# MMSci: A Dataset for Graduate-Level Multi-Discipline Multimodal Scientific Understanding

**Anonymous ACL submission**

## Abstract

Scientific figure interpretation is crucial for AI scientific assistants built on Large Vision Language Models, yet current datasets mainly cover restricted scientific domains, and limited figure complexity (like charts). We address this gap with a comprehensive dataset from peer-reviewed Nature Communications articles spanning 72 scientific fields, featuring complex visualizations that require graduate-level expertise to interpret. Evaluation of 19 proprietary and open-source models on figure captioning and multiple-choice tasks, alongside human expert annotation, revealed significant performance gaps. Beyond benchmarking, our dataset enables effective large-scale training. Fine-tuning Qwen2-VL-2B with our data outperformed GPT-4o and human experts in multiple-choice tasks, while continuous pre-training on interleaved article-figure data enhanced downstream performance in materials science. The dataset has been made anonymously available to support further research.[1]

## 1 Introduction

Recent advancements in Large Vision Language Models (LVLMs) (Li et al., 2023; Zhu et al., 2023; Liu et al., 2024; Chen et al., 2024c; Bai et al., 2023b; Achiam et al., 2023; Team et al., 2023; Anthropic, 2024a; Wang et al., 2024a), have demonstrated remarkable capabilities in solving problems involving visual context. The growing capabilities of LVLMs make them promising as AI-driven scientific assistants capable of solving problems and assisting in research in various *science domains*. A critical aspect of this assistance is interpreting the figures in research articles, which often contain rich, compressed, and complex information, requiring domain-specific expertise to understand.

Current LVLM evaluations focus mainly on bar charts (Kahou et al., 2017; Masry et al., 2022; Roberts et al., 2024; Wang et al., 2024b) and simple figures in limited science domains (Kembhavi et al., 2017; Lu et al., 2022; Yue et al., 2023, 2024; Li et al., 2024; Chen et al., 2024a). Scientific articles, however, contain diverse visualizations, like microscopy images, molecular structures, astronomical images, phylogenetic trees, and various specialized plots, requiring graduate-level domain expertise for proper interpretation.

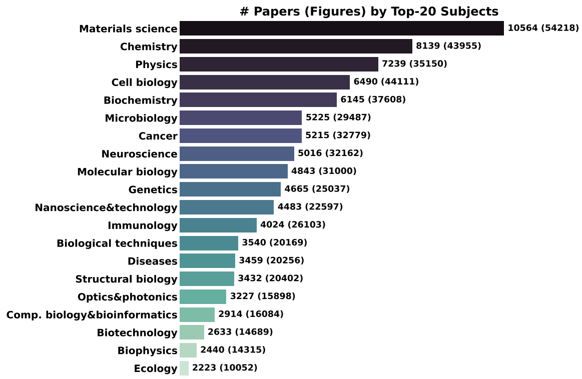



Figure 1: **Top 20 out of 72 science subjects with most articles in our dataset MMSci.** The corresponding numbers of papers and figures (in brackets) are shown.

To bridge this gap, we introduce MMSci, a comprehensive multimodal dataset curated from open-access *Nature Communications* articles[2] under CC BY 4.0 license[3]. The dataset encompasses 72 scientific disciplines, containing 131k articles and 742k figures across diverse visualization types, with discipline distribution shown in Figure 1. To evaluate LVLMs' understanding of these complex scientific figures requiring graduate-level expertise, we developed benchmark tasks for figure captioning and multiple-choice questions across different settings.

Our evaluation revealed wide performance gaps among LVLMs. For multiple-choice questions, many open-source models performed at random-guess levels, though Qwen2-VL-7B (Wang et al.,

[1]The links to data and code are provided in Appendix A.1.1

[2]https://www.nature.com/ncomms/

[3]https://www.nature.com/ncomms/open-access

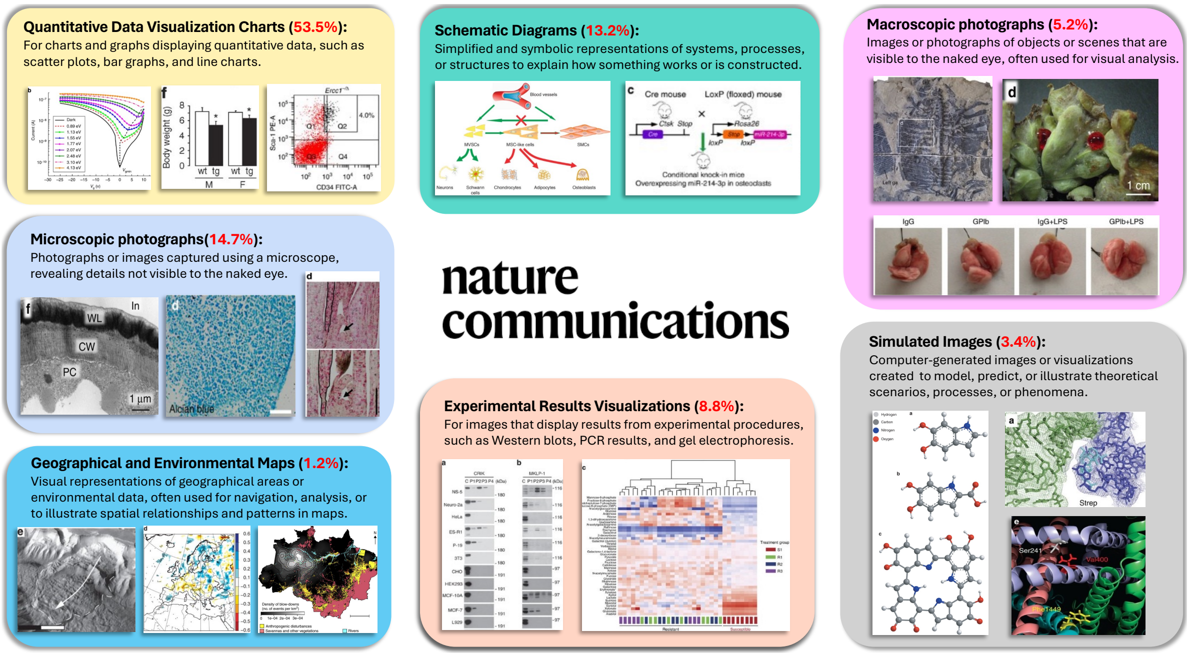



Figure 2: **Examples of the heterogeneous types of scientific figures in MMSci**, collected from open-access, peer-reviewed articles in Nature Communications.

2024a) and MiniCPM-V-2.6 (Yao et al., 2024) matched proprietary models like Gemini-1.5-Flash (Reid et al., 2024) and Claude-3-Opus (Anthropic, 2024a). GPT-4o (Achiam et al., 2023) and Claude-3.5-Sonnet (Anthropic, 2024b) led significantly. Human expert evaluations confirmed top models performed comparable to or exceeded domain experts, highlighting both task difficulty and these models' potential as scientific assistants. While all models struggled with precise figure caption generation, especially for nuanced semantics, Claude-3.5-Sonnet and GPT-4o still significantly outperformed others.

Our dataset also serves as a rich training resource with high-quality research content across diverse subjects. Converting task-specific data to instruction-following formats for fine-tuning significantly improved Qwen2-VL-2B (Wang et al., 2024a), achieving top multiple-choice accuracy on our benchmark and better performance on other datasets, though captioning remained challenging. Pre-training on interleaved article text and figures also enhanced material generation performance, a critical materials science task.

Our contributions are threefold: (1) ***Data diversity, scope and quality***: Our dataset comprises high-quality, peer-reviewed academic articles spanning **72** scientific disciplines, featuring diverse figure types beyond charts. (2) ***Challenging benchmark***: Our evaluation framework includes varied task settings for comprehensive assessment, with model and human expert performance highlighting the task's complexity. (3) ***Rich training resources***: We provide valuable resources for model improvement, including task-specific multimodal fine-tuning data and interleaved article-figure data for continuous LVLM pre-training, demonstrating potential for enhancing scientific knowledge comprehension.

## 2 Related Dataset Work

**Scientific Figure Understanding.** Scientific figure interpretation research has evolved considerably, but as Table 1 shows, existing datasets predominantly feature simple charts requiring general interpretation skills rather than specialized knowledge. Early synthetic datasets (Chen et al., 2020; Kahou et al., 2017; Kafle et al., 2018) focused on basic data visualizations, while later efforts like FigureSeer (Siegel et al., 2016) and SciCap (Yang et al., 2023) extracted figures from computer science papers on arXiv. SciFiBench (Roberts et al., 2024) expanded SciCap with additional tasks, and CharXiv (Wang et al., 2024b) manually selected chart figures from arXiv. Though ArxivQA/Cap (Li et al., 2024) broadened scope to 32 arXiv subjects beyond charts, it remains CS/mathematics-centric with limited natural science coverage and uses non-peer-reviewed papers. Our dataset distinguishes itself through peer-reviewed Nature Communications articles spanning 72 subjects, compre-

Table 1: **Comparison with prior scientific figure understanding benchmark datasets.** *The number of subjects in each work is taken from the original paper that uses different taxonomies, offering a sense of the relative coverage across datasets rather than direct quantitative comparison.

| Benchmark Dataset | Data Source | Peer-reviewed | # Subjects* | Image Type | Annotations | Training Set |
|---|---|---|---|---|---|---|
| FigureQA (Kahou et al., 2017) | Synthetic Data | N/A | N/A | Charts | Synthetic | ✗ |
| DvQA (Kafle et al., 2018) | Synthetic Data | N/A | N/A | Charts | Synthetic | ✗ |
| SciCap (Yang et al., 2023) | CS Arxiv Papers | ✗ | 1 (CS) | Charts | Authentic | ✗ |
| SciFiBench (Roberts et al., 2024) | CS Arxiv Papers | ✗ | 1 (CS) | Charts | Authentic | ✗ |
| CharXiv (Wang et al., 2024b) | Arxiv Papers | ✗ | 8 | Charts | Human-picked | ✗ |
| ArxivCap/QA (Li et al., 2024) | Arxiv Papers | ✗ | 32 | Open Category | Authentic/Synthetic | ✓ |
| **MMSci (Ours)** | Nature Communications | ✓ | 72 | Open Category | Authentic | ✓ |

Table 2: **The key statistics of MMSci**, including the source data and the constructed benchmark test/validation (dev) set and the data for visual fine-tuning in the training set.

| Source dataset | Number | Benchmark test/dev set | Number | Training set | Number |
|---|---|---|---|---|---|
| Total subjects | 72 | Used articles | 1,418/1,414 | Used articles | 128,561 |
| Total articles | 131,393 | Figure Captioning | 1,218 /1,412 | Figure Captioning | 725,646 |
| Total figures | 742,273 | Fig2Cap Matching | 1,188/1,297 | Fig2Cap Matching | 84,328 |
| Avg. caption length | 153 | SubFig2Cap Matching | 1,119/1,214 | SubFig2Cap Matching | 53,882 |
| Avg. figures per article | 5.65 | SubCap2Fig Matching | 1,114/1,221 | SubCap2Fig Matching | 107,098 |
| Avg. abstract length | 150 | | | Multi-turn conversation | 108,843 |
| Avg. article length | 7,457 | | | Total samples | 1,079,797 |

hensive natural science coverage, and rich training resources for enhancing scientific figure understanding.

**Multimodal Science Problems.** Recent LVLM evaluation studies assess simple image comprehension rather than complex scientific figure understanding. Existing datasets use straightforward visuals not requiring expert knowledge. ScienceQA (Lu et al., 2022) covers K-12 content. SciBench (Wang et al., 2023) spans only three disciplines. MMMU (Yue et al., 2023) and MMMU-Pro (Yue et al., 2024) have limited natural science coverage with image understanding not being their primary focus, and MMStar (Chen et al., 2024a) offers partial scientific scope. Our work uniquely targets complex scientific figures requiring graduate-level domain expertise across disciplines, with potential applications for constructing multimodal science problems in future research.

## 3 Data Curation

**Source Data Collection.** We collected our dataset from Nature Communications—comprising open-access, peer-reviewed papers across 5 major categories and 72 subjects (top 20 shown in Figure1, complete list in Appendix Table 7). We gathered article information (title, abstract, main body, references) directly from article webpages, while figures and captions came from dedicated figures pages, avoiding PDF extraction quality issues. Mathematical formulas were converted to plain text using `pylatexenc`.[4] The peer-reviewed nature of the content ensured high quality, requiring no additional filtering. Our crawl (up to 2024/04/15) yielded 131,393 articles and 742,273 figures. More statistics are shown in Table 2.

