AnaFig: A Human-Aligned Dataset for Scientific Figure Analysis

Tan Yue†
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
yuetan@pku.edu.cn

Xuzhao Shi†
Beijing University of Posts and Telecommunications
Beijing, China
sxzs@bupt.edu.cn

Rui Mao
Nanyang Technological University
Singapore, Singapore
rui.mao@ntu.edu.sg

Zilong Song
Beijing University of Posts and Telecommunications
Beijing, China
sozilo@bupt.edu.cn

Zonghai Hu
Beijing University of Posts and Telecommunications
Beijing, China
zhhu@bupt.edu.cn

Dongyan Zhao*
Wangxuan Institute of Computer Technology, Peking University
State Key Laboratory of General Artificial Intelligence
Beijing, China
zhaodongyan@pku.edu.cn

Abstract

Scientific Figure Analysis (SFA) aims to derive analytical insights from figures while incorporating background instructions. Unlike conventional tasks such as figure captioning or description generation, which focus on extracting surface-level information from the sole visual modality, SFA requires an intelligent system to summarize key patterns, infer implications, and contextualize scientific findings from visual and textual inputs. It demands not only visual recognition but also the integration of scientific knowledge, multimodal understanding, and contextual reasoning. In this work, we introduce an SFA dataset, AnaFig, comprising 2,000 high-quality samples across 56 domains. All samples are evaluated by using human-aligned five-dimensional scoring criteria, resulting 10,000 human-annotated score labels. The AnaFig dataset facilitates the assessment of three critical capabilities of multimodal large language models (MLLMs): adherence to complex instructions, multimodal perception, and analytical summarization. By building a new benchmark with widely used MLLMs, this study contributes to scientific knowledge discovery and reasoning, fostering the alignment of MLLMs and human experts in scientific analysis.

CCS Concepts

• **Computing methodologies** → **Language resources**; **Computer vision**; • **Applied computing** → **Physical sciences and engineering**; **Document management and text processing**.

†Equal contribution.
*Corresponding author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758226

Keywords

Scientific Figure Analysis, Benchmark Dataset, Multimodal Large Language Model, Human-Aligned Evaluation

ACM Reference Format:
Tan Yue, Xuzhao Shi, Rui Mao, Zilong Song, Zonghai Hu, and Dongyan Zhao. 2025. AnaFig: A Human-Aligned Dataset for Scientific Figure Analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3746027.3758226

1 Introduction

Scientific Figure Analysis (SFA) is challenging yet significant due to its role in discovering meaningful insights beyond surface-level figure descriptions [6, 7, 14, 37] and its potential to enhance the accessibility and understanding of complex scientific information. Unlike conventional figure captioning tasks, SFA demands a deep understanding of data patterns, relationships, and scientific implications from visual and textual inputs. This necessitates integrating visual recognition with scientific background knowledge [34], contextual reasoning [22], and multimodal comprehension [18].

Multimodal large language models [1, 4] (MLLMs) have significantly advanced image-to-text tasks [38], e.g., image captioning [39] and image description generation [32]. However, their effectiveness in high-level analytical tasks remains unclear. Unlike conventional image-to-text tasks, SFA requires not only extracting visual information but also contextualizing scientific backgrounds, comprehending multimodal data (e.g., charts and accompanying text), and generating summaries that convey analytical insights. As shown in Fig. 1, existing image-to-text datasets are insufficient to capture these complexities [12, 16, 37]. They often focus on descriptive accuracy rather than analytical depth, lacking the scientific context necessary for evaluating a model's ability to reason about data.

To address this gap, we introduce AnaFig[1], a novel dataset for evaluating three key capabilities of MLLMs: 1) adherence to complex scientific instructions, 2) multimodal perception of visual and textual context, and 3) the ability to analyze and summarize scientific information. AnaFig comprises 2,000 samples drawn from

