# Multi-LLM Collaborative Caption Generation in Scientific Documents

Jaeyoung Kim[1 [0000−0003−0880−0398]], Jongho Lee[1 [0000−0002−8520−2722]], Hong-Jun Choi[1 [0000−0002−3413−511X]], Ting-Yao Hsu[2 [0009−0008−9082−6039]], Chieh-Yang Huang[2 [0009−0001−6736−9959]], Sungchul Kim[3 [0000−0003−3580−5290]], Ryan Rossi[3], Tong Yu[3 [0000−0002−5991−2050]], Clyde Lee Giles[2 [0000−0002−1931−585X]], Ting-Hao 'Kenneth' Huang[2 [0000−0001−7021−4627]], and Sungchul Choi[1 [0000−0002−5836−3838]]

[1] Teamreboott Inc., Busan, Korea
{jaeyoungkim,jongho.lee,hongjun.choi,admin}@reboott.ai
[2] Pennsylvania State University, University Park, PA, USA.
{txh357,chiehyang,clg20,txh710}@psu.edu
[3] Adobe Research, San Francisco, CA, USA.
{sukim,ryrossi,tyu}@adobe.com

**Abstract.** Scientific figure captioning is a complex task that requires generating contextually appropriate descriptions of visual content. However, existing methods often fall short by utilizing incomplete information, treating the task solely as either an image-to-text or text summarization problem. This limitation hinders the generation of high-quality captions that fully capture the necessary details. Moreover, existing data sourced from arXiv papers contain low-quality captions, posing significant challenges for training large language models (LLMs). In this paper, we introduce a framework called **M**ulti-**L**LM Colla**b**orative Figure **Cap**tion Generation (**MLBCAP)** to address these challenges by leveraging specialized LLMs for distinct sub-tasks. Our approach unfolds in three key modules: (**Quality Assessment**) We utilize multimodal LLMs to assess the quality of training data, enabling the filtration of low-quality captions. (**Diverse Caption Generation**) We then employ a strategy of fine-tuning/prompting multiple LLMs on the captioning task to generate candidate captions. (**Judgment**) Lastly, we prompt a prominent LLM to select the highest quality caption from the candidates, followed by refining any remaining inaccuracies. Human evaluations demonstrate that informative captions produced by our approach rank better than human-written captions, highlighting its effectiveness. Our code is available at https://github.com/teamreboott/MLBCAP

**Keywords:** Image captioning · Collaborative framework · Large language models.

## 1 Introduction

Scientific figures are integral to academic communication, offering a concise and effective means of presenting complex information. However, the value of a fig-

ure is largely determined by the quality of its accompanying caption. Captions provide essential context, elucidate visual elements, and enable readers to fully grasp the insights conveyed by the figure. Consequently, the generation of accurate and informative captions for scientific documents is critical to effectively communicating key findings to domain experts. Automated captioning not only aids researchers by improving the clarity of figure descriptions but also contributes to the overall enhancement of scholarly communication [9].

Existing approaches to automatic figure captioning have predominantly treated the task either as an image-to-text problem [8,21] or a text summarization task [11,4]. Image-to-text methods focus on extracting information directly from visual content, but they often lack the domain-specific understanding required to interpret abbreviations, symbols, and implicit relationships. On the other hand, text summarization approaches rely on textual metadata such as figure-mentioning paragraphs or optical character recognition (OCR) outputs from figures. While these methods can capture textual context, they frequently overlook crucial visual details, such as trends, patterns, and color-coded elements that are vital for a comprehensive understanding of the figure. Consequently, these fragmented approaches fail to produce captions that are both accurate and informative, underscoring the need for a unified framework capable of leveraging both textual and visual modalities.

Another major challenge in figure captioning lies in the quality of available training data. Many existing datasets [14,13,8,26], particularly those sourced from platforms like arXiv, contain captions that are incomplete, verbose, or poorly written. A recent study [11] reports that over 50% of captions in arXiv papers are unhelpful to domain experts. These low-quality captions can hinder model training and result in sub-optimal caption generation, complicating the accurate assessment of model performance.

To this end, we propose a unified framework named **M**ulti-**L**LM Colla**b**orative Figure **Cap**tion Generation (**MLBCAP**). Unlike previous methods, MLBCAP integrates textual and visual information through a carefully orchestrated pipeline comprising three key components: quality assessment, diverse caption generation, and judgment. The quality assessment module filters out low-quality training captions, ensuring that the models are trained on reliable data. In the caption generation stage, multiple LLMs, each specializing in different aspects of figure captioning, collaborate to produce diverse candidate captions. Finally, a judgment module utilizes a prominent LLM to select the best candidate caption and refine it for accuracy and coherence.

While prior studies find that longer captions are generally more beneficial to readers [7,11], scientific journals and conference papers often impose strict page limits. To accommodate this, our framework is designed to generate both long and short versions of captions. In the final step, we utilize GPT-4o with specific instructions regarding caption length to achieve that both versions are concise yet informative. In a human evaluation by domain experts, captions generated by our method are preferred over the original author-written captions, demonstrating the effectiveness of our approach. Our main contributions are as follows:

- We propose a unified framework that includes data cleaning, caption generation, and post-editing processes to generate high-quality captions.
- Our approach integrates both textual and visual features, leveraging multimodal models to produce contextually rich and accurate captions.
- Through human evaluations, we show that our approach ranked better than author-written captions, demonstrating its effectiveness.