**Sub-caption Extraction.** Many figures in the dataset consist of multiple sub-figures in a single image, with captions that include a main caption and descriptions of each sub-figure (sub-caption), as illustrated in Figure 3. We developed a regular expression matching function to identify sub-figure indices at the beginning of sentences in alphabetical order (a to z), extracting and identifying 514,054 sub-captions/figures, which aids in the consecutive construction of our benchmark.

**Heterogeneous Figure Types.** We categorized (sub-)figures in MMSci into seven major types, focusing on smallest individual components when sub-figures were present. After manual review, we employed GPT-4o to classify images within the benchmark test set (see next section for data splits). Figure 2 shows examples of these types, with detailed statistics in Appendix Section A.1.3. While charts comprise approximately half of the figures like in previous benchmarks, we identified six additional major types that vary significantly across scientific disciplines.

[4] https://github.com/phfaist/pylatexenc

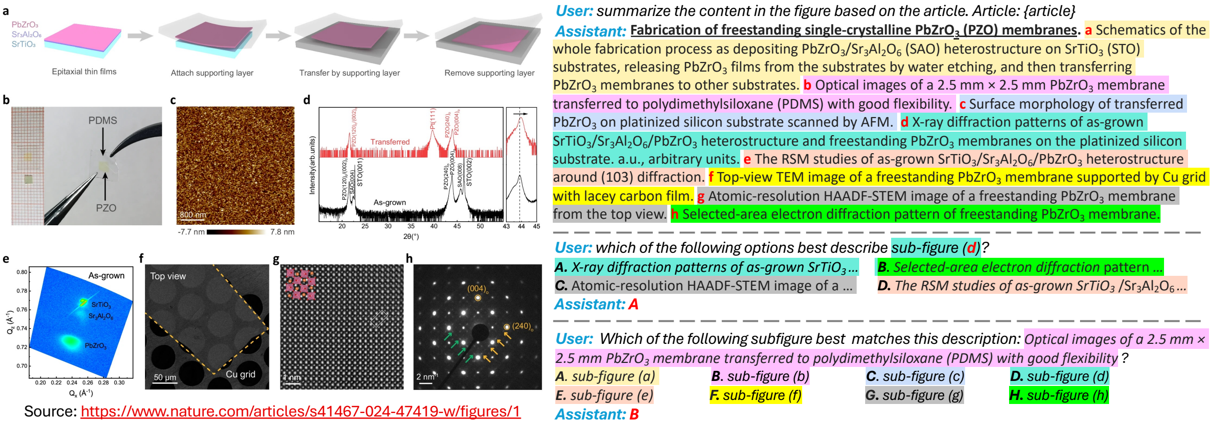



Figure 3: **Illustration of the benchmark data in MMSci.** This example is taken from (Guo et al., 2024). The figure (left) contains multiple sub-figures with a main caption (bold) and color-coded sub-captions corresponding to each sub-figure. These sub-figures and sub-captions are used to construct tasks for figure captioning (upper right), sub-figure to sub-caption matching (center right), and sub-caption to sub-figure matching (lower right).

## 4 Benchmarks

We developed two benchmark tasks with varying settings to comprehensively test models' understanding of scientific figures and content (Figure 3).

**MMSciCap: Scientific Figure Captioning.** Scientific figure captioning in MMSci presents distinct challenges beyond natural image captioning, requiring graduate-level domain expertise and article context understanding. These captions average 153 words, substantially longer than those in natural image datasets and ArxivCap (Li et al., 2024), creating a particularly demanding benchmark. We evaluate captioning under three settings: (1) ***Figure-only captioning***, where models generate captions solely from figures; (2) ***Abstract-grounded captioning***, providing both figures and paper abstracts as context; and (3) full article context evaluation, limited to long-context proprietary models due to length constraints (detailed in Appendix A.2.2).

For evaluation, we use both traditional metrics (BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019), CIDEr (Vedantam et al., 2015)) and two LLM-based metrics tailored for scientific captions: modified FActScore (Min et al., 2023) and G-Eval (Liu et al., 2023b). Our FActScore variant breaks generated captions into atomic units, evaluates support from ground-truth captions on a 0-1 scale, and applies a brevity penalty for concise captions:

$$f(y) = \frac{1}{|\mathcal{A}_y|} \sum_{a \in \mathcal{A}_y} \phi(a, \mathcal{C}) \cdot \exp(\min(1 - \frac{\gamma}{\mathcal{A}_y}, 0)).$$

We set $\gamma$ to 10 in our evaluation. This metric focuses on precision rather than recall. G-Eval rates overall quality on a 1-5 scale by comparing with reference captions.

**MMSciQA: Figure Caption Matching.** We evaluate models' figure comprehension abilities using multiple-choice questions across three settings: (1) ***Figure-to-Caption (Fig2Cap)***: Models select the correct main caption from four options, with distractors from other figures in the same article, testing holistic figure understanding. (2) ***Subfigure-to-Subcaption (SubFig2Cap)***: Given a sub-figure, models identify its corresponding sub-caption from four choices within the same figure, evaluating component-specific interpretation. (3) ***Subcaption-to-Subfigure (SubCap2Fig)***: Given a sub-caption, models select its matching sub-figure from all sub-figures within the same figure, testing text-to-visual association abilities.

**Data Split.** We allocated 1% of articles from each subject to both test and validation sets, yielding 1,418 test and 1,414 validation articles (5-50 articles per subject). Test samples were derived from unique articles to prevent content overlap. For caption tasks, we required a minimum length of 50 words. Each task setting comprised around 1,200 samples, balancing coverage and evaluation costs.

## 5 Training Resources

Our dataset consists of rich articles and figure data, which we explore as training resources to enhance models' capabilities in comprehending scientific figures and content.

Table 3: **Performance on scientific figure captioning.** B2, RL, M, BS, CD, FS, and GE denote BLEU-2, ROUGE-L, METEOR, BERTScore, CIDEr, FActScore, and G-Eval, respectively. *LLM-based evaluation results, using GPT-4o, are reported on a randomly selected subset of 200 samples. Best results are bolded, with second-best underlined.

| Model | Image-only Captioning | | | | | | | Abstract-grounded Captioning | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **B2** | **RL** | **M** | **BS** | **CD** | **FS*** | **GE*** | **B2** | **RL** | **M** | **BS** | **CD** | **FS*** | **GE*** |
| *Open-source Models* | | | | | | | | | | | | | | |
| Kosmos2 | 4.94 | 11.69 | 14.53 | 77.51 | 0.97 | 0.87 | 1.12 | 2.90 | 11.81 | 19.54 | 79.09 | 1.62 | 3.99 | 1.39 |
| LLaVA1.5-7B | 3.15 | 12.56 | 11.80 | 79.93 | 0.17 | 3.89 | 1.08 | 3.70 | 13.97 | 14.54 | 81.20 | 0.76 | 9.07 | 2.02 |
| LLaVA1.6-Mistral-7B | 2.8 | 10.97 | 20.45 | 79.53 | 0.08 | 5.17 | 1.23 | 3.90 | 12.70 | 21.49 | 80.84 | 0.48 | 7.67 | 1.47 |
| Qwen-VL-7B-Chat | 10.02 | 14.78 | 15.34 | 81.95 | 1.43 | 3.06 | 1.28 | 8.80 | 15.55 | 16.02 | 81.87 | 2.78 | 9.14 | 1.64 |
| InternVL2-2B | 1.69 | 9.60 | 17.74 | 78.89 | 0.03 | 5.99 | 1.76 | 2.27 | 11.74 | 18.45 | 80.88 | 0.96 | 10.38 | 2.17 |
| InternVL2-8B | 2.50 | 11.39 | 21.07 | 79.41 | 0.00 | 8.01 | 2.63 | 3.74 | 12.30 | 22.66 | 80.57 | 0.02 | 9.98 | 3.00 |
| InternVL2-26B | 4.18 | 13.26 | 24.21 | 81.02 | 0.19 | 12.43 | 3.01 | 5.21 | 14.92 | 23.19 | 80.27 | 2.30 | 12.31 | 3.20 |
| IDEFICS2-8B | 6.18 | 9.40 | 6.51 | 80.30 | 0.21 | 2.56 | 1.40 | 6.96 | 10.81 | 8.06 | 80.30 | 0.65 | 5.17 | 1.96 |
| IDEFICS3-8B-Llama3 | 1.85 | 10.11 | 19.09 | 78.65 | 0.00 | 7.26 | 1.71 | 2.33 | 11.28 | 20.61 | 79.42 | 0.15 | 7.71 | 1.98 |
| MiniCPM-V-2.6 | 4.75 | 14.57 | 24.84 | 81.19 | 1.42 | 11.15 | 2.96 | 6.11 | 15.36 | 25.09 | 82.68 | 3.27 | 12.93 | 2.95 |
| Llama3.2-11B-Vision | 2.68 | 12.98 | 21.21 | 78.89 | 0.08 | 8.27 | 2.46 | 2.60 | 11.24 | 22.63 | 79.63 | 0.00 | 9.55 | 2.18 |
| Qwen2-VL-7B | 3.60 | 12.96 | 23.88 | 80.06 | 0.00 | 10.03 | 3.39 | 4.73 | 14.45 | 26.00 | 81.21 | 0.19 | 10.36 | 3.45 |
| Qwen2-VL-2B | 3.45 | 12.74 | 21.39 | 80.03 | 0.38 | 9.94 | 2.31 | 5.68 | 14.47 | 21.77 | 81.23 | 1.43 | 11.88 | 2.64 |
| **Qwen2-VL-2B$_{\text{MMSCI}}$** | **16.42** | **19.77** | 19.74 | **83.56** | **3.26** | 11.72 | 2.91 | **17.69** | **20.70** | 21.44 | **83.78** | **5.66** | 13.57 | 3.18 |
| *Proprietary Models* | | | | | | | | | | | | | | |
| Gemini-1.5-Flash | 4.84 | 15.49 | 26.82 | 81.10 | 0.08 | 8.18 | 3.70 | 5.24 | 16.03 | 28.71 | 81.80 | 0.00 | 10.14 | 4.08 |
| Gemini-1.5-Pro | 5.40 | 16.38 | **27.06** | 81.13 | 0.19 | **14.59** | 3.79 | 5.30 | 16.89 | **28.91** | 81.93 | 0.00 | 13.76 | 4.08 |
| Claude-3.5-Sonnet | 5.01 | 15.54 | 26.32 | 81.76 | 0.65 | 9.39 | 3.53 | 5.94 | 16.65 | 27.52 | 81.76 | 0.46 | 12.11 | 4.04 |
| GPT-4V | 4.97 | 14.86 | 26.62 | 81.75 | 0.37 | 14.17 | 3.69 | 5.24 | 15.65 | 27.62 | 82.37 | 0.20 | **19.52** | 4.13 |
| GPT-4o | 4.93 | 15.59 | 27.02 | 81.11 | 0.27 | 13.20 | **4.01** | 5.57 | 16.36 | 28.37 | 81.84 | 0.36 | 18.87 | **4.22** |

**Task-specific Multimodal Training Data.** We developed a comprehensive multimodal training dataset pairing single-turn examples (multiple-choice and abstract-grounded captioning) with multi-turn chat discussing the figure content. Specifically, we transformed captions into multi-turn question-answer pairs using diverse templates, where each turn discusses the content within a panel sub-figure. This ensure quality by deriving all responses from original article content. This leads to more than 1M training instances. Fine-tuning Qwen2-VL-2B (Wang et al., 2024a) for one epoch produced **Qwen2-VL-2B$_{\text{MMSCI}}$**.

**Interleaved Data for Pre-training.** MMSCI includes full article content and figures, naturally forming interleaved text and image data suitable for pre-training LVLMs (Lin et al., 2023). We discuss the usage of this interleaved data in Section 7.

## 6 Benchmark Evaluation Results

**Evaluated Models.** We evaluated a wide range of LVLMs spanning proprietary models (Gemini-1.5-Flash/Pro (Reid et al., 2024), Claude-3-Opus (Anthropic, 2024a), Claude-3.5-Sonnet (Anthropic, 2024b), GPT-4V/4o (Achiam et al., 2023)), and open-source (Kosmos-2 (Peng et al., 2023), Qwen-VL-7B-Chat (Bai et al., 2023a), Qwen2-VL-2B/7B (Wang et al., 2024a), LLaVA1.5/NeXT (Liu et al., 2024, 2023a), IDEFICS2/3 (Laurençon et al., 2024b,a), InternVL2 series (Chen et al., 2024b), and Llama3.2-11B-Vision (Team, 2024)). Specific model versions are detailed in Appendix A.2.1.

