Multimodal ArXiv: A Dataset for Improving Scientific Comprehension of Large Vision-Language Models

Lei Li[*†], **Yuqi Wang**[*†], **Runxin Xu**[‡], **Peiyi Wang**[‡]
Xiachong Feng[†], **Lingpeng Kong**[†], **Qi Liu**[†]
[†]The University of Hong Kong
[‡]Peking University
{nlp.lilei, runxinxu, wangpeiyi9979, xiachongfeng1996}@gmail.com
wangyuqi@connect.hku.hk   {lpk, liuqi}@cs.hku.hk

Abstract

Large vision-language models (LVLMs) excel across diverse tasks involving concrete images from natural scenes. However, their ability to interpret abstract figures, such as geometry shapes and scientific plots, remains limited due to a scarcity of training datasets in scientific domains. To fill this gap, we introduce Multimodal ArXiv, consisting of ArXivCap and ArXivQA, for enhancing LVLMs scientific comprehension. ArXivCap is a figure-caption dataset comprising 6.4M images and 3.9M captions, sourced from 572K ArXiv papers spanning various scientific domains. Drawing from ArXivCap, we introduce ArXivQA, a question-answering dataset generated by prompting GPT-4V based on scientific figures. ArXivQA greatly enhances open-sourced LVLMs' mathematical reasoning capabilities, achieving a 10.4% absolute accuracy gain on a multimodal mathematical reasoning benchmark. Furthermore, employing ArXivCap, we devise four vision-to-text tasks for benchmarking LVLMs. Evaluation results with state-of-the-art LVLMs underscore their struggle with the nuanced semantics of academic figures, while domain-specific training yields substantial performance gains. Our error analysis uncovers misinterpretations of visual context, recognition errors, and the production of overly simplified captions by current LVLMs, shedding light on future improvements.[1]

1 Introduction

Large vision-language models (LVLMs), which integrate large language models (LLMs) (Brown et al., 2020a; Touvron et al., 2023) with pre-trained vision encoders through cross-modal alignment training (Madureira, 2021; Liu et al., 2023b; Li et al., 2023d), have demonstrated remarkable perceptual and cognitive capabilities in processing concrete images from everyday scenes (OpenAI, 2023; Fu et al., 2023; Yang et al., 2023a; Reka, 2024). However, recent studies have shown that open-source LVLMs struggle to understand abstract figures, such as geometric shapes in multimodal mathematical reasoning (Lu et al., 2023; Zhang et al., 2024b) and scientific plots (Yue et al., 2023). The inadequacy of training datasets in scientific domains that involve complex reasoning with abstract figures is the main underlying cause.

To address this, we construct Multimodal ArXiv by utilizing the rich resources in preprints hosted on arXiv to improve the ability to understand scientific literature in LVLMs. We first curate ArXivCap, a diverse scientific figure-caption dataset. In contrast to previous scientific figure datasets, which consist of synthesized figures (Chen et al., 2020) or are restricted to simple captioning scenarios in the computer science domain (Hsu et al., 2021), our dataset is composed of figures extracted from academic papers across a range of domains. ArXivCap has 6.4M images and 3.9M captions from 572K papers. We also keep the subfigure structure, and titles of original papers, thereby supporting diverse evaluation tasks. We further instruct GPT-4V to generate 100K multiple-choice question-answering (QA) pairs for the figures in ArXivCap. The resulting ArXivQA dataset could naturally serve as a pivotal resource for improving the scientific reasoning abilities of LVLMs.

We validate the effectiveness of our Multimodal ArXiv dataset from two dimensions: reasoning ability measured by QA accuracy and generation performance through novel vision-to-text tasks. Our experiments demonstrate that ArXivQA brings a significant 10.4% absolute accuracy boost for Qwen-VL-Chat (Bai et al., 2023b), on the Math-Vista (Lu et al., 2023), a challenging benchmark for multimodal mathematical reasoning. Additionally, detailed analysis uncovers the relationship between paper domains and fine-grained task performance.

---

[*]Equal Contribution.
[1]Datasets and models are released at our project page: https://mm-arxiv.github.io.

Moreover, using ArXivCap, we define four generation tasks of varying complexity to benchmark the ability of LVLMs to comprehend scientific plots: (1) captioning a single academic figure, (2) generating overall summaries for multiple sub-figures, (3) in-context figure captioning given previous figure-caption pairs, and (4) generating paper titles from figure-caption pairs. We examine various LVLMs, including open-source models as well as proprietary models including GPT-4V (OpenAI, 2023) and Gemini 1.0 Pro Vision (Gemini Team, 2023). Evaluation results reveal that despite that current LVLMs still face challenges generating faithful captions for scientific figures, in-domain training on our dataset yields substantial performance improvements across all four tasks. Manual error analysis underscores that LVLMs still suffer from misinterpretation of the visual context, recognition errors, and overly simplified captions, paving the way for future studies.

2 Related Work

Recent advancements in LVLMs have seen notable progress in model architecture, training paradigms, and dataset creation (Zhang et al., 2024a).

Model Architecture LVLMs typically comprise three core modules: (i) a vision encoder for image feature extraction, (ii) a modality alignment module to integrate visual features into the language model embedding space, and (iii) an LLM backbone for decoding multimodal context. CLIP (Radford et al., 2021) is widely used for image encoding, while LLaMA (Touvron et al., 2023) and Vicuna (Chiang et al., 2023) serve as popular choices for LLMs. The alignment module varies from simple linear projections (Liu et al., 2023b; Zhu et al., 2023) to more complex architectures like gated cross-attention layers substantiated by Flamingo and IDEFICS (Alayrac et al., 2022; Awadalla et al., 2023). Innovations such as the Q-Former module in BLIP2 (Li et al., 2023b) and instruction integration in InstructBLIP (Dai et al., 2023) further enhance alignment capabilities. Additionally, Fuyu-8B (Bavishi et al., 2023) introduces a novel framework mapping raw image pixels directly to the LLM embedding space.

Training Paradigms Regarding the training recipes, PaLI-X (Chen et al., 2023b) investigates the scaling effects of both vision encoders and language models, highlighting the advantages of scaling both components. Qwen-VL (Bai et al., 2023b) increases input image resolution and explores different module unfreezing strategies. Alignment methodologies such as RLHF training (Ouyang et al., 2022), e.g., LLaVA-RLHF (Sun et al., 2023), and preference optimization through AI feedback (Li et al., 2023c) demonstrate effectiveness in aligning LVLMs with human preferences.

Dataset Curation Dataset quality significantly impacts LVLM performance. Modality alignment training often utilizes web-scale image-caption pairs such as Laion-400M (Schuhmann et al., 2021), with recent studies favoring cleaned captions (Chen et al., 2023a; Yu et al., 2023). Instruction fine-tuning (IFT) helps LVLMs respond according to user queries, triggering the exploration of high-quality IFT datasets. Efforts include multimodal instruction collections such as MultiInstruct (Xu et al., 2023) and M³IT (Li et al., 2023d), dialog-style datasets such as LLaVA (Liu et al., 2023b) and domain-specific datasets for medical (Li et al., 2023a) and text-rich images (Zhang et al., 2023). In the scientific domain, FigCAP (Chen et al., 2019) and FigureQA (Kahou et al., 2017) are created based on synthetic figures. DVQA (Kafle et al., 2018) creates heuristic-based questions for bar charts only. SciCap (Hsu et al., 2021), SciCap+ (Yang et al., 2023b), and M-Paper (Hu et al., 2023) collect figure-caption pairs from specific domains such as computer science. Compared with these datasets, our ArXivCap is sourced from diverse scientific domains with a much larger scale, enabling more comprehensive improvements and evaluations. Besides, we employ GPT-4V for creating ArXivQA with challenging questions, showcasing its effectiveness in boosting the mathematical reasoning ability of LVLMs.

3 Multimodal ArXiv

This section presents a detailed construction process of our Multimodal ArXiv dataset, consisting of two sets: ArXivCap (§3.1) and ArXivQA (§3.2).

3.1 ArXivCap

Construction Process We outline the creation process of ArXivCap below and Figure 1 gives an overview.