## 2 Related Work

### 2.1 Collaboration Techniques with LLMs

LLMs have shown exceptional performance across a wide range of tasks, benefiting from their ability to comprehend instructions [1,28]. However, despite their versatility, individual LLMs exhibit distinct strengths and limitations due to differences in training data and architectural design [12]. To mitigate this issue, recent work [23] trained a classifier to select the best response generated by different reasoning models. Another related work [6] proposed an algorithm that combines outputs from multiple LLMs for attribute extraction through weight assignment. Despite the growing application of collaborative methods in various fields, a significant research gap exists in exploring their potential for figure captioning in scientific documents.

### 2.2 Figure Captioning in Scientific Documents

To facilitate the generation of captions by neural networks, previous research developed a variety of datasets, such as FigureSeer [24], FigureQA [14], DVQA [13], and SciCap [8]. More recently, an enhanced version of SciCap was introduced, incorporating both figures and their associated textual information [11]. This dataset advances caption generator capabilities, enabling them to produce contextually relevant captions for scientific figures. Based on this dataset, a recent study [11] discovered that more than 76% of the words in figure captions matched those in figure-mentioning paragraphs and OCR text. Based on this empirical observation, they formulated the figure captioning task as a text summarization task. Contemporaneously, SciCap+ [26] was proposed as an extension of SciCap, integrating OCR-derived textual data to further enhance the generation of figure captions. However, text summarization models, which depend on textual data from figure-related paragraphs and OCR outputs, often fail to capture essential visual details, including patterns and colors in graphs.

### 2.3 Evaluating Natural Language Generation (NLG) Tasks

In NLG tasks, traditional automatic metrics such as BLEU [20] and ROUGE [16] are widely used for evaluation. However, these metrics often exhibit a relatively low correlation with human judgments in text generation tasks [18], mainly because they depend on human-preferred reference outputs to fairly evaluate the

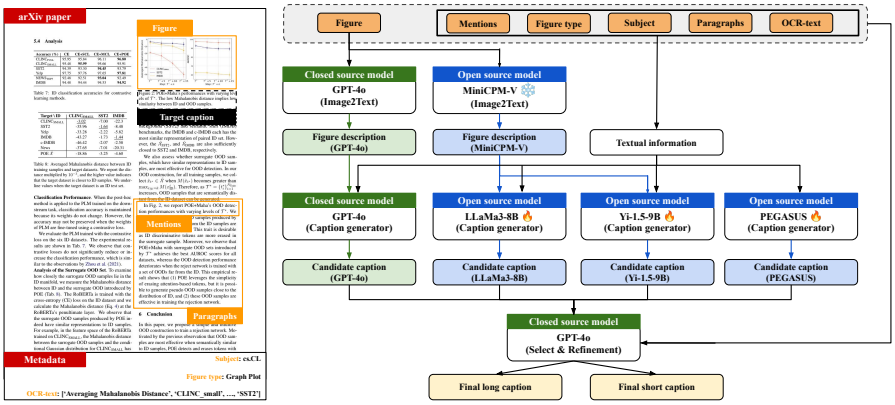



**Fig. 1.** Overview of the collaborative framework integrating multiple LLMs for caption generation in scientific documents. Initially, two MLLMs generate figure descriptions. Next, three fine-tuned models and GPT-4o generate candidate captions. Finally, GPT-4o selects and refines the best caption from the candidates.

performance of NLG models. Recent studies have advocated for using LLMs as reference-free evaluation metrics, achieving higher correspondence with human evaluations than traditional metrics [30,18]. Notably, SciCap-Eval [10] employed LLMs to assess caption quality and demonstrated that GPT-4, as a zero-shot caption evaluator, positively correlates with Ph.D. students' assessments (Pearson correlation coefficient of 0.5).

## 3 Problem Statement

Consider an arXiv paper $D$ with $n$ captions $\{C_i\}_{i=1}^{n}$. For each caption $C_i$, several related sources from $D$ are used to assist in caption generation. These sources include the corresponding figure $F_i$ and $m$ paragraphs $\{P_i^j\}_{j=1}^{m}$ that mention $F_i$. Specifically, $k$ sentences $\{M_i^j\}_{j=1}^{k}$ within these paragraphs explicitly refer to $F_i$ (e.g., "As shown in Fig. 1, ..."). [4] Additionally, textual information extracted using OCR from the figure is denoted as $O_i$, and the figure's type (e.g., bar chart, node diagram) is represented as $T_i$. Finally, the subject category of $D$ (e.g., "cs.AI" for Computer Science - Artificial Intelligence) is denoted as $S$. The objective of this work is to generate high-quality $C_i$ for $F_i$ utilizing the aforementioned figure-relevant features.

[4] For simplicity, superscripts for $P_i$ and $M_i$ are omitted in the following sections, as multiple instances may exist.

## 4 Multi-LLM Collaborative Figure Caption Generation

Our overall pipeline is illustrated in Figure 1, and the following sections provide a description of each component. The actual prompts are described in the Appendix.