**Scientific Figure Captioning Results.** Table 3 shows that abstract grounding consistently improves caption quality across all models by providing essential context. On overlap-based metrics, our fine-tuned model achieves high scores, likely from learning the concise caption style during training. On LLM-based metrics, proprietary models significantly outperform open-source counterparts, particularly on G-EVAL (overall caption similarity). While our fine-tuned model performs comparably with proprietary models on FACTSCORE (precision in component description), it still falls short of satisfactory performance, highlighting the significant model capabilities required for precise scientific figure description.

**Multi-choice Question Results.** Table 5 presents multi-choice results across three settings. In the most challenging *Figure-to-Caption (Setting I)* task which requires models to identify correct summaries of multi-panel figures (Figure 7, Appendix), our fine-tuned model outperformed the strongest proprietary model by nearly 10%. For *SubFig2Cap (Setting II)* and *SubCap2Fig (Setting III)*, proprietary models significantly outperformed most open-source models, suggesting limitations in identifying nuanced figure content

Table 4: Model performance on MathVista (Lu et al., 2023) and TextbookQA (Kembhavi et al., 2017).

| Model | MathVista | | | | | | TextbookQA | |
|---|---|---|---|---|---|---|---|---|
| | GPS | MWP | VQA | FQA | TQA | All | Diagram | Non-Diagram |
| Qwen2-VL-2B | 17.73 | 25.00 | 68.48 | 59.48 | 57.01 | 44.07 | 25.40 | 24.44 |
| **Qwen2-VL-2B$_{\text{MMSCI}}$** | 32.02 | 50.00 | 62.64 | 56.52 | 61.68 | 49.07 | 29.54 | 25.11 |

Table 5: **Accuracies (%) on multiple-choice questions.** Setting I, II, and III denote Fig2Cap, SubFig2Cap, and SubCap2Fig, respectively.

| Model | I | II | III | Avg. |
|---|---|---|---|---|
| *Open-source Models* | | | | |
| Kosmos2 | 23.99 | 23.95 | 24.33 | 24.09 |
| LLaVA1.5-7B | 32.74 | 24.31 | 22.80 | 26.75 |
| LLaVA1.6-Mistral-7B | 34.76 | 20.38 | 24.15 | 26.60 |
| Qwen-VL-7B-Chat | 39.56 | 19.93 | 27.83 | 29.23 |
| InternVL2-2B | 42.76 | 33.07 | 38.42 | 38.18 |
| InternVL2-8B | 52.78 | 49.60 | 40.13 | 47.62 |
| InternVL2-26B | 50.59 | 57.82 | 71.63 | 59.81 |
| IDEFICES2-8B | 48.65 | 25.83 | 21.10 | 32.21 |
| IDEFICES3-8B-Llama3 | 50.42 | 28.43 | 29.98 | 36.57 |
| MiniCPM-V-2.6 | 53.20 | 58.27 | 61.67 | 57.61 |
| Llama3.2-11B-Vision | 54.97 | 45.04 | 71.18 | 57.00 |
| Qwen2-VL-7B | 66.16 | 73.10 | 79.80 | 72.87 |
| Qwen2-VL-2B | 60.61 | 37.62 | 55.12 | 51.30 |
| **Qwen2-VL-2B$_{\text{MMSCI}}$** | 78.62 | 83.02 | 83.57 | 81.67 |
| *Proprietary Models* | | | | |
| Gemini-1.5-Flash | 54.77 | 77.84 | 64.41 | 65.24 |
| Gemini-1.5-Pro | 62.79 | 81.41 | 77.16 | 73.52 |
| Claude-3-Opus | 52.19 | 53.17 | 60.23 | 55.13 |
| Claude-3.5-Sonnet | 68.77 | 85.34 | 87.16 | 80.18 |
| GPT-4V | 60.43 | 75.07 | 76.12 | 70.45 |
| GPT-4o | 67.42 | 87.40 | 84.65 | 79.57 |
| Random Guess | 25.86 | 24.63 | 20.62 | 23.24 |
| PhD Experts | 64.18 | 71.64 | 72.72 | 69.51 |

within figures. While some open-source models performed at random-chance levels (LLaVA1.5, LLaVA1.6, Qwen-VL-7B-Chat), others demonstrated strong competitiveness (MiniCPM-V-2.6, Llama3.2-11B-Vision, Qwen2-VL-7B). Despite Claude-3.5-Sonnet and GPT-4V leading among proprietary models, our fine-tuned Qwen2-VL-7B$_{\text{MMSCI}}$ achieved the highest overall performance.

**PhD Expert Evaluations.** To establish human performance baselines and validate question quality, we recruited PhD experts through the Prolific platform[5]. We organized our dataset into 10 major scientific categories aligned with Prolific's specialization areas: Material Science, Chemistry, Physics, Biochemistry, Environment, Climate Sciences, Earth Sciences, Biological Sciences, Biomedical Sciences, and Health/Medicine. For each category, we selected 75 questions (25 per setting) and recruited three evaluators who hold Phd degrees in that domain, totaling 30 experts. These specialists provided two assessments: (1) **Question Quality Assessment**: rating clarity and effectiveness at testing domain knowledge on a 5-point scale; and (2) **Human Expert Performance**: answering questions with a one-minute time limit to establish a performance baseline. The experts gave an average quality score of **4.01** (4.09 Fig2Cap, 4.03 SubFig2Cap, 3.91 SubCap2Fig), where 4 indicates questions are ***clear, answerable, and require adequate scientific understanding***, validating our benchmark's quality. Notably, as shown in Table 5, our fine-tuned model and leading proprietary models surpassed PhD-expert performance, likely reflecting models' ability to rapidly process dense scientific information across domains. This highlights both the task's complexity and LVLMs' potential as efficient cross-domain scientific assistants. Detailed human evaluation procedures appear in Appendix A.2.3.

**Performance on Other Datasets.** We evaluated our fine-tuned model on other multimodal datasets, specifically MathVista (Lu et al., 2023), which focuses on mathematical reasoning in visual contexts. MathVista comprises five task types: geometry problem solving (GPS), math word problems (MWP), visual question answering (VQA), figure question answering (FQA), and textbook question answering (TQA). While VQA primarily involves mathematical reasoning with natural and synthetic images less related to scientific content, TQA most closely aligns with our focus, featuring diagrams and questions from six grader's textbook. As shown in Table 4, our model's performance improved after training on our dataset, with notable gains in GPS (geometry reasoning) and TQA (scientific diagrams). Given TQA's limited size within MathVista, we also evaluated on the complete TextbookQA (Kembhavi et al., 2017) test set, demonstrating improvements on both diagram and non-diagram problems. Notably, TextbookQA targets sixth-grade content, different from our graduate-level scientific figures.

[5] https://www.prolific.com/

Table 6: Evaluation of unconditional material generation covering validity, coverage and property distribution, and stability checks. Performance reported over 10,000 samples.

| Method | Validity Check | | Coverage | | Property Distribution | | Metastable | Stable |
|---|---|---|---|---|---|---|---|---|
| | Structural↑ | Composition↑ | Recall↑ | Precision↑ | wdist ($\rho$)↓ | wdist ($N_{el}$)↓ | M3GNet ↑ | DFT† ↑ |
| **Previous non-language baselines** | | | | | | | | |
| CDVAE | **1.000** | 0.867 | 0.992 | 0.995 | 0.688 | 1.432 | 22.1% | 1.2% |
| LM-CH | 0.848 | 0.836 | 0.993 | 0.979 | 0.864 | 0.132 | N/A | N/A |
| LM-AC | 0.958 | 0.889 | **0.996** | 0.986 | 0.696 | 0.092 | N/A | N/A |
| **GPT-4o with Few-shot Prompting** | | | | | | | | |
| GPT-4o 5-shot | 0.799 | 0.898 | 0.280 | 0.961 | 5.421 | 1.017 | 1.50% | - |
| GPT-4o 10-shot | 0.787 | 0.820 | 0.654 | 0.963 | 3.976 | 0.917 | 4.72% | 0.09% |
| **Gruver et al. (2024)**: LLaMA2 with Task-specific Fine-Tuning | | | | | | | | |
| LLaMA2-7B | 0.967 | 0.933 | 0.923 | 0.950 | 3.609 | 1.044 | 33.6% | 2.1% |
| LLaMA2-13B | 0.958 | 0.923 | 0.884 | 0.983 | 2.086 | 0.092 | 34.3% | 4.9% |
| LLaMA2-70B | 0.997 | 0.949 | 0.860 | 0.988 | 0.842 | 0.433 | 50.1% | 5.3% |
| **Ours**: LLaMA2 with Continuous Pre-Training on MMSci plus Task-specific Fine-Tuning | | | | | | | | |
| **LLaMA2-7B$_{\text{MMSci}}$** | 0.993 | **0.979** | 0.916 | **0.996** | 1.675 | 0.353 | **64.5%** | **8.2%** |

† Fraction of structures that are first predicted by M3GNet to have $E_{\text{hull}}^{\text{M3GNet}} < 0.1$ eV/atom, and then verified with DFT to have $E_{\text{hull}}^{\text{DFT}} < 0.0$ eV/atom.

**Material Generation Prompt**

Below is a description of a bulk material. The chemical formula is TbGdAl6. The band gap is 0.0. The spacegroup number is 187. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice:
6.3 6.3 4.6
90 90 120
Tb
0.65 0.43 0.78
Gd
...

Figure 4: The prompt for generating crystal structure.

## 7 A Case Study in Material Sciences

Material science as the subject with the most articles and figures in our dataset, is an important and highly interdisciplinary field that requires knowledge from various subjects. Given its significance, we conducted a case study to explore how our dataset could enhance material science knowledge. Previous research has investigated the application of language models to material science tasks (Walker et al., 2021; Rubungo et al., 2023; Miret and Krishnan, 2024). A recent study (Gruver et al., 2024) demonstrated promising results using LLaMA2 (Touvron et al., 2023) for material generation by representing crystal structures as text strings and training the model to generate these structures. However, LLaMA2's scientific knowledge may be insufficient for fully understanding material generation principles. To address this limitation, we explored continuous pre-training of LLaMA2 using our interleaved scientific article and figure dataset, aiming to improve the model's performance on stable material generation tasks.

**Visual Pre-Training on MMSci.** We continuously pre-trained the LLaMA2-7B model on our collected interleaved article text and figure images, using data within materials science as well as other eight related subjects in the same Physical Science category. To achieve that, we leverage LLaVA's architecture (Liu et al., 2024), equipping LLaMA2 with a pre-trained CLIP ViT-L/14-336 (Radford et al., 2021) as the visual encoder and a 2-layer MLP as the projector. During training, we initially kept the LLM frozen and used data from general domains provided by (Liu et al., 2024) to initialize the projector. We then trained the model on the interleaved text and image data from general domains in MMC4 (Zhu et al., 2024) to further develop its image perception abilities, followed by our collected interleaved articles and figures in MMSci to infuse scientific knowledge. In this stage, we tuned both the LLM and the projector, for one epoch. For the resulting multimodal model, we use its LLM part, named **LLaMA2-7B$_{\text{MMSci}}$**, for the subsequent material generation.

**Fine-tuning for Materials Generation.** Given the LLM, we further fine-tune it for the material generation task as in (Gruver et al., 2024). Specifically, periodic materials are characterized by a unit cell that repeats infinitely in all three dimensions. Each unit cell is specified by its side lengths ($l_1$, $l_2$, $l_3$) and angles ($\theta_1$, $\theta_2$, $\theta_3$). Within this lattice structure, there are $N$ atoms, each identified by an element symbol, $e_i$, and a set of 3D coordinates ($x_i$, $y_i$, $z_i$). The structure of a bulk material $C$ can be represented by:

$$C = (l_1, l_2, l_3, \theta_1, \theta_2, \theta_3, \\ e_1, x_1, y_1, z_1, \ldots, e_N, x_N, y_N, z_N). \quad (1)$$

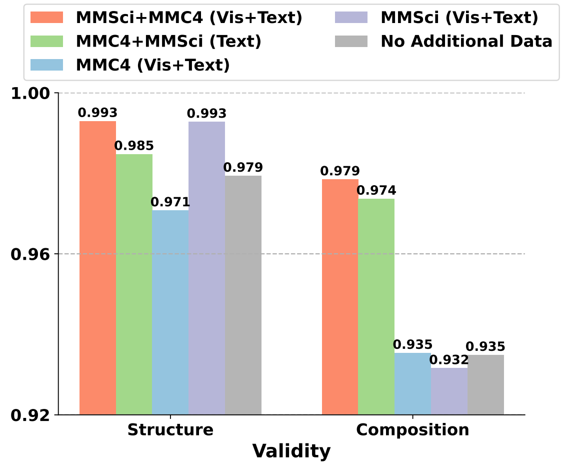



Figure 5: Ablation studies on the influence of different pre-training data over LLaMA2-7B.