Paper Filtering with Publication Type: ArXivCap is extracted from ArXiv (Clement et al., 2019), which is under CC-0 licence for modification and distribution. The raw files of papers posted on
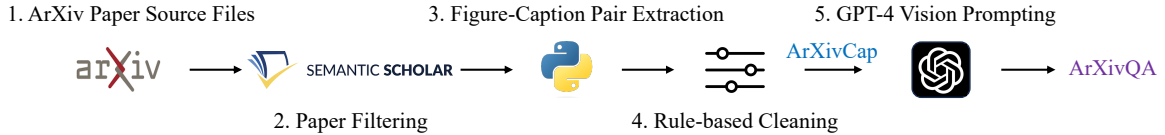
Figure 1: Overview of our dataset curation process. Starting from the ArXiv paper source files, we ensure the paper quality by selecting papers according to publication records. Figure and caption pairs are extracted and then cleaned according to manually designed rules. ArXivQA is generated by prompting GPT-4V with a curated template.
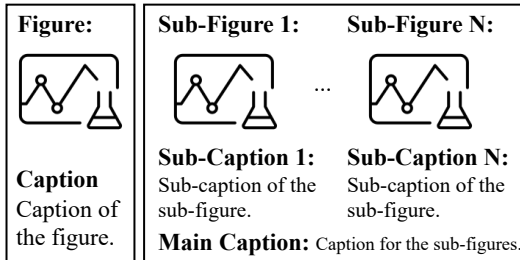
Figure 2: Illustration of two types of figure-caption pairs. (Left) Single-Figure pair. (Right) Multiple-Figure caption pair has multiple sub-figures with corresponding sub-captions and an overall main caption.

Field	Number	Average Len.	Quartile of Len.
Title	572K	10.4	(8, 10, 12)
Abstract	572K	167.6	(126, 165, 207)
Main Caption	3.9M	47.6	(15, 35, 65)
Subcaption	1.0M	4.8	(2, 3, 5)
Chunk Caption	3.9M	48.8	(16, 36, 67)
Images	6.4M	N / A	N / A

Table 1: Word count statistics for title, abstract, captions, and Image number. Chunk caption refers to the combination of subcaptions and the main caption for a multiple-figure case.

ArXiv tar files before June 2023 are downloaded. To ensure the quality of our dataset, we employ a rigorous selection process to filter potentially low-quality papers that might influence the figure-caption pair quality. Firstly, we retrieve meta-information for papers from Semantic Scholar (Kinney et al., 2023), which contains the publication record for each paper. Papers with publication types JournalArticle, Conference, or Review are kept as we assume the peer-review process could ensure the overall figure-caption quality is satisfactory. We further exclude papers with titles exceeding 100 words or abstracts longer than 300 words, in alignment with common submission requirements.

Figure-Caption Pair Extraction: Images and captions are extracted from the original LaTeX files by matching the syntax. We further use a robust tool ImageMagisk (ImageMagick Studio LLC) to convert images into JPEG format for easy processing. The extracted images and captions are stored in a designed chunk structure, which consists of either a single figure-caption pair or multiple figures with their respective sub-captions and a main caption for the overall description. This format is more consistent with the layout of academic papers, and Figure 2 illustrates the chunk structure.

Caption Cleaning and Image Filtering: After a manual inspection of the initially collected dataset, we design several transformations to clean the cap-

tions and filter the images.

Caption Cleaning: (i) Chunks with captions shorter than 5 words are removed; (ii) For captions with LaTeX expressions such as math formulas and references, we apply the pylatexenc[2] to transform the LaTeX to text with math formulas retained, citations to a special symbol <cit.>, references to <ref>. An illustration of caption cleaning can be found in Appendix A.1.

Image Filtering: We remove images that are deemed to be problematic according to the following rules: (i) Images with an aspect ratio larger than 100; (ii) Images with the shortest edge shorter than 224 pixels; and (iii) Images with pixel numbers larger than the decompression bombs threshold.

After these processes, 100 pairs are sampled to perform an additional manual inspection, where we found all of these pairs contained clear images and correct caption descriptions. We provide visualized figure-caption pairs in Appendix A.2.

Statistics of ArXivCap Table 1 lists the dataset statistics. ArXivCap consists of 572K papers, containing 6.4M high-quality images in total with 193M words. A word cloud illustration of captions can be found in the Appendix A.3. Figure 3 demonstrates the paper domain distribution extracted from ArXiv, where we find that our ArXivCap covers 32 domains, such as computer science, mathematics,

---

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

Dataset	Image Number	Paper Number	Image Category	Domain	Real Data
FigCAP (Chen et al., 2020)	219K	N / A	Bar, Line and Pie Charts	N / A	✗
SciCap (Yang et al., 2023b)	2.1M	295K	Open-Category	Computer Science and Machine Learning	✓
M-Paper (Hu et al., 2023)	350K	48K	Open-Category	Mainly "Deep Learning"	✓
ArXivCap (Ours)	6.4M	572K	Open-Category	Open-Domain	✓
FigureQA (Kahou et al., 2017)	140K	N / A	Bar, Line and Pie Charts	N / A	✗
DVQA (Kafle et al., 2018)	300K	N / A	Bar Charts	N / A	✗
ArXivQA (Ours)	32K	16.6K	Open-Category	Open-Domain	✓

Table 2: Comparison with previous scientific figure datasets. Our ArXivCap is the largest captioning dataset and our ArXivQA is the only QA dataset that covers a wide range of domains from real papers.
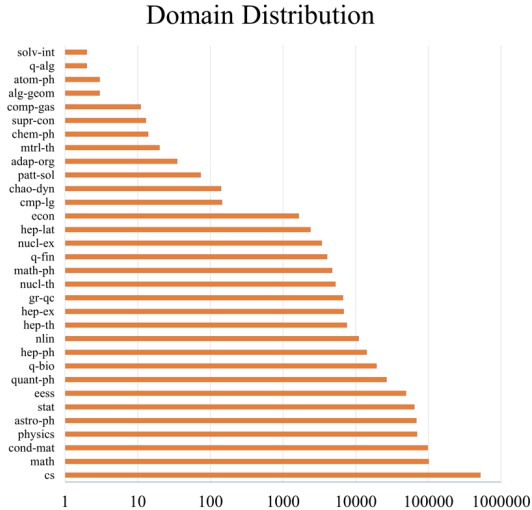
Figure 3: Paper domain distribution of ArXivCap. See Table 10 in Appendix A for the full name of each domain.

Model	Accuracy
InstructBLIP-Vicuna7B	7.0%
LLaVA-1.5-7B	44.2%
LLaVA-1.5-13B	46.8%
OpenFlamingo-9B	9.9%
IDEFICS-Instruct-9B	34.5%
Qwen-VL-Chat	46.6%
Human (100-sample Subset)	80.0%
Human (CS subset)	88.2%

Table 3: Evaluation results on the sampled 1,000 ArXivQA samples.

physics, and economics. As shown in Table 2, compared with previous scientific figure datasets, our ArXivCap is the largest figure-caption dataset collected from real papers and covers a wide range of scientific domains, serving as a valuable resource for improving and benchmarking LVLMs.

3.2 ArXivQA

As our ArXivCap contains diverse images from scientific domains, we assume that learning to answer questions about these figures could boost scientific reasoning ability. Following the successful practice of LLaVA (Liu et al., 2023b), we adopt GPT-4V to generate instruction-tuning datasets for generating the QA pairs based on the figures extracted from scientific papers. Specifically, we design a prompting template to query GPT-4V for generating QA pairs based on 35K images randomly sampled from our ArXivCap. Table 11 in Appendix A.4 provides the template we used for the prompt. The generated pairs are parsed according to the format requirement and we discard the samples without options

and rationales. There are 100K QA pairs after filtering the invalid samples. The dataset comprises questions with an average word count of 16.98 for the question text. On average, there are 4.20 options per question and the average length of the text for a single option is 7.59 words. Appendix A.2 provides samples from the ArXivQA dataset.