[1]https://github.com/yuetanbupt/AnaFig

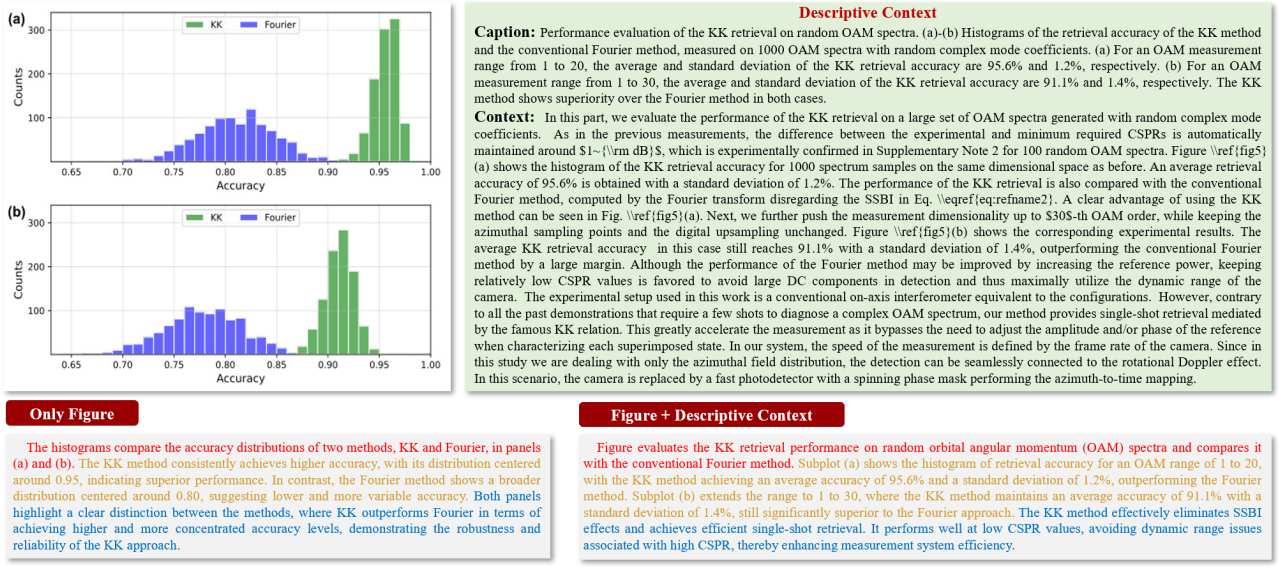

Figure 1: Importance of descriptive contextual information on the quality of analytical summaries. Different color fonts represent the corresponding different qualities of the generated content.

56 domains. Each sample pairs a figure with its corresponding descriptive text as input, sourced directly from research papers. This descriptive text incorporates both the figure's caption and additional descriptive content from the paper, providing a rich context and instructions for MLLMs to interpret. The expected output is an analytical summary that synthesizes key insights from the figure. While figures serve as the primary source for interpretation, the descriptions provide essential instructions, enabling MLLMs to contextualize and derive meaningful insights. We also developed a five-dimensional scoring framework assessing faithfulness, completeness, conciseness, logicality, and analysis. Ground truth summaries are meticulously crafted by human experts through multiple refinement rounds to uphold high-quality standards across these dimensions. Thus, the dataset provides 2,000 expert-written summaries along with 10,000 evaluation score labels, establishing a robust benchmark for assessing MLLM capabilities in more challenging SFA task, which is important for scientific knowledge discovery, multimodal understanding, and AI-human alignment.

As a new benchmark, we test eight widely used MLLMs on AnaFig, namely, Qwen2-VL-2B and 7B, InterVL-2.5-8B, MiniCPM-V-2.6, GPT-4o, Gemini-1.5-flash, Claude-3-haiku, and Claude-3.5-sonnet, and report their performance from the perspectives of text generation (e.g., BLEU, METERO, BERT score, ROUGE, and MLLM score) and summary quality (e.g., the aforementioned five dimensions). Our results reveal a significant performance gap between MLLM-generated analyses and human-level expertise. By evaluating MLLM performance through the five-dimensional framework, we identify specific strengths and weaknesses of each MLLM, e.g., strong in logicality but weak in the depth of analysis. An ablation study further explores the impact of the descriptive text on performance. Finally, we investigate potential biases in MLLM-based evaluators when applied to this task.

The contributions can be summarized as follows: 1) We introduce AnaFig, a novel dataset designed to examine the boundaries of MLLM capabilities in the complex task of SFA. It comprises 2,000 gold-standard analytical summaries, carefully curated by human experts. 2) Human-aligned evaluation criteria for SFA are proposed with 10,000 manually annotated score labels. 3) Leveraging AnaFig, we establish a benchmark by testing prominent MLLMs, providing a comprehensive assessment in this challenging domain.

2 Related Works

2.1 Evaluation of MLLMs in Visual Tasks

Recent advancements in MLLMs [47] have demonstrated significant progress in visual understanding tasks [11, 42, 45], such as visual questions answering [19, 44], image captioning [23, 38]. These models are evaluated primarily on the basis of their ability to correctly answer or generate descriptive text that accurately reflects the visual content of the image. However, the evaluation of MLLMs [9, 15] in more complex visual reasoning tasks, particularly those requiring scientific or analytical insight, remains limited. While some studies have made strides in addressing basic visual recognition and textual generation [25, 32, 33, 36, 41, 46], they often overlook the ability of these models to perform high-level reasoning on visual data, such as interpreting trends, relationships, or implications embedded in scientific figures.