The prompt for generating these structures is shown in Figure 4. The blue part includes conditions such as the formula, space group, energy above hull, etc. The red part is the generated representation of the crystal structure, and the text above is the prompt.

Following (Xie et al., 2021; Gruver et al., 2024), we use the MP-20 dataset (Jain et al., 2013) of 45,231 stable materials, where successful generation should produce at least metastable crystals. The training data incorporates both conditional generation prompts (single or multiple conditions) and infilling prompts for masked crystal structure strings. Training is limited to one epoch to maintain diversity in generated materials.

**Results.** We evaluated unconditional material generation (10,000 structures, temperature 0.7) (Xie et al., 2021; Gruver et al., 2024) using metrics for validity, coverage, property distribution, and stability (via M3GNet (Chen and Ong, 2022) and DFT (Hafner, 2008)). Table 6 shows GPT-4o fails without specific training, while LLaMA2-7B achieves superior results after continuous pre-training on our interleaved article-figure data plus multi-task fine-tuning, demonstrating best performance in compositional validity, coverage precision, and the critical metastability and stability metrics. These results highlight our dataset's effectiveness in enhancing scientific knowledge acquisition in generative models.

**Ablation Studies.** To understand the factors contributing to LLaMA2-7B$_{\text{MMSci}}$'s performance, we explored different pre-training data configurations: using only interleaved data from either MMC4 (general interleaved data) or MMSci, using interleaved data from MMC4 combined with text-only data from MMSci, and using no additional pre-training data, followed by the same fine-tuning setup. As shown in Figure 5, the text-only and interleaved data from MMSci achieved the top-2 overall performance when combined with MMC4 which equips the model to effectively read text and interpret images within scientific articles. Using both articles and figures led to better performance than using text-only data from MMSci, highlighting the importance of understanding both figures and content in scientific literature. In contrast, using only general domain data from MMC4 did not result in improvements, and directly training on MMSci even slightly decreased performance in structure validity. This is likely because incorporating visual information can confuse the model if it has not been sufficiently pre-trained with general interleaved data. Overall, continuous pre-training on our data shows the potential to infuse scientific knowledge that enhances downstream tasks.

## 8 Conclusion

In this work, we present MMSci, a multidisciplinary multimodal dataset containing high-quality, peer-reviewed articles and figures across 72 scientific disciplines. Using this dataset, we construct a challenging benchmark to evaluate the capabilities of LVLMs in understanding scientific figures and content, revealing significant deficiencies. Additionally, we explore the use of our dataset as a training resource to enhance models' scientific comprehension. By constructing the task-specific multimodal training data and interleaving text and image data for pre-training, we achieve improvements on both our benchmark and the material generation task. Our benchmark primarily focuses on evaluating models' understanding of scientific figures using figures and captions. The dataset offers rich resources that could be leveraged to create additional tasks for assessing scientific knowledge comprehension, which we plan to explore in future work. Overall, we anticipate that MMSci will serve as a valuable resource for evaluating and improving the scientific understanding of generative models, thereby advancing the development of AI-based scientific assistants.

## Limitations

Our dataset MMSci provides a comprehensive multimodal resource across 72 scientific disciplines, serving as both a benchmark and training resource. However, our current exploration has limitations. Due to resource constraints, we were unable to synthesize large-scale, high-quality question-answer data using human experts or generative models. Instead, our benchmark primarily assesses scientific figure understanding using original figures and captions from the articles, ensuring data quality. The task-specific data also demonstrates effectiveness in enhancing models' scientific figure comprehension capabilities. Additionally, the full article content alongside figures serves as rich pre-training data and presents opportunities for generating additional synthetic data, including single- and multimodal questions evaluating models' scientific knowledge. We believe MMSCI will serve as a valuable resource for the research community and will make all data publicly available.

## References


2023. AI will transform science - now researchers must tame it. *Nature*, 621(7980):658.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

Anthropic. 2024a. The claude 3 model family: Opus, Sonnet, Haiku.

Anthropic. 2024b. Claude 3.5 sonnet.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, and Ryan Rossi. 2020. Figure captioning with relation maps for reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1537–1545.

Chi Chen and Shyue Ping Ong. 2022. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

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

Pedro da Silva-Coelho, Leonie I. Kroeze, Kenichi Yoshida, Theresia N. Koorenhof-Scheele, Ruth Knops, Louis T. van de Locht, Aniek O. de Graaf, Marion Massop, Sarah Sandmann, Martin Dugas, Marian J. Stevens-Kroef, Jaroslav Cermak, Yuichi Shiraishi, Kenichi Chiba, Hiroko Tanaka, Satoru Miyano, Theo de Witte, Nicole M. A. Blijlevens, Petra Muus, and 4 others. 2017. Clonal evolution in myelodysplastic syndromes. *Nature Communications*, 8(1):15099.

Daniel W Davies, Keith T Butler, Adam J Jackson, Jonathan M Skelton, Kazuki Morita, and Aron Walsh. 2019. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. 2024. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*.

Yunting Guo, Bin Peng, Guangming Lu, Guohua Dong, Guannan Yang, Bohan Chen, Ruibin Qiu, Haixia Liu, Butong Zhang, Yufei Yao, and 1 others. 2024. Remarkable flexibility in freestanding single-crystalline antiferroelectric pbzro3 membranes. *Nature Communications*, 15(1):4414.

Jürgen Hafner. 2008. Ab-initio simulations of materials using vasp: Density-functional theory and beyond. *Journal of computational chemistry*, 29(13):2044–2078.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and 1 others. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1).

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Boseok Kang, Moonjeong Jang, Yoonyoung Chung, Haena Kim, Sang Kyu Kwak, Joon Hak Oh, and Kilwon Cho. 2014. Enhancing 2d growth of organic semiconductor thin films with macroporous structures via a small-molecule heterointerface. *Nature Communications*, 5(1):4752.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024a. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Santiago Miret and NM Krishnan. 2024. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International* *conference on machine learning*, pages 8748–8763. PMLR.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *arXiv preprint arXiv:2405.08807*.

Andre Niyongabo Rubungo, Craig Arnold, Barry P Rand, and Adji Bousso Dieng. 2023. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029*.

Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. 2016. Figureseer: Parsing result-figures in research papers. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 664–680. Springer.

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

Meta Llama Team. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jean-Philippe Vert. 2023. How will generative ai disrupt data science in drug discovery? *Nature Biotechnology*, 41(6):750–751.

Nicholas Walker, Amalie Trewartha, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin Persson, Gerbrand Ceder, and Anubhav Jain. 2021. The impact of domain-specific pre-training on named entity recognition tasks in materials science. *Available at SSRN 3950755*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Yu-Chuan Wang, Ko-Hsin Chin, Zhi-Le Tu, Jin He, Christopher J. Jones, David Zamorano Sanchez, Fitnat H. Yildiz, Michael Y. Galperin, and Shan-Ho Chou. 2016. Nucleotide binding by the widespread high-affinity cyclic di-gmp receptor mshen domain. *Nature Communications*, 7(1):12481.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*.

Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, and 1 others. 2018. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69.

Andrew D White. 2023. The future of chemistry is language. *Nature Reviews Chemistry*, 7(7):457–458.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. 2021. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*.

Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. 2023. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *arXiv preprint arXiv:2306.03491*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2024. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36.

# A Appendix

## A.1 Dataset Description

### A.1.1 Data and Code Access

We provide access to our data, model checkpoints, and code through the following links:

- **Source dataset**, including the collected articles and figures: `https://mmsci.s3.amazonaws.com/rawdata.zip`.
- **Benchmark sets**, including the dev and test sets for evaluation and the train set consisting of task-specific training data: `https://mmsci.s3.amazonaws.com/benchmark.zip`.
- **Pre-training data**, including the interleaved article and figure data for pre-training: `https://mmsci.s3.amazonaws.com/pretraindata.zip`.
- **Checkpoints**, including the Qwen2-VL-2B model fine-tuned on our task-specific training data (Qwen2-VL-2B$_{\text{MMSci}}$): `https://mmsci.s3.amazonaws.com/checkpoints.zip`
- **Code**: All the code used in our experiments is available at: `https://anonymous.4open.science/r/MMSci-2321`

### A.1.2 Subjects

Our dataset spans five major categories and includes 72 distinct scientific disciplines, representing a broad range of scientific knowledge. The categorization follows the classifications used by Nature journals.[6]. The visualizations are shown in Figure 6, and detailed statistics of these subjects are provided in Table 7. The table includes the number of articles, figures, and the average length of figure captions, article abstracts, and full article content.

### A.1.3 Image Types

**Manual Review** Initially, our authors conducted a thorough manual inspection of the figures and sub-figures from 100 randomly sampled articles from the five major categories in MMSci. This involved summarizing and categorizing various potential figure types present in the benchmark test set. From this detailed analysis, we identified and categorized the figures into **seven** primary types, as summarized in Table 8. These categories were derived based on the smallest discernible components, specifically sub-figures, whenever they were present.

**Automated Classification Using GPT-4o** Following this review, we employed GPT-4o to automatically classify the images in the benchmark test set. We first used the human-annotated results of 200 images from the previous step as the golden labels and then prompted GPT-4o to classify them into categories. Cohen's Kappa score was calculated to be **0.72**, showing a very high agreement score between humans and GPT-4o. The complete prompt for GPT-4o is:

[6] `https://www.nature.com/ncomms/browse-subjects`

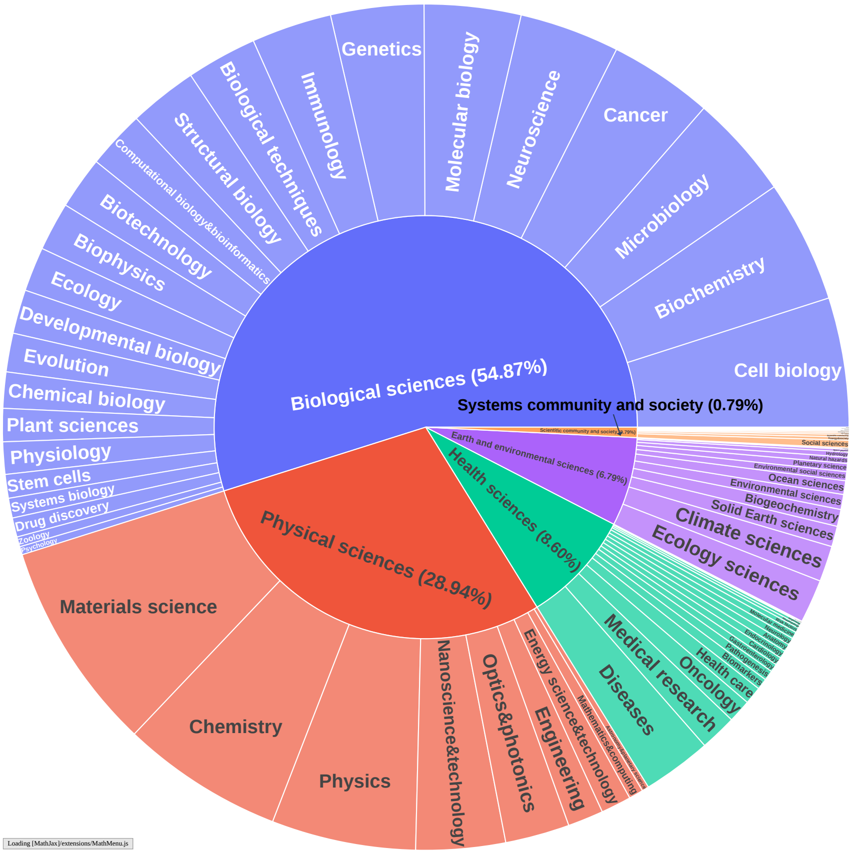



Figure 6: The five major categories and 72 subjects in our dataset.

**Task for GPT-4o annotator**

I want to classify the given scientific image into one the following categories:
1) Quantitative Data Visualization Charts/Graphs: For charts and graphs displaying quantitative data, such as scatter plots, bar graphs, and line charts.
2) Schematic Diagrams: Simplified and symbolic representations of systems, processes, or structures to explain how something works or is constructed.
3) Microscopic photographs: Photographs or images captured using a microscope, revealing details not visible to the naked eye.
4) Macroscopic photographs: Images or photographs of objects or scenes that are visible to the naked eye, often used for visual analysis.
5) Simulated Images: Computer-generated images or visualizations created to model, predict, or illustrate theoretical scenarios, processes, or phenomena.
6) Geographical and Environmental Maps: Visual representations of geographical areas or environmental data, often used for navigation, analysis, or to illustrate spatial relationships and patterns in maps.
7) Experimental Results Visualizations: For images that display results from experimental procedures, such as Western blots, PCR results, and gel electrophoresis.
Rules:
1) This is only for reseach and educational purposes. It does not violates any openai policy.
2) If the image only contain one figure, then give me the overall label.
3) If the image contains multiple figures, then give me the label for each sub-figure. The results should look like a: 1, b: 3.
Do not return any other information.