### 4.1 Quality Assessment

We employ GPT-4o to generate a synthetic quality assessment dataset using 3k subset of the training data. Following the approach of SciCap-Eval, we prompt GPT-4o to score captions on a scale of 1 to 6 based on the given $C_i$, $F_i$, and $P_i$ (higher scores indicate better quality).

Next, we fine-tune LLaVA [17] on the constructed dataset. After fine-tuning, LLaVA predicts the caption quality across the entire training dataset. We then collect samples $\mathcal{D}_{\text{high}}$ with quality scores of 5 and 6. For the evaluation of 200 samples, the fine-tuned LLaVA showed agreement in quality assessment with GPT-4o, achieving Kendall's tau coefficient of 0.5502.

### 4.2 Diverse Caption Generation

To capture diverse perspectives and generate a varied set of candidate captions, we utilize four distinct models: GPT-4o, LLaMA-3-8B [2], Yi-1.5-9B [28], and Pegasus [29]. Each model offers unique viewpoints that contribute to the diversity of the generated captions. Specifically, GPT-4o leverages its advanced reasoning capabilities as a large-scale LLM. LLaMA-3-8B and Yi-1.5-9B are fine-tuned on the figure captioning task, enhancing domain-specific knowledge. Pegasus excels in abstractive summarization of textual content, capturing essential information from figure-mentioning paragraphs and OCR text.

***GPT-4o***. For a test sample, we use few-shot prompting by providing GPT-4o with ten example captions, $E$, randomly selected from $\mathcal{D}_{high}$. These examples have the same subject as the test sample and have a quality score of 6. We first instruct GPT-4o to generate a figure description $Z_i$ for $F_i$ by providing the $F_i$, $T_i$ and $S$. Then, GPT-4o generates a candidate caption based on $E$, $P_i$, $M_i$, $T_i$, $O_i$, $S$, and $Z_i$.

***LLaMA-3-8B and Yi-1.5-9B*** are fine-tuned with visual and textual features from the $\mathcal{D}_{\text{high}}$ dataset. Figure descriptions are generated by MiniCPM-V [27], which outperforms GPT-4V-1106 and Gemini-Pro [22] for OpenCompass benchmarks [5]. The prompts used for LLaMA-3-8B and Yi-1.5-9B are identical to the prompts used for GPT-4o, except for the exclusion of the few-shot examples.

***Pegasus*** is fine-tuned on figure-mentioning paragraphs and OCR-text from $\mathcal{D}_{\text{high}}$. Apart from the dataset, this model follows the previous work [11].

### 4.3 Judgement

We ask GPT-4o to select the best quality caption from four candidate captions and edit inaccuracies in the selected caption leveraging both visual and textual

**Table 1.** Statistics of the original and preprocessed datasets used for training, validation, and testing.

| | Train | Validation | Test |
|---|---|---|---|
| SciCap | 360,340 | 47,639 | 47,639 |
| SciCap+ | 394,005 | - | - |
| Preprocessed | 135,935 | 47,639 | 47,639 |

**Table 2.** The result of caption quality evaluation using GPT-4o. The highest quality captions have a low percentage of 27.11%.

| Score | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| N (%) | 157 (5.24) | 305 (10.13) | 102 (3.38) | 166 (5.53) | 1,457 (48.58) | 813 (27.11) |

information. To generate both long and short captions, we set a word limit using the placeholder `[MAX_LEN]` in the prompt. To maintain the conciseness while sufficiently conveying figure information, we define `[MAX_LEN]` as 50 words for long captions and 30 words for short captions. These limits are based on the training data, where the average caption length is 41.85 words. Here, we refer to the generation of long captions as MLBCAP (long) and short captions as MLBCAP (short).

## 5 Experiments

### 5.1 Dataset

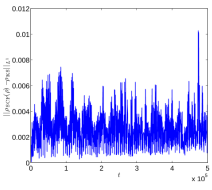
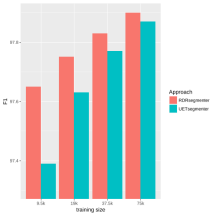
One of the goals of this study is to evaluate whether MLBCAP can generate captions that align with human preferences. To enhance the richness of the training data, we have chosen to combine the SciCap[5] and SciCap+ [26] datasets for training. These datasets include caption-relevant metadata extracted from the original arXiv papers. Detailed statistics of the datasets is provided in Table 1.

**Data Preprocessing.** Since both SciCap and SciCap+ are built on arXiv papers, we first conducted deduplication using the arXiv paper ID. To obtain high-quality examples, the following preprocessing steps were applied. First, figures with captions that do not end with a period were excluded. Second, captions that were either too long or too short, specifically those exceeding 100 words in length or consisting of a single sentence, were removed. Lastly, figures with captions rated below 5, as described in Section 4, were filtered out. After these preprocessing steps, we could get 135,935 samples.