Current MLLM benchmarks largely emphasize surface-level understanding over scientific reasoning and analysis. Thus, specialized benchmarks are needed to assess MLLMs' ability to interpret scientific data and extract insights from complex visual information.

2.2 Existing Image-to-Text Datasets

As shown in Table 1, in the field of image-to-text generation, various datasets have been developed to help train and evaluate MLLM [35],

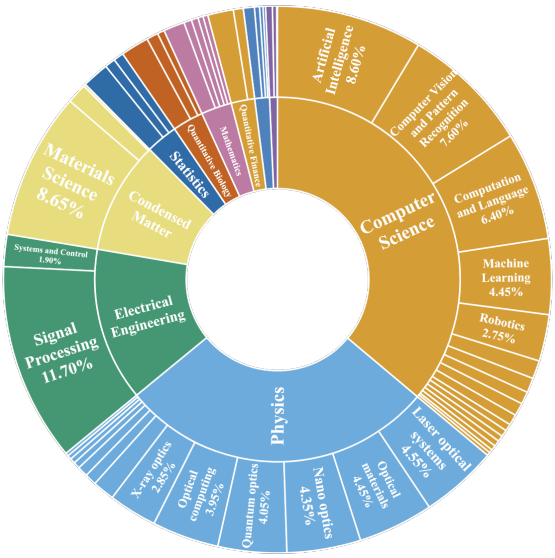

Figure 2: Statistics of figure application domains.

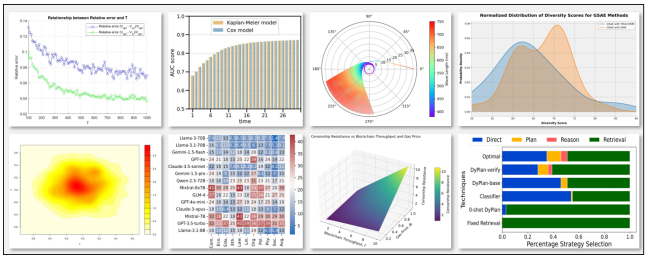

Figure 3: Examples of figure types.

Table 1: The comparison with related datasets. R.F.=Real Figure, Cap.=Caption. Con.=Context, Ana.=Analysis. Five-D evaluation=Five-dimension evaluation.

Dataset	R.F.	Cap.	Con.	Ana.	Evaluation
SCICAP [14]	✓	✓	✗	✗	Rule-based
ChartSumm [29]	✓	✗	✗	✗	Rule-based
Chart-to-text [16]	✓	✓	✗	✗	Rule-based
ChartX [37]	✗	✗	✗	✗	GPT-Score
AnaFig (Ours)	✓	✓	✓	✓	Five-D evaluation

such as the Microsoft COCO [21], Visual Genome [17], SciCap [14], and more recently developed ChartX [37] focusing on figure understanding. These datasets provide a large amount of visual content with corresponding textual descriptions and have contributed significantly to advancing the field of image caption and description generation [30]. However, these datasets have shown limitations when it comes to scientific reasoning tasks, especially those involving complex visual analysis or domain-specific knowledge.

One significant shortcoming of current image-to-text datasets is the reliance on non-expert annotations. While crowdsourced data is valuable for training models in general tasks, it often lacks the depth

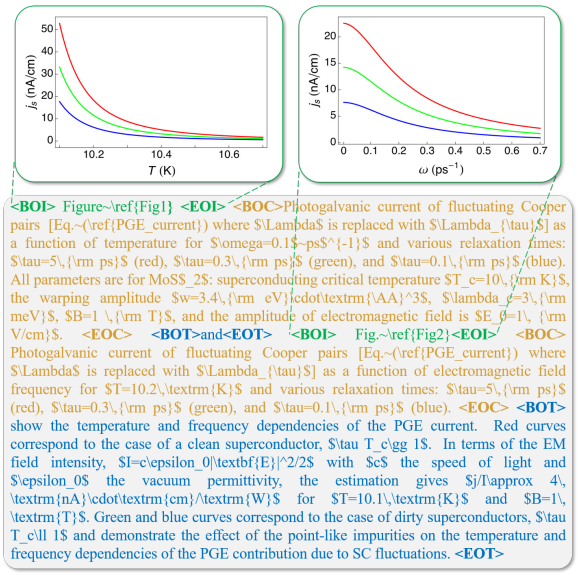

Figure 4: An example input of AnaFig.

required for scientific analysis [10, 26]. In addition, the annotation process for these datasets often lacks systematic criteria for assessing the quality of reasoning. The lack of expert-level annotations and evaluation criteria for existing datasets implies that current datasets do not fully capture the nuances of scientific reasoning in multimodal data [13, 43]. This limitation is particularly evident in SFA, where the ability to reason about trends, relationships, and meanings in data visualizations is critical.