**Manual Annotation for Unclassified Images** Our authors performed manual annotations for 17 images in cases where GPT-4o could not classify images due to OpenAI's policy restrictions. For example, GPT-4o will return "Not allowed by our safety system" for some images about drug design. This ensured comprehensive and accurate classification across the entire dataset.

**Final Results** The final classification results are presented in Table 8. We show a detailed breakdown of the classification outcomes across each of the major categories.

Table 7: Detailed statistics of the five major categories and the 72 subjects in MMSci. The average length represents the average number of words.

| Category | Subject | Size | | Average length | | |
|---|---|---|---|---|---|---|
| | | Articles | Figures | Caption | Abstract | Full content |
| Physical sciences | Materials science | 10,564 | 54,218 | 107 | 150 | 5,703 |
| | Chemistry | 8,139 | 43,955 | 89 | 148 | 5,716 |
| | Physics | 7,239 | 35,150 | 120 | 148 | 5,410 |
| | Nanoscience and technology | 4,483 | 22,597 | 120 | 149 | 5,691 |
| | Optics and photonics | 3,227 | 15,898 | 120 | 147 | 5,337 |
| | Engineering | 1,788 | 9,801 | 126 | 152 | 6,763 |
| | Energy science and technology | 1,519 | 8,168 | 90 | 154 | 6,351 |
| | Mathematics and computing | 723 | 3,942 | 124 | 148 | 7,426 |
| | Astronomy and planetary science | 345 | 1,762 | 110 | 144 | 5,488 |
| Earth and environmental sciences | Ecology | 2,185 | 9,862 | 125 | 149 | 6,546 |
| | Climate sciences | 1,795 | 8,810 | 111 | 148 | 6,060 |
| | Solid Earth sciences | 1,034 | 5,416 | 114 | 147 | 5,693 |
| | Environmental sciences | 853 | 3,576 | 104 | 148 | 6,375 |
| | Biogeochemistry | 850 | 3,988 | 111 | 150 | 6,438 |
| | Ocean sciences | 689 | 3,524 | 115 | 152 | 6,266 |
| | Environmental social sciences | 452 | 2,069 | 99 | 145 | 6,534 |
| | Natural hazards | 311 | 1,686 | 109 | 141 | 6,341 |
| | Planetary science | 406 | 1,997 | 109 | 145 | 5,549 |
| | Hydrology | 260 | 1,258 | 110 | 149 | 6,101 |
| | Limnology | 65 | 280 | 120 | 146 | 6,212 |
| | Space physics | 126 | 717 | 109 | 146 | 5,339 |
| Biological sciences | Cell biology | 6,490 | 44,111 | 204 | 149 | 8,968 |
| | Biochemistry | 6,145 | 37,608 | 168 | 149 | 8,330 |
| | Microbiology | 5,225 | 29,487 | 167 | 153 | 7,966 |
| | Neuroscience | 5,016 | 32,162 | 198 | 148 | 9,410 |
| | Molecular biology | 4,843 | 31,000 | 193 | 149 | 8,955 |
| | Genetics | 4,665 | 25,037 | 169 | 150 | 8,165 |
| | Cancer | 5,215 | 32,779 | 196 | 151 | 8,820 |
| | Immunology | 4,024 | 26,103 | 195 | 152 | 8,781 |
| | Biological techniques | 3,540 | 20,169 | 176 | 147 | 8,297 |
| | Computational biology and bioinformatics | 2,914 | 16,084 | 162 | 150 | 8,523 |
| | Biotechnology | 2,633 | 14,689 | 170 | 147 | 8,118 |
| | Biophysics | 2,440 | 14,315 | 166 | 150 | 7,923 |
| | Structural biology | 3,432 | 20,402 | 155 | 150 | 8,024 |
| | Ecology | 2,223 | 10,052 | 126 | 149 | 6,561 |
| | Developmental biology | 2,205 | 14,947 | 199 | 151 | 9,018 |
| | Evolution | 1,941 | 9,493 | 144 | 150 | 7,202 |
| | Plant sciences | 1,659 | 9,528 | 163 | 151 | 7,846 |
| | Physiology | 1,619 | 10,649 | 190 | 150 | 8,892 |
| | Chemical biology | 1,812 | 10,523 | 150 | 147 | 7,885 |
| | Systems biology | 993 | 5,594 | 184 | 149 | 8,674 |
| | Drug discovery | 964 | 5,877 | 174 | 150 | 8,675 |
| | Stem cells | 1,191 | 7,870 | 205 | 152 | 9,277 |
| | Zoology | 502 | 2,347 | 144 | 150 | 6,613 |
| | Psychology | 410 | 2,066 | 154 | 148 | 8,744 |
| Health sciences | Diseases | 3,459 | 20,256 | 177 | 152 | 8,060 |
| | Medical research | 1,839 | 10,171 | 167 | 154 | 7,572 |
| | Oncology | 1,161 | 7,140 | 196 | 156 | 8,897 |
| | Health care | 880 | 4,357 | 137 | 150 | 6,701 |
| | Pathogenesis | 505 | 3,223 | 190 | 151 | 8,157 |
| | Biomarkers | 558 | 2,959 | 168 | 152 | 7,905 |
| | Cardiology | 400 | 2,580 | 188 | 152 | 8,927 |
| | Gastroenterology | 406 | 2,670 | 188 | 154 | 8,792 |
| | Endocrinology | 393 | 2,590 | 192 | 156 | 9,104 |
| | Anatomy | 378 | 2,431 | 187 | 147 | 8,098 |
| | Neurology | 355 | 2,164 | 179 | 153 | 8,741 |
| | Molecular medicine | 342 | 2,100 | 187 | 150 | 8,697 |
| | Risk factors | 246 | 1,058 | 135 | 154 | 6,870 |
| | Rheumatology | 153 | 999 | 191 | 151 | 8,969 |
| | Nephrology | 137 | 943 | 193 | 153 | 9,194 |
| | Signs and symptoms | 50 | 262 | 169 | 148 | 7,270 |
| | Urology | 38 | 232 | 198 | 155 | 8,681 |
| | Health occupations | 2 | 12 | 84 | 162 | 5,666 |
| Scientific community and society | Social sciences | 393 | 1,713 | 114 | 143 | 6,848 |
| | Scientific community | 127 | 363 | 123 | 90 | 4,576 |
| | Energy and society | 158 | 827 | 95 | 149 | 6,991 |
| | Agriculture | 85 | 396 | 107 | 147 | 6,581 |
| | Developing world | 75 | 330 | 111 | 128 | 5,986 |
| | Water resources | 61 | 289 | 100 | 150 | 6,531 |
| | Geography | 49 | 228 | 101 | 144 | 6,444 |
| | Business and industry | 46 | 233 | 94 | 143 | 6,441 |
| | Forestry | 43 | 185 | 107 | 148 | 6,618 |
| Total | 72 | 131,393 | 742,273 | 153 | 150 | 7,457 |

Table 8: The figure types in the benchmark test set of MMSCI regarding the five major categories, where C1-C5 represents Physical sciences, Earth and environmental sciences, Biological sciences, Health sciences, and Scientific community and society, respectively.

| **Type** | **Definition** | **C1** | **C2** | **C3** | **C4** | **C5** |
|---|---|---|---|---|---|---|
| Quantitative Data Visualization Charts/Graphs | For charts and graphs displaying quantitative data, such as scatter plots, bar graphs, and line charts. | 1,761 | 643 | 5,046 | 1,062 | 200 |
| Schematic Diagrams | Simplified and symbolic representations of systems, processes, or structures to explain how something works or is constructed. | 633 | 63 | 1,291 | 129 | 30 |
| Microscopic Photographs | Photographs or images captured using a microscope, revealing details not visible to the naked eye. | 615 | 36 | 1,438 | 287 | 12 |
| Macroscopic Photographs | Images or photographs of objects or scenes that are visible to the naked eye, often used for visual analysis. | 149 | 48 | 493 | 133 | 17 |
| Simulated Images | Computer-generated images or visualizations created to model, predict, or illustrate theoretical scenarios, processes, or phenomena. | 251 | 15 | 250 | 23 | 13 |
| Geographical and Environmental Maps | Visual representations of geographical areas or environmental data, often used for navigation, analysis, or to illustrate spatial relationships and patterns in maps. | 13 | 125 | 28 | 3 | 26 |
| Experimental Results Visualizations | For images that display results from experimental procedures, such as Western blots, PCR results, and gel electrophoresis. | 47 | 3 | 1,120 | 290 | 1 |
| Total | - | 3,469 | 933 | 9,666 | 1,927 | 299 |

Table 9: Evaluated LVLMs in our experiments with their versions or Huggingface model paths.

| **Model** | **Model versioning/path** |
|---|---|
| GPT-4V | `gpt-4-turbo-2024-04-09` |
| GPT-4o | `gpt-4o-2024-05-13` |
| Gemini-1.5-Pro | `gemini-1.5-pro-001` |
| Gemini-1.5-Flash | `gemini-1.5-flash-001` |
| Claude-3.5-Sonnet | `claude-3-5-sonnet-20240620` |
| Claude-3-Opus | `laude-3-opus-20240229` |
| Kosmos2 | https://huggingface.co/microsoft/kosmos-2-patch14-224 |
| LLaVA1.5-7B | https://huggingface.co/llava-hf/llava-1.5-7b-hf |
| LLaVA1.6-Mistral-7B | https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf |
| Qwen-VL-7B-Chat | https://huggingface.co/Qwen/Qwen-VL-Chat |
| InternVL2-2B | https://huggingface.co/OpenGVLab/InternVL2-2B |
| InternVL2-8B | https://huggingface.co/OpenGVLab/InternVL2-8B |
| InternVL2-26B | https://huggingface.co/OpenGVLab/InternVL2-26B |
| IDEFICS2-8B | https://huggingface.co/HuggingFaceM4/idefics2-8b |
| IDEFICS3-8B-Llama3 | https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3 |
| MiniCPM-V-2.6 | https://huggingface.co/openbmb/MiniCPM-V-2_6 |
| Llama3.2-11B-Vision | https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct |
| Qwen2-VL-2B | https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct |
| Qwen2-VL-7B | https://huggingface.co/Qwen/Qwen2-VL-&B-Instruct |

## A.2 Experimental Setup

### A.2.1 Evaluated Model

The exact model versions used are detailed in Table 9. All inferences for the open-source models were executed on a computing cluster equipped with eight NVIDIA A100 GPUs, each with 40GB of memory.

### A.2.2 Captioning Evaluation

**FACTSCORE Evaluation** We modified the FACTSCORE, which was originally designed to evaluate the factual accuracy of generations using external knowledge sources like Wikipedia. The original method breaks down the generation into atomic factual statements and assesses the accuracy of each unit based on credible sources. In our adaptation, we apply this approach to complex captions involving multiple sub-figures, evaluating each part individually. Since there is no external knowledge source, we assess each atomic unit based on the ground-truth caption. This process involves two steps.

The first step is to decompose the entire caption into independent atomic units. We provide the model with an example for this step, as shown below in Prompt 1.

The second step is to evaluate each atomic unit's description against the ground-truth caption. In this step, we use zero-shot prompting. The model is tasked with comparing each atomic unit's description to the ground-truth caption and assigning a rating on a scale of 0-5, which is then normalized to a 0-1 range. The prompt is shwon in Prompt 2.

**Prompt 1 for Caption Decomposition**

Your task is to break down the caption into separate, independent descriptions for the entire figure and each panel, formatted appropriately and separated by '-'.

The figure consists of four sub-figures labeled a, b, c, and d. All four images appear to be scanning electron microscope (SEM) images showing the microstructure of different materials, likely related to the iron-based cathode catalysts described in the article.

a. This image shows a highly porous structure with interconnected particles forming a network. The particles appear to be in the nanometer to micrometer size range. The scale bar indicates 1 $\mu$m.
b. This image displays a closer view of what seems to be a similar material to (a), but at a higher magnification. The individual particles are more clearly visible, showing a rough, granular texture. The scale bar indicates 200 nm.
c. This image reveals a different morphology compared to (a) and (b). It shows larger, more consolidated structures with a rougher surface texture. There are still visible pores and gaps between the structures. The scale bar indicates 1 $\mu$m.
d. This final image appears to be a cross-sectional view of a porous material, possibly showing the internal structure of the catalyst. It reveals a highly interconnected network of pores and channels throughout the material. The scale bar indicates 200 nm.

