of AstroMMBench in HE subdomain.