As a preliminary study, we sample 1,000 samples from ArXivQA and prompt open-sourced LVLMs to predict answers given the questions and options. A simple prompt is designed to employ GPT-4 for extracting the answer label from the model generations. For human performance, we ask four authors to perform predictions on a 100-sample subset (where 17 samples are from the CS domain). Each of them is asked to answer 50 samples and the accuracy scores are obtained by averaging two annotators. As shown in Table 3, most models struggle to perform satisfactorily on the ArXivQA dataset, falling far behind human performance. This verifies our premise that current open-sourced LVLMs fail to understand scientific figures. We also notice that simply increasing the model scale from 7B (LLaVA-1.5-7B) to 13B (LLaVa-1.5-13B) does not yield a significant boost, which indicates that the ability for multimodal mathematical reasoning cannot be simply acquired from the LLM-side only.

Model	Figure QA	Geometry Problem Solving	Math Word Problem	Textbook QA	Visual QA	ALL
IDEFICS-Instruct-9B[†]	41.4	22.0	18.2	34.6	44.6	33.7
InstructBLIP-Vicuna13B[†]	41.4	19.9	<u>45.5</u>	45.8	57.6	39.3
LLaVa-v1.5-13B[†]	44.0	26.7	40.9	45.8	44.6	39.3
Qwen-VL-Chat-7B	<u>48.3</u>	19.1	22.7	46.7	57.6	40.0
Qwen-VL-Chat-7B$_{ArXivCap}$	39.7	19.8	27.2	39.7	52.1	36.2
Qwen-VL-Chat-7B$_{ArXivQA}$	44.8	34.0	27.3	<u>70.0</u>	<u>64.1</u>	50.2
Qwen-VL-Chat-7B$_{ArXivCap + ArXivQA}$	44.0	37.6	27.3	68.2	63.0	<u>50.4</u>
Bard[†]	38.8	<u>51.1</u>	27.3	64.5	51.1	50.0
GPT-4V[†]	**52.6**	**51.8**	**54.5**	**83.2**	**66.3**	**61.9**

Table 4: Evaluation on MathVista dataset. ArXivCap and ArXivQA together enhance Qwen-VL-Chat's overall performance, surpassing that of the commercial model Bard. [†] denotes results based on the original predictions from Lu et al. (2023). The best results are highlighted in **bold**, while the second-best scores are marked with <u>underline</u>.

4 Experiments

We conduct experiments to (i) validate the effectiveness of ArXivQA for boosting multimodal scientific reasoning for open-source LVLMs (§4.1) and (ii) benchmark LVLMs capability to comprehend scientific figures with ArXivCap (§4.2).

4.1 Boosting LVLMs with ArXivQA

4.1.1 Experimental Settings

We adopt Qwen-VL-Chat-7B (Bai et al., 2023b) as the backbone due to its support for interleaved image-text input formats and high-resolution images. We fine-tune it on our ArXivCap (Qwen-VL-Chat-7B$_{ArXivCap}$), ArXivQA (Qwen-VL-Chat-7B$_{ArXivQA}$) and combination of these two datasets (Qwen-VL-Chat-7B$_{ArXivCap + ArXivQA}$) for three epochs with a learning rate of 1e-5 following the original paper. We combine the answer and the rationale in ArXivQA to form the target output during training. Models are evaluated on MathVista (Lu et al., 2023), a benchmark that requires fine-grained, deep visual understanding and compositional reasoning. MathVista contains 6,141 examples, consisting of five multimodal tasks Figure QA, Geometry Problem Solving, Math word problem, Text Book QA, and Visual QA. We select 478 multiple-choice questions in the `testmini` split to avoid the inconsistency of answer parsing. We compute the accuracy scores and adopt the provided prediction files for calculating the baseline performance.

4.1.2 Results

As shown in Table 4, fine-tuning on our Multimodal ArXiv, especially on the ArXivQA dataset, consistently boosts the performance, helping the open-sourced Qwen-VL-Chat achieve a comparable overall MathVista reasoning performance. Due to the wide coverage of the scientific figures, the performance gain mainly comes from significantly improved Geometry Problem Solving, Textbook QA, and Visual QA tasks. For example, after fine-tuning on the ArXivQA dataset, the accuracy is increased from 19.1% to 34.0% and from 46.7% to 70.0% on Geometry Problem Solving and Textbook QA tasks, respectively. The improvement on Math Word Problem is marginal, where we think the domain-specific data augmentation can be further explored with a curated filtering dataset on our dataset (Gao et al., 2023). On the contrary, the accuracy of Figure QA deteriorates slightly compared with the original backbone model, which we attribute to the fact that most of the plots in the Figure QA evaluation are sampled from synthesized datasets such as DVQA (Kafle et al., 2018), exhibiting great gaps between real-world paper figures.

4.1.3 Analysis

We investigate how different subject domains affect mathematical reasoning ability using pairs of questions and answers (QA). We focus on six domains with more than 5K samples each. From each domain, we randomly choose a subset of 5K samples to ensure fairness in comparison. We then fine-tune the Qwen-VL-Chat base model using QA pairs from each domain and observe how it affects the model's accuracy compared to its original state. Figure 4 demonstrates the relative accuracy changes (i.e., $\frac{\text{Accuracy after Fine-tuning}}{\text{Original Accuracy}} - 1$) after training the model on QA pairs from each domain. Our findings reveal several key points: (i) QA pairs from the Computer Science (CS) domain are highly effective for improving mathematical reasoning ability, achieving a notable 27.09% relative improvement. We attribute this to the compositional nature of the CS area. (ii) The most beneficial do-

main varies depending on the specific task. For instance, QA pairs from astrophysics domains enhance geometry problem-solving, while those from Condensed Matter improve performance in math word problems. (iii) Most domains hurt the Figure QA task. This suggests that synthetic Figure QA might not be the best benchmark for assessing realistic reasoning ability. These findings underscore the efficacy of generated QA pairs and offer valuable insights for future research, such as adjusting task-specific weights in the dataset accordingly.

4.2 Benchmarking LVLMs on ArXivCap

4.2.1 Evaluated Tasks

Four vision-to-text tasks to benchmark LVLMs' ability to comprehend scientific figures.

Single-Figure Captioning Similar to the traditional image captioning setup (Lin et al., 2014), single-figure captioning requires the model to generate a caption for the given single figure. The captions generated by the model are expected to encapsulate the nuanced details within these figures, including numbers and mathematical formulas, presenting a unique challenge for models to identify and articulate these elements accurately. Formally, given an image-caption pair (I, C), the LVLM $\mathcal{M}$ is asked to generate the caption given an instruction prompt P_s to hint the goal of scientific captioning:

$$\hat{C} = \mathcal{M}(I, P_s),$$

where $\hat{C}$ would be evaluated according to the ground-truth C.

Multiple-Figure Captioning We introduce a more intricate challenge involving applying reasoning across multiple images. This task, termed Multiple-Figure Captioning, necessitates the model to craft a comprehensive summary caption for subfigures. As exemplified in Figure 2, the model is tasked with generating an overarching caption for two or more subfigures, leveraging visual clues to draw comparisons and formulate summary captions. Formally, given a list of figures $L = (I_1, \ldots, I_n)$, the model is asked to generate the ground-truth main caption C by considering all the semantics in the figures with a task prompt P_m:

$$\hat{C} = \mathcal{M}(L, P_m) = \mathcal{M}(I_1, \ldots, I_n, P_m).$$

Contextualized Captioning Inspired by the evolving in-context learning capabilities of LLMs (Brown et al., 2020b; Dong et al., 2022),

we introduce a contextualized captioning task to examine the in-context learning ability of LVLMs. In this task, the model is presented with a set of figure-caption pairs, and its goal is to generate a caption for a given image based on the provided demonstrations. Given a sequential image-captions pairs $S = \{(I_i, C_i)\}_{i=1}^n$ consisting of n pairs of image I_i and the corresponding C_i, the contextualized image captioning task can be formalized as follows:

$$\hat{C}_n = \mathcal{M}(I_1, C_1, \ldots, I_{n-1}, C_{n-1}, I_n, P_c).$$

The model is supposed to leverage the context history to enhance the accuracy and coherence of the generated caption.