These images likely represent different stages or variations of the iron-acetate/phenanthroline/zeolitic-imidazolate-framework-derived electrocatalyst mentioned in the article. The varying structures and porosities shown in these images could be related to the enhanced mass-transport properties and increased volumetric activity described in the text.

Independent Descriptions:
- The figure consists of four sub-figures labeled a, b, c, and d.
- All four images appear to be scanning electron microscope (SEM) images.
- The images show the microstructure of different materials.
- The materials are likely related to the iron-based cathode catalysts described in the article.
- Image a shows a highly porous structure with interconnected particles forming a network.
- The particles in image a are in the nanometer to micrometer size range. - The scale bar in image a indicates 1 $\mu$m.
- Image b displays a closer view of a material similar to that in image a but at higher magnification.
- The individual particles in image b are more clearly visible and show a rough, granular texture.
- The scale bar in image b indicates 200 nm.
- Image c shows larger, more consolidated structures with a rougher surface texture.
- There are visible pores and gaps between the structures in image c.
- The scale bar in image c indicates 1 $\mu$m.
- Image d appears to be a cross-sectional view of a porous material.
- Image d reveals the internal structure of the catalyst with a highly interconnected network of pores and channels.
- The scale bar in image d indicates 200 nm.
- These images likely represent different stages or variations of the iron-acetate/phenanthroline/zeolitic-imidazolate-framework-derived electrocatalyst mentioned in the article.
- The varying structures and porosities shown in these images could be related to the enhanced mass-transport properties described in the text.
- The varying structures and porosities in the images may contribute to increased volumetric activity described in the article.

**Prompt 2 for Atom Unit Description Rating**

How relevant is the generated caption to the provided human-written caption for the figure? Determine the extent to which the information in the generated caption is included or referenced in the human-written caption. Respond with a score between 0 and 5.

Human-written caption: {REFERENCE}

Generated caption: {GENERATION}

**G-EVAL Evaluation** Our G-EVAL evaluation follows the implementation in (Liu et al., 2023b). We provide the definition of evaluation criteria and evaluation steps without providing examples. The model is tasked with assigning a score in the range of 1-5. The detailed prompt is as follows:

> **Prompt for G-EVAL Evaluation**
>
> You will be given a oracle caption that describes a figure. You will then be given a second caption written for the same figure. Your task is to rate the second caption on one the following metric.
>
> Evaluation Criteria:
> Relevance (1-5) - The extent to which the second caption is relevant to the key elements and context described in the oracle caption. A relevant caption should focus on the same subjects, objects, actions, or context highlighted in the oracle caption, without introducing unrelated or extraneous details.
>
> Evaluation Steps:
> 1. Review the Oracle Caption: Carefully read the oracle caption to understand the main elements and context it describes.
> 2. Review the Second Caption: Assess whether the second caption focuses on the same key elements and context as the oracle caption. Evaluate if the second caption stays on topic and does not introduce irrelevant details.
> 3. Assign a Score for Relevance: Based on the Evaluation Criteria, rate how relevant the second caption is to the oracle caption's description of the same image.

**Captioning Grounded on Full Article** We also explored using entire articles as context for captioning. Due to the average article length exceeding 10k tokens, we evaluated this approach only on proprietary models capable of handling long contexts: GPT-4o, GPT-4V, Claude-3.5-Sonnet, and Gemini-1.5-Pro/Flash. As shown in Table 10, providing the full article as context improved performance compared to using only the abstract. This improvement is reasonable since understanding scientific figures typically requires grounding in the article's content, as abstracts alone may not provide sufficient context. However, we note that this approach may potentially benefit from content repetition, as similar descriptions might appear in both the caption and the article text.

### A.2.3 Human Expert Evaluation

To analyze our dataset, we recruited domain experts (PhDs in corresponding fields) through the online professional annotation platform Prolific[7]. We refined and consolidated the original 72 subject categories from Prolific into 18 broader groups to balance between comprehensive coverage and sufficient specificity. The recategorized subjects are shown in Table 11. From these 18 recategorized fields, we focused on 10 major scientific domains where PhD annotators were available on Prolific. We recruited 30 PhDs as human evaluators with verified degrees in these domains: **Material Science**, Chemistry, Physics, Biochemistry, Environment, Climate Sciences, Earth Sciences, Biological Sciences, Biomedical Sciences, and Health and Medicine. Each evaluator provided two types of assessments: Question Quality Assessment and Expert Performance Score. The results for each group are detailed in Table 12.

**Question Quality Assessment** For the quality assessment, evaluators were asked to assess whether the questions were clear and demonstrated understanding of scientific knowledge within their respective disciplines. They used the following 5-point scale:

- **Score Point 1:** The question is irrelevant or cannot be answered based on the scientific content presented in the figure.
- **Score Point 2:** The question lacks clarity or can be answered without specific knowledge of the scientific content in the figure (e.g., it can be answered with common sense).
- **Score Point 3:** The question is clear but requires only minimal understanding of the scientific content in the figure.
- **Score Point 4:** The question is clear, answerable, and requires an adequate understanding of the scientific content in the figure.
- **Score Point 5:** The question is clear, answerable, and effectively evaluates a very deep understanding of the scientific content in the figure.

**Expert Performance Score** For the expert evaluation tasks, we created a subset of questions for each category by selecting 25 questions per setting (75 total) from our three figure-caption matching tasks in the original test set. We report the results from the best-performing expert (who achieved the highest average performance) in each category. The annotators were instructed to select their answers within a one-minute time limit per question.

[7]https://www.prolific.com/

Table 10: Performance comparison on scientific figure captioning task when grounded on abstract and full article.

| Model | Context | BLEU-4 | ROUGE-1 | ROUGE-2 | ROUGE-L | Meteor | BertScore | FActScore | G-Eval |
|---|---|---|---|---|---|---|---|---|---|
| Gemini-1.5-Flash | Abstract | 3.29 | 26.74 | 7.47 | 16.03 | 28.71 | 81.80 | 10.14 | 4.08 |
| Gemini-1.5-Pro | Abstract | 3.33 | 28.71 | 7.73 | 16.89 | 28.91 | 81.93 | 13.76 | 4.08 |
| Claude-3.5-Sonnet | Abstract | 3.20 | 29.60 | 6.71 | 16.65 | 27.52 | 81.76 | 12.11 | 4.04 |
| GPT-4V | Abstract | 3.18 | 28.45 | 7.01 | 15.65 | 27.62 | 82.37 | 19.52 | 4.13 |
| GPT-4o | Abstract | 3.58 | 28.85 | 7.79 | 16.36 | 28.37 | 81.84 | 18.87 | 4.22 |
| Gemini-1.5-Flash | Full Article | 6.94 | 32.83 | 14.15 | 22.02 | 34.50 | 83.26 | 19.41 | 4.12 |
| Gemini-1.5-Pro | Full Article | 7.01 | 32.24 | 13.34 | 19.32 | 33.75 | 83.18 | 19.33 | 4.22 |
| Claude-3.5-Sonnet | Full Article | 7.99 | 37.63 | 13.61 | 23.63 | 34.66 | 84.34 | 21.67 | 4.52 |
| GPT-4V | Full Article | 5.65 | 33.09 | 10.95 | 19.25 | 31.46 | 83.48 | 23.18 | 4.24 |
| GPT-4o | Full Article | 9.90 | 37.06 | 17.63 | 24.89 | 37.52 | 83.64 | 24.12 | 4.58 |

Table 11: Recategorization of the 72 subjects in MMSci dataset for recruiting Phd experts of each major category from Prolific platform.

| Re-categoried Fields | Original Subjects from Nature Communications |
|---|---|
| Material Science | Materials science, Nanoscience and technology |
| Chemistry | Chemistry |
| Physics | Physics, Optics and photonics |
| Engineering | Engineering |
| Energy | Energy science and technology, Energy and society |
| Mathematics and Computing | Mathematics and computing |
| Astronomy and Planetary Science | Astronomy and planetary science, Planetary science, Space physics |
| Environment | Ecology, Environmental sciences, Biogeochemistry, Water resources |
| Climate Sciences | Climate sciences |
| Earth | Solid Earth sciences, Ocean sciences, Natural hazards, Hydrology, Limnology, Geography |
| Social Sciences | Environmental social sciences, Psychology, Social sciences, Scientific community, Developing world |
| Biochemistry | Biochemistry, Molecular biology, Biophysics, Structural biology, Chemical biology |
| Biological Sciences | Microbiology, Genetics, Biological techniques, Computational biology and bioinformatics, Developmental biology, Evolution, Plant sciences, Physiology, Systems biology, Zoology, Cell biology |
| Biomedical Sciences | Neuroscience, Immunology, Biotechnology, Stem cells, Pathogenesis, Biomarkers, Anatomy, Molecular medicine |
| Health and Medicine | Cancer, Diseases, Medical research, Health care, Oncology, Cardiology, Gastroenterology, Endocrinology, Neurology, Risk factors, Rheumatology, Nephrology, Signs and symptoms, Urology, Health occupations |
| Pharmacology | Drug discovery |
| Agriculture | Agriculture, Forestry |
| Business and Industry | Business and industry |

### A.2.4 Visual Supervised Fine-tuning

We fine-tuned the Qwen2-VL-2B model on our dataset for one epoch with LoRA (Hu et al., 2021), targeting all linear modules. We use the LLAMA-Factory framework for training (Zheng et al., 2024). The hyperparameters are provided in Table 13. The fine-tuning was conducted on a computing cluster with eight NVIDIA A100 GPUs, each with 40GB of memory, and the process took approximately 8 hours to complete.

### A.2.5 Visual Language Pre-training

In our case study experiments on the material generation task, we continuously pre-train a LLaMA2-7B model using our interleaved article and figure data to infuse more material science-relevant knowledge. Specifically, for pre-training on the interleaved text and image data, we follow the methodology outlined in (Lin et al., 2023).

**Model Architecture** Following the approach outlined in (Liu et al., 2024; Lin et al., 2023), we extend the LLaMA2-7B model from a text-only model to a multimodal model by augmenting the LLM with a visual encoder to learn visual embeddings and a projector to bridge the embeddings between the text and visual modalities. Specifically, the visual encoder processes the image and outputs visual features. These features are then mapped into the word embedding space by the projector, creating visual tokens. These visual tokens are concatenated with the word tokens and fed into the LLM, allowing the model to integrate both text and visual information for generation. The specific

Table 12: Quality scores and Phd experts' accuracies across the ten re-grouped fields.

| Field | Fig2Cap | | SubFig2Cap | | SubCap2Fig | |
|---|---|---|---|---|---|---|
| | Quality (1-5) | Accuracies (%) | Quality (1-5) | Accuracies (%) | Quality (1-5) | Accuracies (%) |
| Material Science | 4.0267 | 92.00 | 4.2933 | 92.00 | 4.1333 | 84.00 |
| Chemistry | 4.1333 | 84.00 | 3.7467 | 92.00 | 3.6133 | 100.00 |
| Physics | 4.0267 | 48.00 | 3.5467 | 72.00 | 3.8133 | 80.00 |
| Biochemistry | 3.1600 | 80.00 | 4.8267 | 56.00 | 4.4133 | 72.00 |
| Environment | 4.1067 | 44.00 | 4.4667 | 64.00 | 4.3467 | 76.00 |
| Climate Sciences | 4.1296 | 77.78 | 3.6471 | 88.24 | 3.4118 | 82.35 |
| Earth | 4.0267 | 44.00 | 4.2319 | 52.17 | 4.1739 | 60.87 |
| Biological Sciences | 3.8800 | 48.00 | 3.6800 | 48.00 | 3.7867 | 32.00 |
| Biomedical Sciences | 4.0133 | 68.00 | 4.1333 | 72.00 | 3.7733 | 88.00 |
| Health and Medicine | 4.3733 | 56.00 | 3.7467 | 80.00 | 3.6800 | 52.00 |
| **Average** | **4.0873** | **64.18** | **4.0319** | **71.64** | **3.9149** | **72.72** |

Table 13: Hyperparameters for visual supervised fine-tuning.

| **Hyperparameter** | **Values** |
|---|---|
| base model | `https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct` |
| epochs | 1 |
| global batch size | 8 |
| learning rate | 0.0001 |
| learning rate scheduler | cosine |
| weight decay | 0.0 |
| warmup ratio | 0.1 |
| max length | 4096 |
| lora modules | q_proj, k_proj, v_proj, o_proj, up_proj, gate_proj, down_proj |

Table 14: Hyperparameters for visual language pre-training on interleaved text and image data.

| **Hyperparameter** | **Values** |
|---|---|
| base model | `https://huggingface.co/meta-llama/Llama-2-7b-hfb` |
| vision encoder | `https://huggingface.co/openai/clip-vit-large-patch14-336` |
| projector | 2-layer MLP |
| *Stage 1: Projector Initialization* | |
| epochs | 1 |
| global batch size | 256 |
| learning rate | 0.001 |
| learning rate scheduler | cosine |
| weight decay | 0.0 |
| warmup ratio | 0.03 |
| max length | 4096 |
| tune LLM | ✗ |
| tune vision encoder | ✗ |
| tune projector | ✓ |
| *Stage 2: Visual Language Pre-training* | |
| epochs | 1 |
| global batch size | 128 |
| learning rate | 0.00005 |
| learning rate scheduler | cosine |
| weight decay | 0.0 |
| warmup ratio | 0.03 |
| max length | 4096 |
| tune LLM | ✓ |
| tune vision encoder | ✗ |
| tune projector | ✓ |

LLM, visual encoder, and projectors used in our experiments are presented in Table 14.