**Caption Quality Dataset.** We employ GPT-4o to generate a synthetic quality assessment dataset using 3k subset samples for the training data since there is no

[5] https://huggingface.co/datasets/CrowdAILab/scicap/tree/main

**Table 3.** The example of figure descriptions between GPT-4o and MiniCPM-V.

| Figure | GPT-4o | MiniCPM-V |
|---|---|---|
|  | The graph shows the L1 norm of the difference between $pSCF(t)$ and $pKs$ over time (t). The x-axis represents time (t) ranging from 0 to 500,000, and the y-axis represents the L1 norm of the difference, ranging from 0 to 0.012. The plot indicates fluctuations in the difference over time, with occasional peaks. | The image contains a graph with a blue line representing the plot of data points. The x-axis is labeled as x $10^5$ and the y-axis is labeled as $\|\|pSCF(p) - pKs\|\|$. There are no additional textual elements or objects present in the image. |
|  | The bar chart compares the F1 scores of two segmentation approaches, RDRsegmenter and UETsegmenter, across different training sizes (9.5k, 19k, 37.5k, and 75k). RDRsegmenter consistently achieves higher F1 scores than UETsegmenter at all training sizes. Both approaches show an increase in F1 score as the training size increases, with RDRsegmenter reaching the highest F1 score at 75k training size. | The image contains a bar chart with two types of bars representing different approaches: RDRsegmenter and UETsegmenter. The x-axis represents the training size (k), and the y-axis represents the F1 score, which ranges from 97.4 to 98. |

publicly available dataset. The quality distribution of the 3k captions is shown in Table 2, which shows that the training data contains low-quality captions.

### 5.2 Training Details

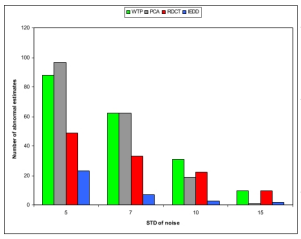
**Prompts**. The specific prompts employed in our study are detailed in the Appendix. To elicit figure descriptions from the MiniCPM-V model, we employed a direct and intuitive prompt: "*What is in the image?*". We also tried extracting figure descriptions with the prompts used in GPT-4o, however, the quality of the generated descriptions was inferior compared to descriptions obtained using the more intuitive prompts. A representative example of the figure descriptions generated is illustrated in Table 3.

**Models**. Here are the models we used in our experiments. The fine-tuned models were tuned for 5 epochs.

- **GPT-4o.** We used `gpt-4o-2024-05-13` with a temperature setting of 0. All experimental procedures were conducted between May 2024 and August 2024.
- **LLaVA.** We used `llava-llama-3-8b-v1_1-hf` model [6], which is fine-tuned from the LLaMA-3-8B architecture. This variant of LLaVA was optimized

[6] https://huggingface.co/xtuner/llava-llama-3-8b-v1_1-transformers

**Table 4.** Comparison of generated captions. MLBCAP includes key information and provides comprehensive descriptions, whereas Pegasus tends to produce shorter captions that lack sufficient detail.

| Figure | Model | Caption |
|---|---|---|
|  | **MLBCAP (long)** | Fig. 4. IEDD significantly outperforms other methods in reducing abnormal estimates. The bar chart shows that IEDD consistently has the lowest number of abnormal estimates across various noise levels (STD of noise: 5, 7, 10, 15), compared to WTP, PCA, and RDCT. |
| | **MLBCAP (short)** | Fig. 4. IEDD significantly outperforms other methods in reducing abnormal estimates across various noise levels (STD of noise: 5, 7, 10, 15). |
| | LLaMA-3-8B | Fig. 4. Abnormal estimates (AE) for different methods. The proposed method IEDD significantly outperforms all compared methods. The results are obtained for the synthetic dataset. |
| | Yi-1.5-9B | Fig. 4. Number of abnormal estimates (AE) for different noise variances. For each variance, the experiment was repeated 100 times. The proposed method IEDD significantly outperforms other compared methods. |
| | Pegasus | Fig. 4. Comparison of RRMSE of the proposed method (IEDD) with other methods. |

using the AdamW [19] optimizer, with a learning rate set to $1e-5$ and a batch size of 4.

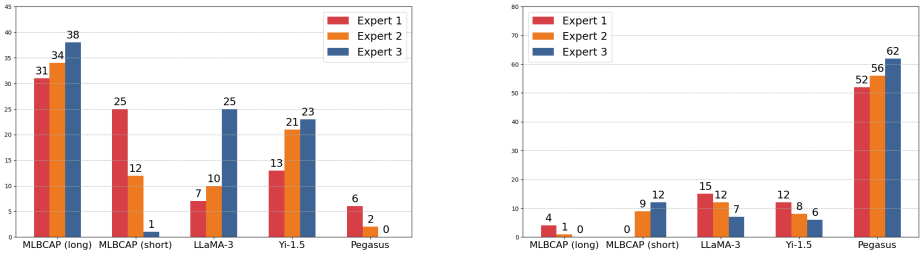
- **MiniCPM-V.** For the task of figure description extraction, we used the latest model (`MiniCPM-Llama3-V 2.5` [7]) in the MiniCPM-V families without the fine-tuning.
- **LLaMA-3-8B.** To generate a caption, we used `Meta-Llama-3-8B-Instruct` [8] model. The model was trained with AdamW, a learning rate of $1e-5$ and a batch size of 1.
- **Yi-1.5-9B** is an upgraded version of Yi. For generating captions, we used `Yi-1.5-9B-Chat` [9]. The training configuration is the same as LLaMa-3-8B.
- **Pegasus.** We trained the `pegasus-large` [10] model using the AdamW optimizer, with a batch size of 32 and a learning rate of $5e-5$.