Title Generation This task requires a nuanced understanding of figures and captions to distill essential observations into a high-level summary of the presented results for LVLMs. Specifically, instead of producing the captions for the figures, this task requires the model to connect different figures and corresponding captions to infer the paper title. Let $S = \{(I_i, C_i)\}_{i=1}^m$ be a sequence of m figure-caption pairs in the extracted paper. Note that I_i could be a single figure or a multiple-figure, and we reuse I_i for simplicity here. The title generation asks $\mathcal{M}$ to generate the title for the paper given a task prompt P_t:

$$\hat{T} = \mathcal{M}(I_1, C_1, \ldots, I_m, C_m, P_t).$$

The prediction $\hat{T}$ is evaluated by comparing it to the original title T.

4.2.2 Experimental Settings

Dataset We divide ArXivCap into training and test sets with a 9:1 ratio for evaluation. The test set includes: 161.3K samples for single-figure captioning, 12.8K samples for multiple-figure captioning, 57.2K samples for contextualized captioning, and 57.2K samples for title generation.

Evaluated Models We select various LVLMs covering different architectures. (1) LVLMs designed for dealing with a single image, BLIP2-OPT-6.7B (Li et al., 2023b) , InstructBLIP-Vicuna7B (Dai et al., 2023), LLaVA-1.5-7B/13B (Liu et al., 2023a). Due to the ability limitation, we only benchmark these models on the single image captioning task; (2) LVLMs capable of handling interleaved text-image inputs, such as OpenFlamingo-9B (Alayrac et al., 2022; Awadalla
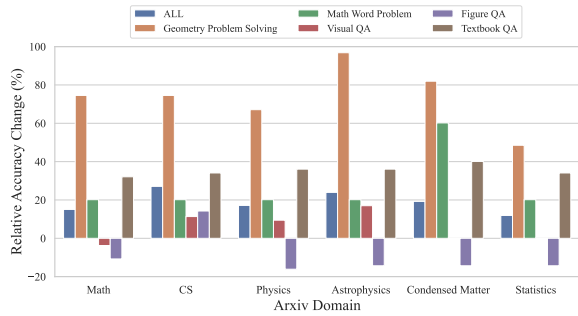
Figure 4: Relative accuracy changes brought by the training on different domain ArXivQA samples.

et al., 2023), IDEFICS-Instruct-9B (Laurençon et al., 2023), Qwen-VL-Chat-7B (Bai et al., 2023b). These models are evaluated on all the tasks we proposed; (3) Proprietary models such as Gemini 1.0 Pro Vision and GPT-4V. Due to the large scale of our test set, we randomly sample a subset consisting of 500 instances for evaluating these two models to reduce costs, with corresponding scores colored in grey. Details of evaluated models and the task prompts used are provided in Appendix B.

Training Settings To investigate whether in-domain training can enhance the model's capabilities, we train the Qwen-VL-Chat-7B on ArXivCap using the same setting as in §4.1.1. To fit the input length limit, we set the maximum number of figures per sample to four. The training process takes 70 hours with 8 NVIDIA A100s.

Metrics BLEU-2 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and BERT-Score (Zhang et al., 2020) are adopted as the automatic evaluation metrics. We also explore using GPT-4 to assist in caption evaluation. Our findings in Appendix B.3 indicate that ROUGE-L and BLEU-2 scores are highly correlated with GPT-4's annotations. We primarily use these three metrics due to their convenience. A manual error analysis is conducted to supplement the automatic metrics (§4.3).

4.2.3 Results

Results of Single-Figure Captioning The evaluation results for the single-figure captioning task are presented in Table 5. Despite achieving near-perfect performance on conventional image captioning tasks like MSCOCO (Lin et al., 2014), open-source LVLMs, such as LLaVA models, face challenges when applied to academic figures. For closed models, GPT-4V performs comparably with

Model	BLEU-2	ROUGE-L	BERT-S
BLIP-2-OPT-6.7B	2.1	7.1	81.1
InstructBLIP-Vicuna7B	3.7	10.1	83.3
LLaVA-v1.5-7B	2.3	10.6	83.0
LLaVA-v1.5-13B	2.6	10.7	83.3
OpenFlamingo-9B	5.7	9.9	82.4
IDEFICS-Instruct-9B	2.5	9.1	83.5
Qwen-VL-Chat-7B	4.4	11.1	81.8
Qwen-VL-Chat-7B$_{ArXivCap}$	**8.9**	**15.8**	**83.3**
Gemini 1.0 Pro Vision	5.6	14.5	82.2
GPT-4V	5.5	14.2	83.3

Table 5: Evaluation results of single figure captioning. Grey results are obtained from a 500-sample subset. Despite most LVLMs struggle to produce high-quality captions of scientific figures, training with ArXivCap significantly boosts the performance.

Model	BLEU-2	ROUGE-L	BERT-S
Qwen-VL-Chat-7B	4.4	11.1	81.8
+ Title	5.7	13.1	81.6
+ Title and Abstract	6.0	12.7	81.4
Qwen-VL-Chat-7B$_{ArXivCap}$	8.9	15.8	83.3
+ Title	**12.9**	**18.6**	**83.8**
+ Title and Abstract	12.7	18.5	83.8

Table 6: Evaluation results of single figure captioning with paper meta information.

Gemini 1.0 Pro Vision. Furthermore, continuous training on our dataset yields a significant performance boost for this task. For instance, fine-tuning results in a notable increase in the BLEU-2 score from 4.4 to 8.9, indicating a promising avenue for enhancing academic figure comprehension through domain-specific training. We also investigate whether providing additional context information, such as the paper title and abstract, could help models generate better figure captions. As shown in Table 6, adding the title is beneficial evidenced by the boosted scores, while providing abstracts brings negligible gains.

Results of Multiple-Figure Captioning As shown in the first block of Table 7, similar to single-figure captioning, multiple-image captioning poses a challenge for current open-source LVLMs. For instance, Qwen-VL-Chat achieves only a 3.0 BLEU-2 and a 7.2 ROUGE-L score on this task, considerably lower than its performance in single-figure captioning. In contrast, GPT-4V consistently demonstrates proficiency in both tasks, suggesting a balanced ability to capture semantics across multiple images. Notably, training on our ArXiv-Cap dataset yields more pronounced improvements

Model	Multiple-Figure Captioning			Contextualized Captioning			Title Generation		
	BLEU-2	ROUGE-L	BERT-S	BLEU-2	ROUGE-L	BERT-S	BLEU-2	ROUGE-L	BERT-S
OpenFlamingo-9B	3.7	11.3	81.9	20.0	20.5	83.7	2.7	17.7	82.7
IDEFICS-Instruct-9B	3.6	10.8	82.8	**20.7**	**22.6**	**85.7**	3.5	18.4	85.8
Qwen-VL-Chat-7B	3.0	7.2	79.7	17.0	22.1	85.0	2.6	15.8	85.1
Qwen-VL-Chat-7B$_{ArXivCap}$	**10.6**	**18.0**	**83.6**	16.1	21.2	84.8	**6.7**	**23.5**	**86.8**
Gemini 1.0 Pro Vision	6.1	16.2	83.1	10.2	20.2	84.5	5.7	21.8	85.9
GPT-4V	5.7	14.7	83.0	9.6	20.1	84.7	4.0	20.2	86.0

Table 7: Evaluation results of three newly defined tasks. The best results are highlighted in **bold**.

Model	BLEU-2 ($\Delta \downarrow$)	ROUGE-L ($\Delta \downarrow$)
Qwen-VL-Chat-7B	17.0	22.1
+ random contexts	5.7 (66.5%)	13.0 (38.1%)
+ shuffle order	12.0 (29.4%)	15.1 (31.7%)
Qwen-VL-Chat-7B$_{ArXivCap}$	16.1	21.2
+ random contexts	7.5 (53.4%)	14.3 (32.5%)
+ shuffle order	14.1 (12.4%)	19.5 (8.0%)

Table 8: Contextualized captioning performance is influenced by the order. After tuning on the ArXivCap, the model is more robust to the order of the history captions.

for this task, culminating in Qwen-VL-Chat even surpassing the performance of the GPT-4V model. This enhancement underscores the pivotal role of our dataset in facilitating LVLMs to enhance reasoning capabilities over multiple images, leading to more effective summarization of scientific figures.