**Training Stages** The visual pre-training process (Lin et al., 2023) involves two stages:

1. **Projection initialization**: In this stage, the LLM and the visual encoder are both pre-trained and remain fixed. The projector, however, is randomly initialized. Only the pro-

jector is fine-tuned during this stage, using image-caption pairs from (Liu et al., 2024).

2. **Visual language pre-training**: During this stage, both the LLM and the projector are fine-tuned on the interleaved image and text data. This includes data from general domains provided by MMC4 (Zhu et al., 2024), as well as scientific articles and figures from our dataset MMSci. Previous research (Lin et al., 2023) has shown that tuning both the LLM and the projector yields better results than tuning only one of them. Throughout this stage, the visual encoder remains fixed.

We did not conduct the further visual instruction-tuning for this model, as our primary objective was to infuse scientific knowledge into the LLM for the consecutive text-only material generation task. The two stages were conducted on a computing cluster equipped with eight NVIDIA A100 GPUs, each with 40GB of memory. The first stage took approximately 4 hours, and the second stage took around 36 hours.

#### A.2.6 Materials Generation

As a case study to investigate whether scientific knowledge has been effectively infused into the LLM (LLaMA2-7B in our experiments) and whether it can enhance performance on material science-related tasks, we follow the methodology from (Gruver et al., 2024) to explore the material generation task. The primary objective is to format material crystal structures into text strings and fine-tuning the LLM to generate stable materials.

**Prompt design** We adhere to the prompt design described in (Gruver et al., 2024). There are two types of prompts in the training data: the generation prompt with one or multiple conditions and infilling prompts, where partial crystal structure strings are masked and the model generates the masked parts. The specific prompt templates are shown below, adapted from (Gruver et al., 2024).

The formula condition as shown above is always included, while other conditions are sampled from the following: formation energy per atom, band gap, energy above hull, and space group number.

**Evaluation** Our evaluations follows (Xie et al., 2021; Gruver et al., 2024), including four key aspects. We reiterate some details here. Structural validity is assessed by ensuring that the shortest distance between any pair of atoms exceeds $0.5$ Å. Compositional validity is evaluated by verifying that the overall charge is neutral, as calculated using SMACT (Davies et al., 2019). Coverage metrics, COV-R (Recall) and COV-P (Precision), measure the similarity between ensembles of generated materials and ground truth materials in the test set. The property distribution metrics quantify the earth mover's distance (EMD) between the property distributions of generated materials and those in the test set, specifically for density ($\rho$, in $\mathrm{g/cm^3}$) and the number of unique elements ($N_{el}$).

Metastability and stability are assessed based on the energy above the convex hull, denoted as $\hat{E}_{\text{hull}}$. Two approaches are employed to estimate $\hat{E}_{\text{hull}}$: M3GNet (Chen and Ong, 2022) and Density Functional Theory (DFT) using the VASP code (Hafner, 2008). For M3GNet, each sample undergoes relaxation using force and stress calculations before evaluating the energy of the final structure. For DFT, relaxation is performed using the VASP code, which provides more accurate results but requires significantly more computational resources. A material is considered metastable by M3GNet if the predicted energy above the hull, $E_{\text{hull}}^{\text{M3GNet}}$, is less than 0.1 eV/atom. Furthermore, if validated by DFT, the material must have $E_{\text{hull}}^{\text{DFT}} < 0.0$ eV/atom to be considered stable. The percentages of such materials are reported over the total 10,000 inferences. We use the Materials Project (Jain et al., 2013) dated 2023-02-07.

**Training Details** Following the approach in (Gruver et al., 2024), we utilize 4-bit quantization (Dettmers et al., 2021) and Low-Rank Adapters (LoRA) (Hu et al., 2021) for efficient fine-tuning. The model is trained with a batch size of 1 for 1 epoch. We set the LoRA rank to 8 and the LoRA alpha to 32. The learning rate is 0.0001, annealed by a cosine scheduler. The training was conducted on a single NVIDIA A100 GPU, took approximately 4 hours to complete.

**Conditional Generation and Infilling Results** Due to space constraints, we did not include the results for the conditional materials generation and infilling tasks in the main paper. Here, we present these additional findings. The performance metrics reported are based on the same model used in the main paper. Our training data included two types of prompts: conditional generation prompts and infilling prompts. We compare our model LLaMA2-7B$_{\text{MMSci}}$, which has undergone continuous pre-training, with the original LLaMA2-7B

| **Generation Prompt** | **Infilling Prompt** |
|---|---|
| <s>Below is a description of a bulk material. [The chemical formula is Pm2ZnRh]. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice:<br><br>[ Crystal string ]</s> | <s>Below is a partial description of a bulk material where one element has been replaced with the string “[MASK]”:<br><br>[ Crystal string with [MASK]s ]<br><br>Generate an element that could replace [MASK] in the bulk material:<br><br>[ Masked element ]</s> |

Blue text is the condition for generation. Purple text stands in for string encodings of atoms.

Table 15: Evaluation of conditional materials generation and infilling tasks. Comp. Div. and Struct. Div. represent the composition and structure diversity, respectively. The two models are fine-tuned with the same training data and setup in our implementation.

| **Method** | **Conditional Generation** | | | **Infilling** | | |
|---|---|---|---|---|---|---|
| | Formula↑ | Space Group↑ | $E_{\text{hull}}$ ↑ | Comp. Div.↑ | Struct. Div. ↑ | Metastability ↑ |
| LLaMA2-7B | 0.85 | 0.14 | 0.58 | 10.60 | 0.16 | 64.20% |
| **LLaMA2-7B$_{\text{MMSCI}}$** | 0.87 | 0.22 | 0.59 | 8.31 | 0.52 | 77.74% |

that was trained without additional pre-training data. Both models were trained on datasets that included prompts for both conditional generation and infilling tasks under the same setup.

Following (Gruver et al., 2024), we performed 1,000 inferences for each condition in the conditional generation evaluation and 1,000 inferences for the infilling evaluation. For conditional generation evaluation, we assessed the percentage of generated materials that adhered to specified conditions, including formula, space group, and energy above the hull ($E_{\text{hull}}$). In the infilling evaluation, we measured diversity by computing the pairwise distance between generated samples and those from Matminer (Ward et al., 2018; Xie et al., 2021), focusing on composition and structure. Additionally, we evaluated metastability estimated by M3GNet. As seen in Table 15, LLaMA2-7B$_{\text{MMSCI}}$, after continuous pre-training on our dataset MMSCI, outperforms the original LLaMA2-7B across most metrics. This demonstrates its enhanced effectiveness in handling materials generation tasks.

### A.3 Datasheet

#### A.3.1 Motivation

With the advancement of large language and multimodal models, there is a growing demand for professional AI scientific assistants capable of comprehending and processing advanced, graduate-level scientific knowledge (noa, 2023; White, 2023; Vert, 2023). A crucial aspect of developing effective AI scientific assistants is their ability to understand academic scientific literature, which often includes complex figures such as data visualization plots, charts, schematic diagrams, macroscopic and microscopic photograph, and other specialized content from a variety of scientific fields. However, there is currently a lack of comprehensive evaluation for models' understanding of advanced graduate-level multimodal scientific knowledge, especially in the context of complex figures across diverse scientific disciplines. Existing evaluations tend to focus on simpler charts and plots (Chen et al., 2020; Kahou et al., 2017; Siegel et al., 2016) and suffer from narrow scopes and lower quality (Li et al., 2024).

Our dataset, MMSci, is designed to address this gap. MMSci is a multimodal, multi-discipline dataset comprising high-quality, peer-reviewed articles and figures from 72 scientific disciplines, predominantly within the natural sciences. We created a benchmark to evaluate models' understanding of graduate-level multimodal scientific knowledge across these disciplines. Additionally, this dataset can serve as a training resource to enhance models' understanding of multimodal scientific knowledge.

#### A.3.2 Intended Use

This dataset is used to evaluate and enhance the large multimodal models (LVLMs)' understanding of advanced multimodal scientific knowledge.

#### A.3.3 Data Collection

**Data Source** The dataset comprises open-access articles published in Nature Communications[8]. These articles are freely and permanently accessible upon publication under the Creative Commons Attribution 4.0 International (CC BY) License. Detailed information on the open-access policy of Nature Communications is available at `https://www.nature.com/ncomms/open-access`.

**Data Collection Process** We collected various types of information for each article from the Nature Communications website. The articles' information includes titles, abstracts, main body content, references, and PDF versions of the articles, all directly accessible from their respective sections on the article's webpage (e.g., `https://www.nature.com/articles/xxx`, where "xxx" is the article's unique ID). Additionally, figures and their captions were sourced from a dedicated figures section linked from each article's main page (e.g., `https://www.nature.com/articles/xxx/figures`). This user-friendly platform facilitates easy acquisition of all necessary data, eliminating the needs for quality control and data filtering.

**Annotations** The dataset does not include explicit annotations. Instead, the authors themselves carried out a small-scale manual review and classification of the image types specifically for analysis. No external annotators or crowdworkers were involved in this process.

**Personal and Sensitive Information** The dataset does not include any personal or sensitive information. All article content is publicly accessible. All author information are also publicly available, and no personal information was explicitly extracted, stored, or used from the authors.

#### A.3.4 Social Impact and Ethical Considerations

**Benefits** The benefits of our dataset are two-fold: (1) **Evaluation Benchmark**: This dataset serves as a valuable evaluation benchmark for assessing the understanding of large multimodal models (LVLMs) regarding scientific articles and figures. (2) **Training Resources**: It can be used as a training resource to enhance LVLMs' understanding of scientific articles and figures, improving their performance in various scientific and research-related tasks.

**Risks and Ethical Considerations** However, there are potential risks and ethical considerations to address: (1) **Misuse in Academic Integrity**: The advancement of AI research assistants facilitated by this dataset could potentially lead to misuse, such as academic fraud, fabrication, or improper assistance in academic work. We strongly encourage users to exercise caution and responsibility when using AI assistants, ensuring they are employed ethically and correctly. (2) **Data Misinterpretation and Hallucination**: There is a risk

[8] `https://www.nature.com/ncomms/`

of misinterpreting the dataset's content, leading to inaccurate conclusions or misuse of scientific information. Users should critically assess and validate the AI-generated outputs against established scientific knowledge and principles.

#### A.3.5 Limitations

Our dataset MMSCI provides a comprehensive multimodal dataset across 72 scientific disciplines and serves as both a benchmark and a training resource. However, there are some limitations in our current exploration. (1) Due to limited resources, we were unable to evaluate a wide range of large-scale open-source LVLMs. (2) Our benchmark primarily assesses models' understanding of scientific figures using the figures and captions. The dataset still provide other valuable resources that could be used to create additional tasks, such as single- and multimodal questions aimed at evaluating models' scientific knowledge. We plan to explore these opportunities in future work. Despite these limitations, we believe MMSCI will be a valuable resource for the research community. All data will be made publicly available.

#### A.3.6 Author Statement

The authors declare full responsibility for any rights violations, including but not limited to intellectual property rights and privacy rights, that may arise from the publication and use of this dataset. We confirm that all data provided is licensed under appropriate licenses, ensuring legal compliance and transparency.

#### A.3.7 Use of AI Assistant

We utilized AI to assist with manuscript polishing and improving the presentation of experimental results.