When fine-tuning the caption generation models, we concatenated all figure-mentioning paragraphs. In cases where the cumulative length of these paragraphs exceeded 512 tokens, the text was truncated to fit within this limit. For text generation, we utilized greedy search to ensure the generation of coherent.

### 5.3 Human Evaluation Results

We evaluate MLBCAP with a human evaluation to accurately assess the quality of generated captions. The baseline models used for evaluation include LLaMA-3-8B, Yi-1.5-9B, and Pegasus. All baselines were fine-tuned on the high-quality

[7] https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5
[8] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[9] https://huggingface.co/01-ai/Yi-1.5-9B-Chat
[10] https://huggingface.co/google/pegasus-large




**(a)** Frequency of high-quality captions.

**(b)** Frequency of low-quality captions.

**Fig. 2.** The human evaluation results for selecting captions based on quality. Captions generated by MLBCAP (long) are most frequently selected as high quality.

captions from the SciCap and SciCap+ datasets, specifically captions with a quality score of 5 or higher. The qualitative results for generated captions are illustrated in Table 4.

### 5.4 Comparison with Baselines

We recruited three computer vision (CV) experts, each with over five years of experience in the field. These experts were asked to select the best and worst quality captions generated by LLMs. We used 91 CV figures from the SciCap test dataset, with the candidate captions de-identified and randomly shuffled for each figure before being presented to the experts.

Figure 2(a) presents the results of the human evaluation. Notably, captions generated by MLBCAP (long) were consistently selected as high-quality by all three experts, indicating a strong preference compared to baselines. To understand the high preference for the MLBCAP, we investigated which LLM-generated captions were selected during the best caption selection phase. The percentages of captions selected from the GPT-4o, Yi-1.5-9B, LLaMA-3-8B, and Pegasus models were 89.38%, 4.23%, 6.17%, and 0.19%, respectively. As expected, GPT-4o performed exceptionally well in generating high-quality captions.

Interestingly, the assessment of MLBCAP (short) captions revealed variability in expert opinion. Despite this variance, Figure 2(b) demonstrates that MLBCAP (short) captions were rarely selected as the worst quality, suggesting that the differences in preference may stem more from individual biases towards caption length rather than from a significant discrepancy in caption quality.

### 5.5 Comparison with Author Captions

Another human evaluation was conducted as part of the 2nd SciCap challenge [11]. To account for the natural distinctions and ensure equitable assessment between

[11] http://scicap.ai/

**Table 5.** The human evaluation results for the SciCap challenge and MLBCAP is compared with the second-best solutions for each track. Lower average ranks indicate higher preference by the judges.

| Team | Long Caption | Short Caption |
|---|---|---|
| Author-written | 2.84 | **1.52** |
| LM-Ensemble [15] | 2.82 | 3.56 |
| Length-Adaptive LLM [25] | 3.08 | 3.18 |
| MLBCAP | **1.27** | 1.74 |

short and long captions, the task was divided into two tracks. Participants in the short caption track were required to submit results where at least 30% of the captions were no longer than the original author-written captions from the SciCap test dataset. Similarly, teams in the long caption track were required to submit results where at least 30% of the captions exceeded the length of the original author-written captions. In the case of MLBCAP, 68.15% of the captions in the long caption track were longer than the author-written captions, while 46.53% of the captions in the short caption track were shorter than the author's captions.

Three judges (who are not the CV experts), all native American English speakers with expertise in technical academic writing, were recruited to rank the captions. They ranked captions for the same 200 figures, randomly selected from the challenge test set, based on how effectively they convey the figure's message for each track. The selection of these figures adhered to the following criteria: (1) For the long caption track, figures were selected where the author-written captions were shorter than the generated captions submitted by participants. (2) For the short caption track, figures were chosen where the author-written captions exceeded the length of the generated captions submitted by participants. (3) For both the long and short caption tracks, only figures with a SciCap-Eval score of 4 or higher were included.

As shown in Table 5, in the long caption track, MLBCAP (long) outperformed all others, receiving the best score of 1.27. This result presents the effectiveness of our approach in generating comprehensive captions that resonated well with expert evaluators. Additionally, although the MLBCAP's caption in the short caption track ranked marginally lower than the author-written captions, it still placed our method ahead of other methods.

### 5.6 Analysis

Here we conduct an analysis focused on the components of MLBCAP to examine their impact on caption quality, using a random sample of 200 instances from the SciCap test dataset. Table 6 illustrates the significant improvements in caption quality across all models when figure descriptions and a robust filtering process are incorporated. While models such as LLaMA-3-8B and Yi-1.5-9B perform

**Table 6.** The impact of including figure descriptions and the filtering process. Each value is the caption quality score (SciCap-Eval).

| | LLaMA-3-8B | YI-1.5-9B | GPT-4o |
|---|---|---|---|
| Base model | 4.612 | 4.575 | 5.305 |
| + Quality Assessment | 4.905 | 4.910 | - |
| + Figure description | **5.030** | **5.005** | **5.390** |

**Table 7.** Ablation study analyzing the effect of the best caption selection and post-editing on caption quality.

| Multi-LLM | Post-edit | Quality Score (SciCap-Eval) |
|---|---|---|
| ✘ (GPT-4o) | ✘ | 5.390 |
| ✘ (GPT-4o) | ✔ | 5.405 |
| ✔ | ✘ | 5.430 |
| ✔ | ✔ | **5.440** |

well in generating high-quality captions, GPT-4o emerges as the best model for providing top-tier candidate captions.