Results of Contextualized Captioning In the middle block of Table 7, we find that IDEFICS-Instruct-9B achieves the best performance on this task. This achievement is largely attributed to its remarkable proficiency in leveraging contextual cues, stemming from its extensive pre-training involving interleaved image-text pairs (Laurençon et al., 2023). Interestingly, fine-tuning on ArXivCap results in marginal performance declines across all metrics, with GPT-4V achieving the lowest scores as well. This phenomenon can be attributed to the tendency of sequential captions to exhibit similar patterns, thereby favoring models that effectively leverage contextual cues. We perform two more challenging evaluations by (i) providing context pairs from another paper and (ii) randomly shuffling the order of figure-caption pairs in the context. As shown in Table 8, the performance with random contexts degrades significantly, validating our previous hypothesis. Instead, the fine-tuned model demonstrates more robust captioning results under these settings, evidenced by the slight 8% drop on ROUGE-L compared to the 31% of the original model with shuffled context orders.
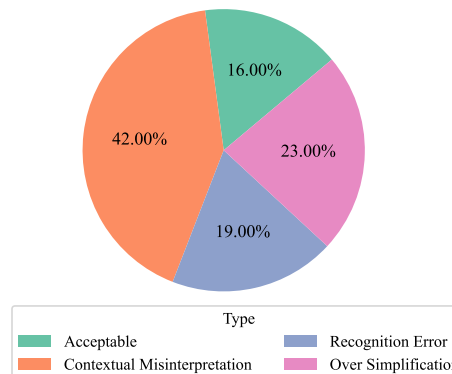
Figure 5: Manual analysis of the generated captions.

Results of Title Generation The results are presented in the last block of Table 7. Notably, the title generation task poses a formidable challenge, evident in the significantly lower overall BLEU-2 score compared to the captioning tasks. This suggests the inherent difficulty in generating precise predictions for paper titles. A contrasting picture emerges when considering the ROUGE-L and BERT-Score metrics, which either closely align or surpass the performance on captioning tasks. This underscores the model's proficiency in producing semantic-related results given the presented figures. Consistent with the previous two tasks, fine-tuning the model on our dataset yields substantial enhancements for the title generation task. The BLEU-2 score jumps impressively from 2.6 to 6.7, while the ROUGE-L score sees a commendable increase from 15.8 to 23.5. These findings highlight the challenge of title generation for current LVLMs and the effectiveness of our dataset in improving the model's capability to generate accurate titles.

4.3 Analysis

Manual Evaluation of Generated Captions We conduct a manual inspection for single-figure captioning results. To ensure a more informed evaluation, we focus on a paper from the CS domain, leveraging our domain knowledge to assess caption quality better. The quality of generated captions
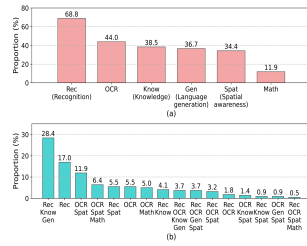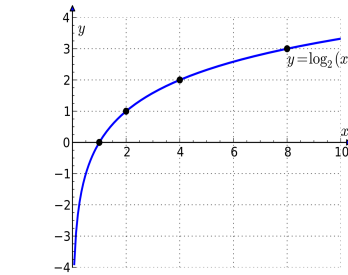
Figure 6: ArXivQA enables the model not only to answer questions related to scientific figures in papers (left) but also to improve mathematical understanding ability (right). The model not only selects correct options but also gives reasonable rationale.

is assessed by scrutinizing the figure, the ground-truth caption, the paper title, and the abstract. We categorize captions into the following quality types according to our preliminary inspection: (1) *Acceptable*, where captions accurately encapsulate the scientific figure's essence, aligning with the intended information of the ground-truth; (2) *Over Simplification*, instances where the model oversimplifies content, offering a broad overview while neglecting specific details and nuances present in the ground truth; (3) *Recognition Error*, where the model inaccurately recognizes and describes key visual and textual elements in the scientific figure, such as colors, numerical values, or textual context; and (4) *Contextual Misinterpretation*, where the model misinterprets the specific context of the scientific figure, resulting in captions relevant in a generic sense but inaccurate for the given figure. Visualized generated captions of different types are shown in Figure 14 of Appendix C.1. The results of 100 manually examined captions are depicted in Figure 5, revealing that only 16% of captions are deemed acceptable when compared to human-written ones. Among unsatisfactory captions, contextual misinterpretation emerges as the dominant issue, suggesting a need for incorporating more contextual information as suggested in Table 6. Oversimplification is another concern, with generic captions identified. Additionally, 23% and 19% of examined captions suffer from the oversimplification issue and recognition errors in reported numbers/texts in the caption, respectively. The former is attributed to the highly frequent simple caption in the training dataset and the latter issue could be

addressed through potential integration with OCR results. Our manual evaluation suggests future efforts may benefit from incorporating additional context clues, such as paper metadata, improving the model's fundamental perception abilities, and utilizing external information.

Case Study of MathVista We conduct case studies to illuminate the tuning effects facilitated by our ArXivQA dataset. In the left segment of Figure 6, ArXivQA helps the model accurately answer a question related to the presented bar plot. The right part in Figure 6 demonstrates that ArXivQA can enhance algebraic reasoning abilities. Here, a question involving the derivative of a function is correctly answered, accompanied by a lucid reasoning rationale. Figure 15 in Appendix C.2 highlights a challenging geometry problem where both models generate hallucinated outputs. These illustrative cases collectively affirm the efficacy of our dataset.

5 Conclusion

Our work introduces Multimodal ArXiv, comprising ArXivCap and ArXivQA, aims at advancing the scientific comprehension of LVLMs. Experiments show that fine-tuning on ArXivQA notably enhances LVLMs' mathematical reasoning capabilities. Moreover, our comprehensive evaluations across four vision-to-text tasks on ArXivCap underscore the challenges in understanding scientific figures for LVLMs, while highlighting the substantial improvements achieved by in-domain training. Our error analysis offers valuable insights for the ongoing development of LVLMs.

Limitations

Our study has several limitations worth noting. Firstly, our exploration may not encompass the full spectrum of LVLMs due to the rapid evolution of architectures and training methodologies such as parameter-efficient tuning (Hu et al., 2022; Ma et al., 2024). Nevertheless, we believe our dataset could still be effective for other LVLMs and the findings are generalizable. We show that our ArXivQA dataset could also boost LLaVA-series models across scientific understanding benchmarks in Appendix D. Secondly, our Multimodal ArXiv dataset sources from ArXiv papers due to their accessibility and open-source licenses. This approach may overlook the diversity of disciplines and data modalities present in the broader scientific literature. Future research could incorporate a broader range of datasets and domains to enrich the coverage of scientific knowledge, and explore dynamic data selection methods to improve performance and sample efficiency (Li et al., 2021; Chen et al., 2024).

Acknowledgements

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv preprint*, abs/2204.14198.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt.

2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv preprint*, abs/2308.01390.

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

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

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. 2019. Figure captioning with reasoning and sequence-level training. *ArXiv preprint*, abs/1906.02850.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *ArXiv preprint*, abs/2311.12793.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *ArXiv preprint*, abs/2402.12501.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel M. Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, A. J. Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim M. Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023b. Pali-x: On scaling up a multilingual vision and language model. *ArXiv preprint*, abs/2305.18565.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Colin B Clement, Matthew Bierbaum, Kevin P O'Keeffe, and Alexander A Alemi. 2019. On the use of arxiv as a dataset. *ArXiv preprint*, abs/1905.00075.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv preprint*, abs/2305.06500.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A survey for in-context learning.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv preprint*, abs/2306.13394.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2023. G-llava: Solving geometric problem with multi-modal large language model.

Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805.

Google. 2023. Bard.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2023. mplug-paperowl: Scientific diagram analysis with the multimodal large language model.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

ImageMagick Studio LLC. Imagemagick.

Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: understanding data visualizations via question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 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*, abs/1710.07300.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023.