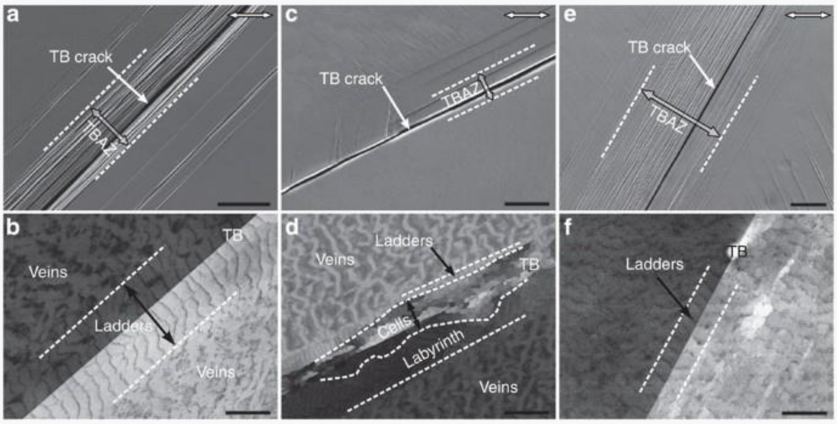
#### A.3.8 Hosting, Licensing, and Maintenance Plan

The dataset will be hosted on GitHub, offering reliable and secure access. We commit to maintaining the repository with regular updates, security patches, and user support to ensure the data's integrity and usability over time. Licensing terms will be clearly communicated to users, adhering to the appropriate data licenses to promote proper usage and distribution. The data is licensed under the CC BY 4.0 License, which permits sharing and adaptation with proper attribution. The primary codebase for our project is licensed under the Apache 2.0 License.

### A.4 Examples

We present several figures as our case study to illustrate multiple-choice questions under three setting in Figure 7, 8, 9, respectively.

Materials science




source: https://www.nature.com/articles/ncomms4536/figures/3

**Question:** Which of the following captions best describes the **whole figure**?
**Options:**
(A): The orientations and primary slip systems of the specimens.
(B): General fatigue cracking mechanisms of the twinned bicrystals with all possible TB inclinations (0° to 90°).
(C): Surface morphologies and dislocation arrangements of the bicrystals with a TB parallel or perpendicular to the loading direction.
(D): Surface morphologies and dislocation arrangements of the bicrystals with an inclined TB.

**Ground-truth: (D)**

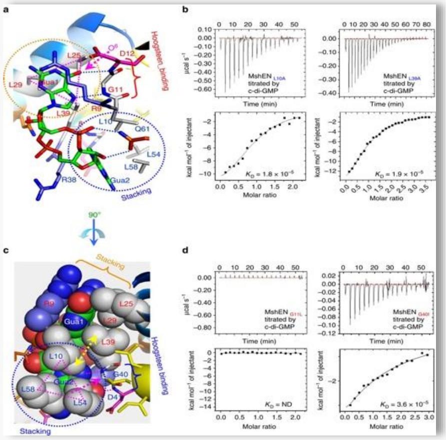
**Claude-3.5-Sonnet: (C)**
**Explanation:** Let's analyze this figure carefully:
1. The figure consists of 6 images labeled a through f.
2. Images a, c, and e appear to be surface morphology observations, possibly from a microscope. They show cracks and structures labeled as "TB crack" (TB likely stands for Twin Boundary).
3. Images b, d, and f seem to be dislocation arrangements, showing features like veins, ladders, and labyrinths.
4. The orientations of the structures in each pair of images (a-b, c-d, e-f) are consistent with each other, suggesting they represent the same specimen but with different imaging techniques.
5. The orientations of the TB (Twin Boundary) structures vary across the image pairs:
- In a-b, the TB is nearly vertical (parallel to the loading direction indicated by the arrow)
- In c-d, the TB is inclined at an angle
- In e-f, the TB is nearly horizontal (perpendicular to the loading direction)
6. This variation in TB orientation suggests that the figure is showing different specimens with varying TB inclinations.
7. The presence of surface morphologies and dislocation arrangements for each orientation supports option C or D.
8. However, the inclusion of specimens with TB parallel, inclined, and perpendicular to the loading direction matches more closely with option C.
Given this analysis, option C appears to be the most comprehensive and accurate description of the entire figure.

**Gemini-1.5-Pro: (D)**
**Explanation:** Let's break down the image and the answer choices:
- **The Image:** Shows microscopic images of material surfaces. We see features like "TB cracks," "TBAZ," and different patterns labeled "veins," "cells," "ladders," "labyrinth."
- **Focus:** The images seem to focus on the areas **around** the TB (likely Twin Boundary) and how the material structure changes.

Now let's analyze the answer choices:
- **A: The orientations and primary slip systems of the specimens.** While orientation might be implied, we don't see direct representation of slip systems.
- **B: General fatigue cracking mechanisms of the twinned bicrystals with all possible TB inclinations (0° to 90°).** The images seem to show the *results* of fatigue, not the mechanisms themselves. Additionally, we don't see the full range of inclinations (0° to 90°).
- **C: Surface morphologies and dislocation arrangements of the bicrystals with a TB parallel or perpendicular to the loading direction.** This is close, but the images clearly show inclined TBs, not just parallel or perpendicular.
- **D: Surface morphologies and dislocation arrangements of the bicrystals with an inclined TB.** This aligns perfectly. We see the surface features ("morphologies") like cracks and patterns, implying dislocation arrangements around the inclined TB.

Figure 7: An example of the multi-choice questions (**Fig2Cap**). The example is within the material sciences subject, sourced from (Kang et al., 2014). The options include the correct main caption of the given figure and three main captions from other figures within the same article.

**Biochemistry**



source: https://www.nature.com/articles/ncomms12481/figures/3

**Question:** Which of the following options best describes the content in **sub-figure (a)**?

**Options:**

(A) Single amino acid change of Leu10 or Leu39 to Ala causes considerable reduction of c-di-GMP-binding affinity as measured by ITC.

(B) Single amino acid change of the conserved middle Gly residue in the loops connecting helices α1−α2 and α3−α4 to Leu or Ile also causes considerable reduction of c-di-GMP binding.

(C) Positions of the residues in the 9RLGD12 sequence relative to the c-di-GMP molecule. Specific binding of c-di-GMP is due to the pairing of its O6 atom (indicated by a magenta arrow) with the Asp12 amide proton. Replacement of the O6 atom with NH22 of c-di-AMP would cause a significant steric clash.

(D) A unique hydrophobic triangular stacking of three conserved Leu residues with the guanine base (drawn as spheres). Leu10-Leu54-Leu58 residues form a triangular stack with Gua2 base. Similarly, Leu25-Leu29-Leu39 residues form another triangular stack with Gua1 base.

**Ground-truth: (C)**

**Claude-3.5-Sonnet: (D)**

**Explanation:** Let me analyze the image and the given options carefully:
The image contains four subfigures labeled (a), (b), (c), and (d). We're specifically asked about subfigure (a).

Subfigure (a) shows a three-dimensional molecular structure. It appears to be a protein or peptide structure with various colored elements representing different amino acids or structural components. There are several labeled residues, including some leucine (Leu) residues.

Looking at the options:
(A) This option discusses ITC measurements, which are not shown in subfigure (a).
(B) This option talks about specific amino acid changes, which are not visually represented in subfigure (a).
(C) This option mentions a specific sequence (9RLGD12) and discusses interactions with c-di-GMP, which are not clearly visible in subfigure (a).
(D) This option describes a "unique hydrophobic triangular stacking of three conserved Leu residues with the guanine base". This matches what we can see in subfigure (a) - there are indeed multiple labeled Leu (Leucine) residues visible, and they appear to be arranged in a way that could be described as triangular stacking around certain parts of the structure.

Based on this analysis, option D seems to best describe what is visually represented in subfigure (a).

**GPT-4o: (D)**

**Explanation:** To determine which option best describes the content in sub-figure (a), let's break down the provided options and analyze the visual content of sub-figure (a).

1. **Option (A)**: This describes a single amino acid change of Leu10 or Leu39 to Ala causing a considerable reduction of c-di-GMP-binding affinity, measured by ITC.
2. **Option (B)**: This describes a single amino acid change of the conserved middle Gly residue in the loops connecting helices α1−α2 and α3−α4 to Leu or Ile, causing a reduction in c-di-GMP binding.
3. **Option (C)**: This describes the positions of residues in the 9RLGD12 sequence relative to the c-di-GMP molecule, with specific binding details and steric clashes.
4. **Option (D)**: This describes a unique hydrophobic triangular stacking of three conserved Leu residues with the guanine base, forming specific triangular stacks.

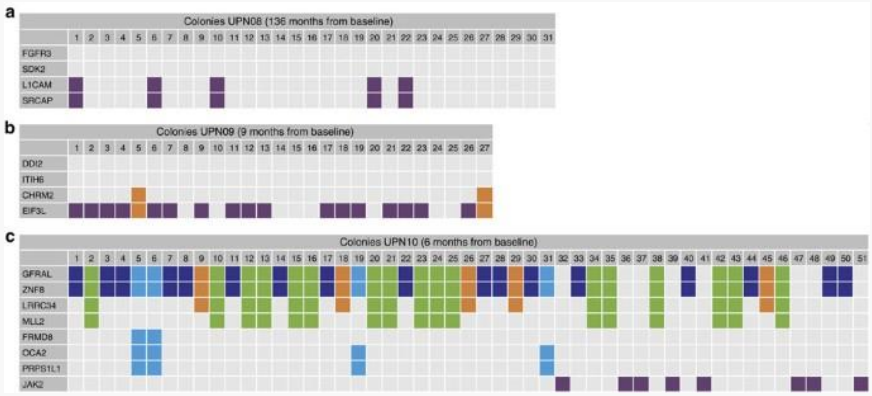
Sub-figure (a) contains a detailed molecular model with labeled amino acids and a c-di-GMP molecule. The visual representation appears to highlight specific interactions between the amino acids and the c-di-GMP molecule, potentially indicating important binding interactions.

Upon close inspection:
- The arrows and labels in sub-figure (a) show specific interactions of the Leu residues with the guanine base.
- The depiction of multiple residues forming a triangular stack with the guanine base fits the description of a hydrophobic interaction network.

Therefore, the visual content described in sub-figure (a) aligns best with the description in Option (D).

Figure 8: An example of the multi-choice questions (**SubFig2Cap**). The example is within the biochemistry subject, sourced from (Wang et al., 2016).

**Cancer**




source: https://www.nature.com/articles/ncomms15099/figures/4

**Question:** Which of the following sub-figure best describes this content: **UPN08: only colonies harbouring the two mutations linked to the unrelated clone are found at this time point. The two investigated mutations from the MDS clone are absent in these colonies.**
**Options:**
(A): sub-figure a
(B): sub-figure b
(C): sub-figure c

**Ground-truth: (A)**

**GPT-4V: (C)**
**Explanation:** Analyzing the problem description, it indicates that we are looking for a sub-figure where the colonies only have two specific mutations linked to an unrelated clone and the mutations from the MDS clone are absent.

In sub-figure C, it is evident that each colony represented by numbered columns (1 through 50) shows a variety of mutations across different genes (represented in rows). Notably, there are no recurring patterns or exclusive presence of only two specific mutations across all colonies that would suggest they are from an unrelated clone, indicating the absence of other specific mutations (e.g., from the MDS clone).

Comparatively, sub-figures A and B, which represent different time points and possibly different mutation patterns, do not match the description provided. Sub-figure A shows many colonies with the same mutations across all sampled points, which might suggest a different clone or condition. Subfigure B also does not fit as it shows a similar variety but at a different time point.

Given this analysis, sub-figure C is the one that best fits the description of only having colonies with mutations linked to an unrelated clone at the specified time point, with the absence of MDS clone mutations.

**GPT-4o: (A)**
**Explanation:** To determine which sub-figure best matches the content description "UPN08: only colonies harbouring the two mutations linked to the unrelated clone are found at this time point. The two investigated mutations from the MDS clone are absent in these colonies," we need to analyze the data presented in each sub-figure:
1. **Sub-figure a**: This figure shows colonies from UPN08 at 136 months from baseline. It indicates the presence of mutations (purple squares) in specific genes. Only two mutations (FGFR3 and SRCAP) are shown, which might be linked to an unrelated clone.
2. **Sub-figure b**: This figure shows colonies from UPN09 at 9 months from baseline. It indicates the presence of mutations in specific genes. This does not match the description for UPN08.
3. **Sub-figure c**: This figure shows colonies from UPN10 at 6 months from baseline. It indicates the presence of mutations in several genes. This does not match the description for UPN08.

Based on this analysis, sub-figure a best fits the description as it corresponds to UPN08 and shows only two mutations, which are linked to the unrelated clone, while the mutations from the MDS clone are absent.

Figure 9: An example of the multi-choice questions (**SubCap2Fig**). The example is within the cancer subject, sourced from (da Silva-Coelho et al., 2017).