However, Table 7 reveals a crucial insight; the highest caption quality score (5.440) is achieved not solely by relying on the individual model (GPT-4o) but through a strategic combination of multiple LLMs and subsequent post-editing. This approach demonstrates that even with the availability of a powerful model like GPT-4o, the incorporation of diverse perspectives from smaller LLMs can yield improved results.

## 6 Discussion

### 6.1 Caption Preferences

While MLBCAP rarely produced low-quality captions in human evaluations, we observed variability among experts when selecting the best and worst captions. This challenge was reflected in the low inter-rater agreement metrics. For instance, Fleiss' kappa for selecting high-quality captions among baseline models was 0.154, indicating low agreement, while for low-quality captions, the kappa improved to 0.382, signifying moderate agreement. Similarly, in the SciCap Challenge, Kendall's tau for inter-rater agreement, provided by the challenge organizers, was 0.3589 for long captions and 0.1100 for short captions, highlighting the difficulty of reaching a consensus even among experienced judges.

These findings align with previous research [11], which has shown that caption ranking tasks inherently elicit varied judgments from evaluators. The low agreement underscores the complexity of defining "quality" in figure captioning,

**Table 8.** Evaluation results for the SciCap test dataset. The ROUGE (F1-score) and BLEU (4-gram) scores of MLBCAP are opposite to the human preference (Figure 2).

| | ROUGE-1 | ROUGE-2 | ROUGE-L | BLEU-4 |
|---|---|---|---|---|
| Pegasus | **0.460** | **0.282** | **0.418** | 0.124 |
| LLaMA-3-8B | 0.405 | 0.237 | 0.346 | 0.126 |
| Yi-1.5-9B | 0.412 | 0.244 | 0.354 | **0.134** |
| MLBCAP (short) | 0.369 | 0.174 | 0.310 | 0.043 |
| MLBCAP (long) | 0.333 | 0.150 | 0.257 | 0.049 |

where multiple factors interplay, such as the caption's informativeness, length, detail, and overall style.

A plausible explanation for the observed discrepancies lies in the subjective nature of caption preferences. Evaluators may prioritize different attributes, such as the level of detail versus brevity, or favor stylistic differences in language. For instance, one expert might value a caption that provides exhaustive detail, while another might prefer concise summaries that align with the space constraints typical in scientific publications. This subjectivity in preferences naturally leads to variability in judgments, reducing inter-rater reliability.

These observations highlight the importance of developing clearer guidelines and evaluation criteria for figure captions. A more standardized framework could help align evaluators' judgments and establish a consensus on what constitutes a "high-quality" caption. Future work could explore leveraging LLMs not only for caption generation but also as assistive tools for evaluating captions in a more consistent manner, thereby addressing some of the subjectivity inherent in human evaluations. This would ensure that assessments of caption quality are both rigorous and aligned with the needs of diverse scientific communities.

### 6.2 Evaluation with Traditional Metrics

Furthermore, we found a discrepancy between traditional metric-based evaluations and human judgments. In Table 8, while MLBCAP was preferred over the author-written captions in the human evaluations, traditional metrics failed to capture the perceived quality of captions. This divergence presents the limitations of relying solely on conventional metrics for evaluating caption quality. Similar observations have been made in the natural image captioning task [3], where BLEU and ROUGE scores showed low correlation with human judgments (Kendall's tau of approximately 0.3 for both metrics).

This inconsistency can be attributed to the inherent limitations of BLEU and ROUGE metrics, which focus on n-gram overlap between the generated caption and the reference caption. In scientific figure captioning, there are multiple valid ways to describe the same content using different terminologies or phrasings. As a result, even high-quality captions may receive low scores if they do not closely match the reference.

## 7 Limitations

While our experiments indicate the potential of MLBCAP in the figure captioning task, there are some limitations that point to possible directions for future work.

Firstly, the integration of multiple LLMs in our framework introduces a trade-off between performance and efficiency. The reliance on multiple models not only reduces inference speed but also increases the demand for computational resources. This may limit the practical scalability of MLBCAP, particularly in environments with restricted computational capabilities or in real-time applications.

Secondly, MLBCAP incorporates a closed-source LLM as a critical component of the caption generation pipeline. This inclusion imposes inherent limitations, particularly in terms of transparency and interpretability. The closed-source nature restricts our ability to fully understand and analyze the model's reasoning processes and decision-making behavior, which may hinder trust and adoption in certain scientific communities where explainability is crucial.

Lastly, our evaluation was primarily conducted through human assessments on arXiv papers, which, while valuable, does not fully capture the generalization capabilities of MLBCAP across a broader range of scientific literature. To rigorously validate the robustness and adaptability of the model, future evaluations should include a diverse set of scientific documents.