Obelics: An open web-scale filtered dataset of inter-leaved image-text documents.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *ArXiv preprint*, abs/2306.00890.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. Dynamic knowledge distillation for pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023c. Silkie: Preference distillation for large visual language models. *ArXiv preprint*, abs/2312.10665.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023d. M^{3}IT: A large-scale dataset towards multi-modal multilingual instruction tuning. *ArXiv preprint*, abs/2306.04387.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.

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 math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv preprint*, abs/2310.02255.

Pan Lu, Swaroop Mishra, Tony 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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Xinyu Ma, Xu Chu, Zhibang Yang, Yang Lin, Xin Gao, and Junfeng Zhao. 2024. Parameter efficient quasi-orthogonal fine-tuning via givens rotation. *ArXiv preprint*, abs/2404.04316.

Brielen Madureira. 2021. Flamingos and hedgehogs in the croquet-ground: Teaching evaluation of NLP systems for undergraduate students. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 87–91.

OpenAI. 2023. Gpt-4v(ision) system card.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

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

Team Reka. 2024. Reka Core, Flash, and Edge: A series of powerful multimodal language models.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv preprint*, abs/2111.02114.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented rlhf. *ArXiv preprint*, abs/2309.14525.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *ArXiv preprint*, abs/2305.17926.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11445–11465.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023a. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *ArXiv preprint*, abs/2309.17421.

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

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. 2023. Capsfusion: Rethinking image-text data at scale. *ArXiv preprint*, abs/2310.20550.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *ArXiv preprint*, abs/2311.16502.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ArXiv preprint*, abs/2403.14624.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding.

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*, abs/2304.10592.

A Details of Multimodal ArXiv

A.1 Caption Cleaning

We apply a Python tool to clean the original caption and Table 9 illustrates the caption before and after cleaning.

Before Cleaning	After Cleaning
A 1995 Hale Telescope Hα image of the Guitar Nebula (20 angstrom filter at 6564 angstroms). The cometary neck connecting to a spherical bubble are clearly evident. Credit: \cite{cha02}.	A 1995 Hale Telescope Hα image of the Guitar Nebula (20 angstrom filter at 6564 angstroms). The cometary neck connecting to a spherical bubble are clearly evident. Credit: <cit.>.
As Fig. \ref{z0} except at $z \sim 6$ ($z = 4.37$ in the EdS model).	As Fig. <ref> except at $z \sim 6$ ($z = 4.37$ in the EdS model).

Table 9: Caption before and after cleaning using pylatexenc.

A.2 Illustration Cases of Multimodal ArXiv

We provide illustrated cases from our dataset for a better understanding. Figure 7 demonstrates a typical single figure-caption pair, and Figure 8 shows the multiple figure-caption case. Figure 10, 11, 12 and 13 illustrate the cases from our ArXivQA dataset, covering different figure types and containing diverse questions.
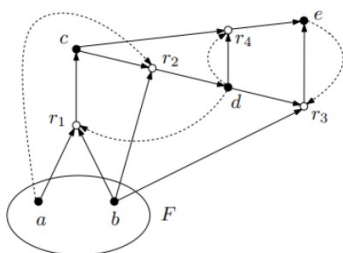
A.3 Caption Word Cloud

We visualize the word cloud of captions in our ArXivCap dataset in Figure 9. It can be seen that the captions have a diverse vocabulary for describing the different figures in the academic papers.

A.4 ArXivQA Prompting Template

The prompt used to query GPT-4V is provided in Table 11.

Figure:

Caption: Example of a graphical representation of a CRS. The CRS consists of five chemicals $X = \{a, b, c, d, e\}$ and four reactions $a + b \rightarrow c, c + b \rightarrow d, b + d \rightarrow e$ and $c + d \rightarrow e$, which are catalyzed by d, a, e and d, respectively. The food set is given by $F = \{a, b\}$.

Figures and captions are from paper arxiv:1908.04642.

Figure 7: A single-figure caption pair in our ArXivCap dataset.

Domain	Full Name
dg-ga	Differential Geometry
acc-phys	Accelerator Physics
solv-int	Exactly Solvable and Integrable Systems
q-alg	Quantum Algebra and Topology
atom-ph	Atomic, Molecular and Optical Physics
alg-geom	Algebraic Geometry
comp-gas	Cellular Automata and Lattice Gases
supr-con	Superconductivity
chem-ph	Chemical Physics
mtrl-th	Materials Theory
adap-org	Adaptation, Noise, and Self-Organizing Systems
patt-sol	Pattern Formation and Solitons
chao-dyn	Chaotic Dynamics
cmp-lg	Computation and Language
econ	Economics
hep-lat	High Energy Physics - Lattice
nucl-ex	Nuclear Experiment
q-fin	Quantitative Finance
math-ph	Mathematical Physics
nucl-th	Nuclear Theory
gr-qc	General Relativity and Quantum Cosmology
hep-ex	High Energy Physics - Experiment
hep-th	High Energy Physics - Theory
nlin	Nonlinear Sciences
hep-ph	High Energy Physics - Phenomenology
q-bio	Quantitative Biology
quant-ph	Quantum Physics
eess	Electrical Engineering and Systems Science
stat	Statistics
astro-ph	Astrophysics
physics	Physics
cond-mat	Condensed Matter
math	Mathematics
cs	Computer Science

Table 10: Name of each domain.

A.5 Quality Analysis of ArXivQA

To evaluate the quality of ArXivQA, we manually assess it from six different aspects. We develop

Table 11: Prompt used for GPT-4V to generate QA pairs based on scientific figures.

a quality examination guideline for annotators, as shown in Table 12, which addresses various aspects of the QA pairs. We sample 100 examples and ask four authors to conduct the quality analysis. The four authors are divided into two groups, with each group tasks with evaluating 50 examples across six sub-aspects, according to the grading protocol.

The evaluation results are presented in Table 13. We find that most samples feature clear, high-quality images with clear and high-quality images, with unambiguous question and option descriptions. However, a small fraction of the generated questions may be unanswerable due to mis-recognizing elements in the figures, as reflected by lower factual alignment scores. Additionally, we consider samples with an aggregate score of 5 or higher from both annotators to be of sufficient quality. Under this stringent criterion, 79 out of 100 samples meet the threshold, demonstrating that the dataset's quality is generally satisfactory.

B Evaluation Details

B.1 Details of Evaluated Models

BLIP2 (Li et al., 2023b), introduces a lightweight Q-Former designed to bridge modality gaps and leverages frozen LLMs. Leveraging LLMs, BLIP-2 can conduct zero-shot image-to-text generation using natural language prompts. We select the BLIP2-OPT-6.7B version for evaluation.[3]

InstructBLIP (Dai et al., 2023) employs an instruction-aware visual feature extraction module based on BLIP2 (Li et al., 2023b) and is trained with unified multimodal instruction tuning datasets.
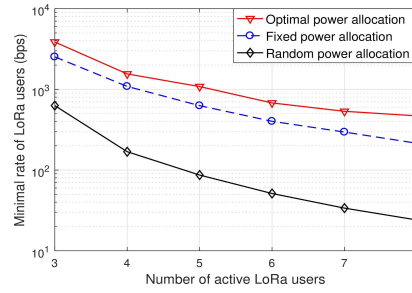
---

[3] https://huggingface.co/Salesforce/blip2-opt-6.7b

Sub-Figure 1:
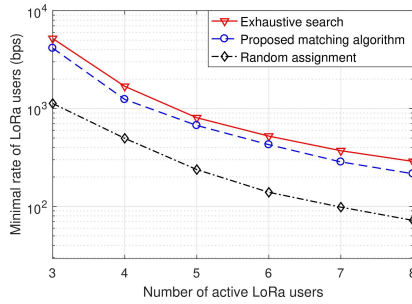
Sub-Figure 2:

Sub-Caption 1: Performance of the proposed low-complexity channel assignment scheme based on matching theory.

Sub-Caption 2: Performance of proposed optimal power allocation scheme for users assigned to the same channel.

Main Caption: Active LoRa nodes are randomly distributed in a circle with the radius of 10 km, center frequency is 868 MHz, the number of available channels is 3 with bandwidth of 125 kHz for each, the transmit power of each LoRa users ranges from 0 dBm to 20 dBm.

The figure and caption are from paper arxiv:1810.10761.