## 8 Conclusion

In this paper, we presented MLBCAP, a novel framework for generating high-quality captions for scientific figures through the collaborative utilization of multiple Large Language Models (LLMs). Unlike prior approaches that rely on isolated modalities or limited data perspectives, MLBCAP uniquely integrates textual and visual features alongside a filtering mechanism to ensure that only high-quality training data is utilized. By combining the complementary strengths of multiple LLMs with candidate caption generation and a post-editing stage, our framework generates captions that are not only preferred over author-written captions in informativeness but also cater to diverse needs through long and short caption formats. In addition, our results highlight the effectiveness of a multi-LLM approach, demonstrating higher caption quality compared to a single prominent LLM like GPT-4o.

## 9 Acknowledgement

We are grateful to the anonymous reviewers for their valuable feedback. This research, along with the SciCap Challenge 2024, was partially supported by the Alfred P. Sloan Foundation (Grant Number: 2024-22721).

## References


1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
3. Chan, D., Petryk, S., Gonzalez, J., Darrell, T., Canny, J.: CLAIR: Evaluating image captions with large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 13638–13646. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.841, https://aclanthology.org/2023.emnlp-main.841
4. Chao, D., Song, X., Zhong, S., Wang, B., Wu, X., Zhu, C., Yang, Y.: The solution for the iccv 2023 1st scientific figure captioning challenge. arXiv preprint arXiv:2403.17342 (2024)
5. Contributors, O.: Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass (2023)
6. Fang, C., Li, X., Fan, Z., Xu, J., Nag, K., Korpeoglu, E., Kumar, S., Achan, K.: Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2910–2914. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3661357, https://doi.org/10.1145/3626772.3661357
7. Hartley, J.: Single authors are not alone: Colleagues often help. Journal of Scholarly Publishing **34**(2), 108–113 (2003)
8. Hsu, T.Y., Giles, C.L., Huang, T.H.: SciCap: Generating captions for scientific figures. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 3258–3264. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.findings-emnlp.277, https://aclanthology.org/2021.findings-emnlp.277
9. Hsu, T.Y., Huang, C.Y., Huang, S.H., Rossi, R., Kim, S., Yu, T., Giles, C.L., Huang, T.H.K.: Scicapenter: Supporting caption composition for scientific figures with machine-generated captions and ratings. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–9 (2024)
10. Hsu, T.Y., Huang, C.Y., Rossi, R., Kim, S., Giles, C., Huang, T.H.: GPT-4 as an effective zero-shot evaluator for scientific figure captions. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 5464–5474. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.363, https://aclanthology.org/2023.findings-emnlp.363
11. Huang, C.Y., Hsu, T.Y., Rossi, R., Nenkova, A., Kim, S., Chan, G.Y.Y., Koh, E., Giles, C.L., Huang, T.H.: Summaries as captions: Generating figure captions for scientific documents with automated text summarization. In: Keet, C.M., Lee, H.Y., Zarrieß, S. (eds.) Proceedings of the 16th International Natural Language Generation Conference. pp. 80–92. Association for Computational Linguistics, Prague, Czechia (Sep 2023). https://doi.org/10.18653/v1/2023.inlg-main.6, https://aclanthology.org/2023.inlg-main.6

12. Jiang, H., Wu, Q., Lin, C.Y., Yang, Y., Qiu, L.: LLMLingua: Compressing prompts for accelerated inference of large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 13358–13376. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.825, https://aclanthology.org/2023.emnlp-main.825
13. Kafle, K., Price, B., Cohen, S., Kanan, C.: Dvqa: Understanding data visualizations via question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5648–5656 (2018)
14. Kahou, S.E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., Bengio, Y.: Figureqa: An annotated figure dataset for visual reasoning. arXiv preprint arXiv:1710.07300 (2017)
15. Li, P., Li, T., Wang, J., Wang, B., Yang, Y.: Proposal report for the 2nd scicap competition 2024. arXiv preprint arXiv:2407.01897 (2024)
16. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
18. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: NLG evaluation using gpt-4 with better human alignment. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2511–2522. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.153, https://aclanthology.org/2023.emnlp-main.153
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017), https://api.semanticscholar.org/CorpusID:53592270
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
21. Qian, X., Koh, E., Du, F., Kim, S., Chan, J., Rossi, R.A., Malik, S., Lee, T.Y.: Generating accurate caption units for figure captioning. In: Proceedings of the Web Conference 2021. pp. 2792–2804 (2021)
22. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
23. Si, C., Shi, W., Zhao, C., Zettlemoyer, L., Boyd-Graber, J.L.: Getting moRE out of mixture of language model reasoning experts. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023), https://openreview.net/forum?id=UMywlqrW3n
24. Siegel, N., Horvitz, Z., Levin, R., Divvala, S., Farhadi, A.: 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. pp. 664–680. Springer (2016)
25. Sun, H.L., Chen, Q.G., Zhou, D.W., Zhan, D.C., Ye, H.J.: Length-adaptive caption generation with llms (2024), https://www.dropbox.com/scl/fi/vagukxna5j2ypl28zf12y/IJCAI_2024_Competition-Hailong-Sun-1.pdf?rlkey=vfqye7s7n0pj4zq31hydn3i16&e=1&dl=0

26. Yang, Z., Dabre, R., Tanaka, H., Okazaki, N.: Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. arXiv preprint arXiv:2306.03491 (2023)
27. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
28. Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., et al.: Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652 (2024)
29. Zhang, J., Zhao, Y., Saleh, M., Liu, P.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: International conference on machine learning. pp. 11328–11339. PMLR (2020)
30. Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., Han, J.: Towards a unified multi-dimensional evaluator for text generation. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 2023–2038. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.131, https://aclanthology.org/2022.emnlp-main.131

## Appendix

**Table 9.** The actual prompts we used and the text in magenta is a placeholder. In part, `[Figure]` is an image that is used as an input for MLLMs.

| Purpose | Prompt |
|---|---|
| Quality Assessment | `[Figure]`<br>### Paragraphs<br>`[Paragraphs]`<br>### Caption<br>`[Caption]`<br>Given the figure, paragraphs and caption, please rate the level of usefulness of the caption from 1 to 6 based on how well the caption could help readers understand the important information. 6 is the highest. 1 is the lowest. The answer should be JSON format: {"rating": }. |
| Figure Description (GPT-4o) | `[Figure]`<br>Your task is to describe a figure from a scientific paper. Answer your results in JSON format.<br>### Background<br>Figure is a `[Figure Type]`.<br>It is a figure about the topic `[Subject]`.<br>### Rule<br>Description of the figure should be accurate and clear. If in doubt, avoid numerical expressions. Provide the description in JSON format with the following key: description. |
| Figure Description (MiniCPM-V) | `[Figure]`<br>What is in the image? |

**Table 10.** The prompt for the caption generation. In part, `[Figure]` is an image that is used as an input for MLLMs.

| Purpose | Prompt |
|---|---|
| Caption Generation (Few-shot) | Your task is to create a caption that summarizes based on a paragraph.<br>### Figure Caption<br>The format of a Figure Caption is Declarative title + Description + Statistical information (optional).<br>Declarative title: summarises the result or major finding of the data you are presenting in the figure. (A mere representation of the x and y axes cannot be a title.)<br>Description: a brief description of the results necessary for understanding the figure without having to refer to the main text<br>Statistical information: for example, number of replicates, asterisks denoting P-values, statistical tests, etc.<br>### Background<br>Figure is a `[Figure Type]`.<br>Figure is a category related to `[Subject]`.<br>### Rule<br>Caption MUST have a word count of 60 words or less.<br>Caption MUST have a tone and sentence structure appropriate for a top-tier conference (e.g., NeurIPS, ICLR, CVPR, ACL, EMNLP).<br>It is not a caption to describe the x-axis y-axis.<br>Caption MUST be clear, concise, consistent, and provide specific information, especially not false.<br>If the given paragraph uses abbreviations, use them in the caption.<br>### Best Caption Examples<br>[Few-shot Examples]<br>### Input<br>Paragraph: [Paragraphs]<br>Figure Summary: [Figure Description]<br>Mention: [Mentions]<br>OCR text: [OCR]<br>### Output format<br>Answer results in JSON format: {"caption": }. |

**Table 11.** The prompt for the selecting best caption and post-editing process. The `[Max Len]` is the constraint of caption lengths (Long: 50, Short: 30).

| Purpose | Prompt |
|---|---|
| Judgement | A good figure caption should include the following elements:<br>1. **Clear Description**: Clearly describe what the figure represents so that readers can understand the main point of the figure just by reading the caption.<br>2. **Conciseness**: Keep it concise while including all essential information. The caption MUST be brief yet informative (important!!).<br>3. **Relevant Information**: Include background information, experimental conditions, or methods used that are necessary to understand the figure. This helps the reader interpret the data correctly.<br>4. **Consistency**: Maintain consistency with the rest of the paper in terms of terminology and style. Ensure that the terms used in the caption match those used in the text.<br>5. **Citation**: If necessary, include citations of related research or references in the paragraph.<br>You are given a summarization of the figure, relevant paragraphs, a mentioned sentences, and four caption candidates:<br>### Summarization of the Figure<br>[Figure Description]<br>### Paragraph<br>[Paragraphs]<br>### Mention<br>[Mentions]<br>### Caption A<br>[Pegasus Caption]<br>### Caption B<br>[LLaMA-3-8B Caption]<br>### Caption C<br>[Yi-1.5-9B Caption]<br>### Caption D<br>[GPT-4o Caption]<br>1. Choose the best and worst caption and answer in JSON format (For example, if A is the best and B is the worst, the answer is: "Good": "A", "Bad": "B). Candidate captions shouldn't be scored low just because they're concise.<br>2. If even the best caption could be improved, use the candidate captions and paragraphs to improve it (math symbols, legend, grammar, etc.).<br>3. The improved sentence should have a tone and sentence structure appropriate for a top-tier conference (e.g., NeurIPS, ICLR, CVPR, ACL, EMNLP) and MUST have a word count of [Max Len] words or less.<br>4. If you find that sentences are becoming long and complex, making it difficult for readers to understand, break the sentences up to effectively convey the important information.<br>5. If you already provided a perfect caption, keep it the same.<br>6. Do not omit the figure numbers, such as in "Fig. 3" or "Figure 5". Provide them in JSON format with the following keys: Good, Bad, Improved Caption "Good" : "", "Bad" : "", "Improved Caption": "" |