Figure 8: A multiple-figure caption pair in our ArXivCap dataset.

1. Ensuring Factual Integrity and Clear Presentation:
- Factual Alignment: Ensure questions and options are grounded in accurate reflections of the chart data.
- Visual Clarity: Maintain high-resolution charts to ensure that all pertinent details are discernible.
- Unambiguous Textual Information: Employ precise and unambiguous language to formulate questions and answers, thereby mitigating potential misinterpretations.
2. Ensuring Relevance and Integrated Comprehensiveness:
- Question and Option Relevance: Charts must align with their questions, and all options should be applicable and relevant to the given data.
- Comprehensive Integration: Guarantee the provision of comprehensive information necessary for the interpretation of the chart and the resolution of the question, ensuring a cohesive amalgamation of textual and visual data.
3. Promoting Fairness and Avoiding Bias:
- Equitable Content: Strive for impartiality in the dataset to prevent bias and ensure fair representation of diverse groups and perspectives.

Grading Protocol:
Each criterion is to be rigorously evaluated for each dataset entry. The assessment is to be conducted on a qualitative scale with three distinct levels: High, Medium, and Low. These levels will denote the degree of conformity to the respective criterion:
- 1: High: The dataset entry exhibits exemplary adherence to the evaluation criterion, demonstrating a robust and comprehensive alignment with the specified standard.
- 0.5: Medium: The dataset entry meets the evaluation criterion to a moderate extent, indicating a satisfactory but not optimal congruence with the standard.
- 0 Low: The dataset entry falls short of the evaluation criterion, signaling a need for significant improvements to meet the standard.

Table 12: Annotator guideline of ArXivQA manual quality examination.

Figure 9: Word cloud visualization of captions in our ArXivCap dataset.

Aspect	Avg Score
Factual Alignment	0.6975
Visual Clarity	0.9925
Unambiguous Textual Information	0.9825
Question and Option Relevance	0.9375
Comprehensive Integration	0.905
Equitable Content	1.0
Score Sum	5.515

Table 13: Manual quality analysis of ArXivQA. The average scores for each aspect are presented.

We choose `InstructBLIP-Vicuna-7B` for evaluation.[4]

LLaVA (Liu et al., 2023b), adopts Vicuna models as the backbone LLM and is trained on the ChatGPT/GPT-4 generated instruction tun-

---

[4] https://huggingface.co/Salesforce/instructblip-vicuna-7b
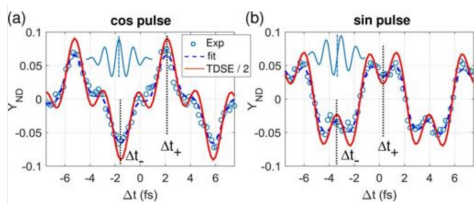
Question: What parameter is being measured in both experiments depicted in the figures?
Options:
A. The intensity of the laser pulse
B. The time delay in femtoseconds (fs)
C. The yield normalized differential (Y_ND)
D. The amplitude of the experimental setup
Label: C
Rationale: The Y-axis in both figures is labeled as Y_ND, indicating the yield normalized differential as the measured parameter.

Figure 10: A case from our ArXivQA dataset.
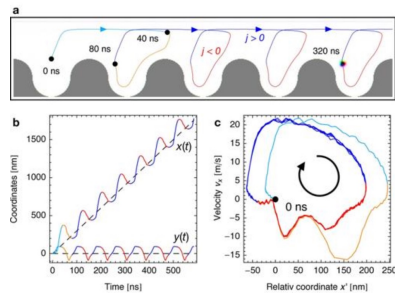
Question: What can be inferred about the particle's motion in figure c?
Options:
A. The particle moves in a straight line with constant velocity.
B. The particle exhibits circular motion with changing velocity.
C. The particle's velocity components are independent of each other.
D. The particle experiences uniform acceleration in a circular path.
Label: B
Rationale: In figure c, the trajectory forms a closed loop with a direction of motion indicated, suggesting circular motion. The varying distance between successive loops implies that the velocity is changing.

Figure 11: A case from our ArXivQA dataset.
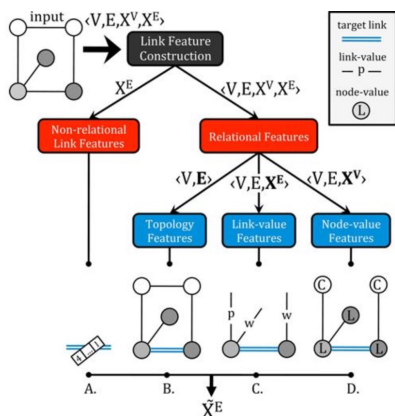
Question: Which of the following best describes the 'target link' in the context of this figure?
Options:
A. A link that connects nodes with the highest values
B. The link under examination for feature construction
C. A predefined link that guides the construction of other links
D. The most frequently occurring link in the network

Label: B
Rationale: The 'target link' is shown with a 'link-value' p, separated from the 'node-value' L, suggesting it is the specific link being analyzed for its features.

Figure 12: A case from our ArXivQA dataset.
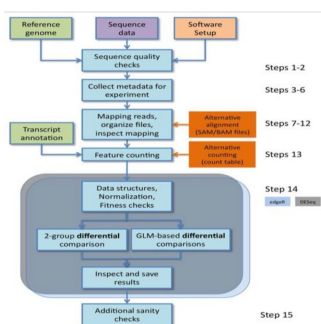
Question: Which steps in the workflow are directly involved in preparing the data for differential expression analysis?
Options:
A. Steps 1-2
B. Steps 7-12
C. Steps 13-14
D. Step 15
Label: C
Rationale:
The steps directly involved in preparing the data for differential expression analysis include "Feature counting" to get a count table and the subsequent data structuring and normalization processes listed under "Step 14". These are the steps that directly precede the actual differential expression analysis methods like "2-group differential comparison" and "GLM-based differential comparisons".

Figure 13: A case from our ArXivQA dataset.

ing dataset. LLaVA-v1.5 (Liu et al., 2023a) improves on LLaVA models by employing curated task datasets and an enhanced modality alignment module. We evaluate both `LLaVA-v1.5-7B`[5] and `LLaVA-v1.5-13B`.[6]

Flamingo (Alayrac et al., 2022) pioneers the development of LVLMs by introducing a cross-

---

[5] https://huggingface.co/liuhaotian/llava-v1.5-7b

[6] https://huggingface.co/liuhaotian/llava-v1.5-13b

gated layer for LLMs to produce visual-grounded text. The training dataset consists of interleaved visual data and text from the web pages, enabling it to generate free-form text as the output. We select the open-source implementation `OpenFlamingo-9B` (Awadalla et al., 2023) for evaluation.[7]

IDEFICS is another open-sourced implementation of Flamingo (Alayrac et al., 2022). Trained on publicly available image-text alignment pairs and instruction tuning datasets, it demonstrates comparable results with the original closed-source model on various image-text benchmarks. We select the `IDEFICS-Instruct-9B` for evaluation.[8]

Qwen-VL-Chat (Bai et al., 2023b) is a bilingual LVLM that supports both English and Chinese built on the Qwen LLM (Bai et al., 2023a). During the training phase, Qwen-VL-Chat adopts a packing strategy to create multiple images as inputs, improving its ability to understand the vision context. We select the `Qwen-VL-Chat-7B` for evaluation.[9]

GPT-4V (OpenAI, 2023), the proprietary vision-language models developed by OpenAI, which are shown to be powerful on various multi-modal tasks (Yang et al., 2023a). The API version we queried is `gpt-4-vision-preview`.

Bard (Google, 2023), a commercial LVLM developed by Google. We utilize the unofficial API[10] querying the model with our task prompts, accessed on `2023-11-17`.

Gemini 1.0 Pro Vision (Reid et al., 2024), a upgraded LVLM by Google. We utilize the official API querying the model with our task prompts, accessed on `2024-05-20`.

B.2 Task Prompts

We evaluate all the models with the same task prompts in our experiments, and the prompts for our four tasks are listed below:

Single-Figure Captioning: `Create a caption for the provided figure.`

Multiple-Figure Captioning `Create a caption for the provided figures.`

Contextualized Captioning: We reuse the prompts in previous captioning tasks depending on the current figure type.

Title Generation: `According to the figures and captions, generate a title for this paper. Title:`

B.3 GPT-4 Evaluation of Caption

In addition to BLEU-2, ROUGE-L, and BERT-S, we also utilize GPT-4 to evaluate a sample of 500 generated captions. Specifically, we employ GPT-4 for the evaluation of single-figure caption tasks following FairEval (Wang et al., 2023). The template for prompting GPT-4 to evaluate generated captions is presented in Table 14. GPT-4 is asked to perform an analysis and then produces a quality score, which is subsequently mapped to a scale from 1 to 5. The results are presented in Table 15. We observe that the ROUGE-L metric exhibits the highest correlation with the GPT-4 Score (Pearson r = 0.91), followed by BLEU-2 (Pearson r = 0.64). BERT-S instead demonstrates a moderate correlation (Pearson r = 0.39). The uniformly low GPT-4 scores across all models suggest that they struggle to produce satisfactory captions, which is consistent with the findings in our main paper. Notably, training on ArXivCap results in a significant 12% improvement in the GPT-4 score compared to the original Qwen-VL-Chat model, leading to the most favorable outcomes in this evaluation.

C Error Analysis

C.1 Caption Type Illustration

We illustrate captions of four types in our main paper in Figure 14. The *Acceptable* caption provides a comprehensive description of the figure presented. The *oversimplified* caption is too short compared with the original human-written caption. Furthermore, as shown in the third block in Figure 14, *Contextual Misinterpretation* refers to captions with unmentioned content in the figure, such as the dataset colored in red. *Recognition Error* denotes the model wrongly identified the number or text in the figure, such as the misidentified model name in the last block of Figure 14.

C.2 Failure Sample of MathVista

Figure 15 shows a challenging geometry mathematic reasoning problem where both models fail to produce the correct answer. Echoing the quantitative results in our main paper, we believe future

Table 14: Prompt template designed for GPT-4 to evaluate generated captions based on the paper title, abstract, and ground truth.

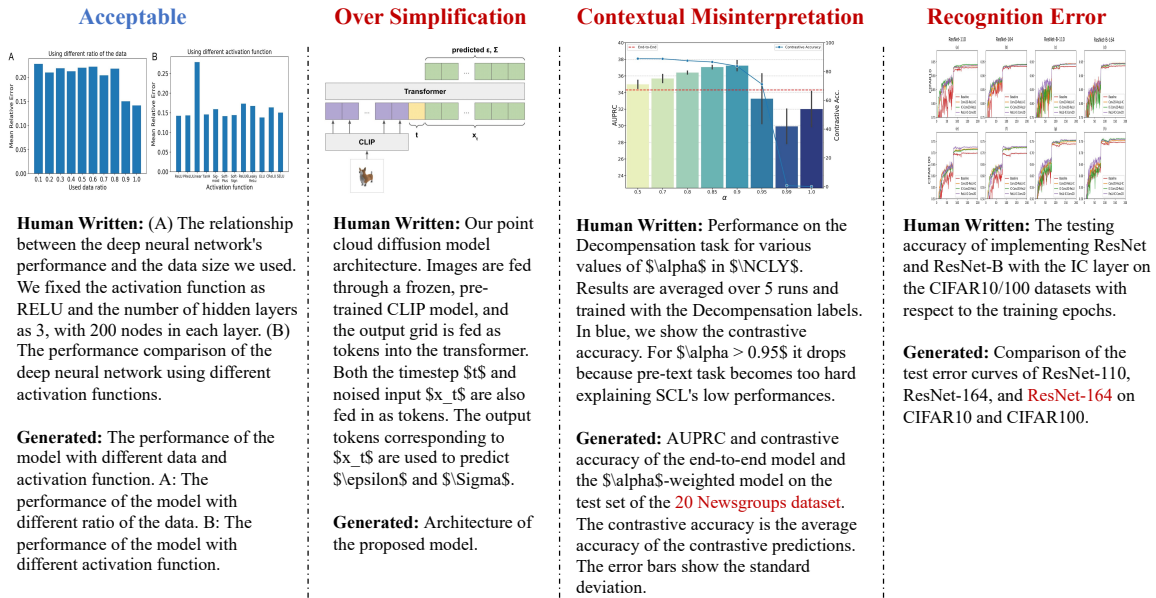Figure 14: Illustration of acceptable and three error types of generated captions.

Model	BLEU-2	ROUGE-L	BERT-S	GPT-4 Score
BLIP-2-OPT-6.7B	1.5	6.6	81.3	1.18
InstructBLIP-Vicuna7B	3.5	10.3	83.6	1.48
LLaVA-1.5-7B	2.3	10.4	83.3	1.80
LLaVA-1.5-13B	2.7	11.0	83.6	1.69
OpenFlamingo-9B	5.8	10.3	82.7	1.52
IDEFICS-Instruct-9B	2.1	9.3	83.8	1.55
Qwen-VL-Chat	4.7	11.1	82.0	1.81
Qwen-VL-Chat tuned w/ ArXivCap	8.6	15.3	83.2	2.03

Table 15: Results of 500 single-figure captions generated by various models.

studies can incorporate more focused corpus for enhancing the geometry and mathematical reasoning ability of LVLMs.

D Results with LLaVA Backbone

We investigate whether ArXivQA could also enhance other LVLMs, such as LLaVA models (Liu et al., 2023b). To maintain model performance, we mix our ArXivQA dataset with the LLaVA SFT 665K-instruction tuning dataset. The LLaVA-v1.5-7B is adopted as the backbone and the model is trained following the original recipe. The results on various benchmarks are listed in Table 16. We find that not only the scientific reasoning performance is improved on multimodal reasoning tasks (MathVista (Lu et al., 2023), MMMU (Yue et al.,

Model	MathVista	MMMU(val)	ScienceQA(IMG only)	MM-Vet
LLaVA-v1.5-7B	26.6	35.3	66.8	30.5
Original SFT +ArXivQA	**28.2**	**36.0**	**68.3**	**32.4**

Table 16: After fine-tuning with a combination of ArXivQA and original SFT data, the LLaVA model shows boosted mathematical reasoning abilities across benchmarks.
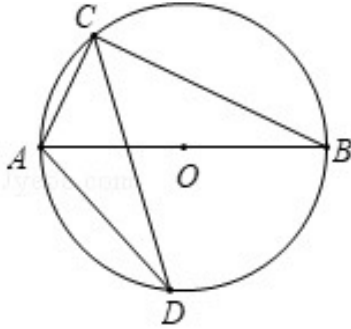
Question: As shown in the figure, AB is the diameter of ⊙O, CD is the chord of ⊙O, ∠ADC = 26.0, then the degree of ∠CAB is (). Choices: A. 26° B. 74° C. 64° D. 54°

Ground-Truth: D. 54°

Qwen-VL-Chat: We know that a chord of a circle intersects the circle at two points, and the measure of the angle formed by the chord and the radius is equal to the fraction of the circumference of the circle that the chord represents. In this figure, ∠ ADC=26.0°, and ∠ACB= ∠CAD, so ∠ACB + ∠AD=90°. Since ∠ACB = ∠AD, we have ∠ACB + ∠ACB=90°, so ∠ACB=45°. Then ∠CAB = 90°- ∠ACB = 90°- 45°= 45°. Therefore, the degree of ∠CAB is 45°. Thus, the answer is (A).

Qwen-VL-Chat-Tuned: The answer is B. ΔABC is a 74°-64° triangle, and ΔBCD is a 54°-26° triangle. Adding these two angles gives us 74° + 54° = 128°, which is not possible since the sum of the angles in a triangle must equal 180°. Therefore, ∠CAB must be 74°, which is the only angle that fits the given conditions.

Figure 15: A failure case on the geometry problem-solving task.

2023), and ScienceQA (Lu et al., 2022)), but the overall capability on MM-Vet (Yu et al., 2024) is also boosted. Together with our results using Qwen-VL-Chat, these findings indicate that our ArXivQA dataset can enhance different model backbones and is beneficial across various benchmarks.