3 AnaFig-Dataset

3.1 Dataset Collection

As shown in Fig. 2, the AnaFig dataset is constructed from academic papers, including 10 major domains (e.g. Computer Science, Physics, etc.) and 56 sub-domains (e.g. Quantum optics, Photoexcitations, etc.). All data are collected from an open access academic website (Arxiv). We filter out figures with low resolution or poor image quality. To ensure data diversity, we manually remove specific types of figures, such as flowcharts, and sample schematics. Finally, we retain 2,000 high-quality samples cover 18 figure types (shown in Fig. 3), selected from over 6,000 academic papers.

3.2 Data Formatting

In real-world scenarios, most figure and text data do not match each other exactly. The interleaved figure and text processing paradigm is more generalized. Fig. 4 shows the figure-text interleaved input structure of AnaFig. Figures and text are integrated at the token level, with special tokens delineating the different modalities. The input format is as follows: <BOI>Figure<EOI><BOC> Caption<EOC><BOT>Context<EOT>. <BOI> and <EOI> mark the beginning and end of a figure's representation, <BOC> and <EOC> similarly denote the caption, and <BOT> and <EOT> enclose the surrounding scientific context, including any necessary instructions. If there are multiple figures, we instruct the model to summarize

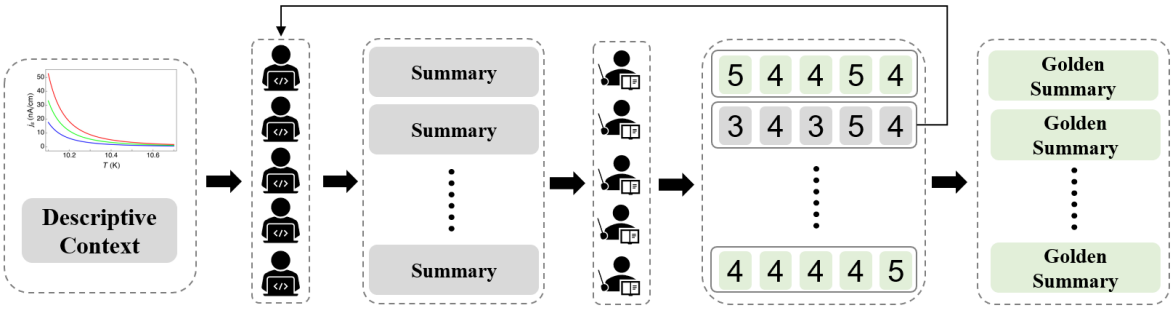

Figure 5: AnaFig dataset annotation and scoring process.

Faithfulness	**5 points:** The summary is scrupulously faithful to the content of the figure and related descriptions, and is completely free of error, misinformation, or speculation. **4 points:** The summary is generally faithful to the content of the figure, with only very minor deviations or inaccuracies that do not affect overall understanding. **3 points:** The summary is mostly faithful to the figure, but contains some significant biases or inaccurate information. **2 points:** The summary contains multiple deviations or errors, and some parts are grossly inconsistent with the information in the figure. **1 point:** The summary does not match the information in the figure, contains numerous errors or speculative information, and does not reflect the true content.
Completeness	**5 points:** The summary covers all the key information and trends in the figure and is complete without any omissions. **4 points:** The summary covers most of the important information, but slightly omits some minor information. **3 points:** The summary covers the main points of the figure, but some important information is not included. **2 points:** The summary covers only some of the information in the figure, leaving out several important aspects or details. **1 point:** The summary of most of the important content was ignored, with only a small portion of the content or details being focused on.
Conciseness	**5 points:** The summary is concise and clear, without redundancy, and effectively conveys the core information of the figure in a minimum number of words. **4 points:** The summary is concise, with a small amount of redundancy, but it does not detract from the clear communication of the information. **3 points:** The summary has some redundant or repetitive content, which slightly detracts from the overall effectiveness of the communication. **2 points:** The summary contains a lot of redundant information and the core content is not sufficiently salient, which affects comprehension. **1 point:** The summary is very lengthy and cluttered, making it difficult to highlight the main information of the figure.
Logicality	**5 points:** The summary is well-structured, logically clear, and linguistically fluent; the unfolding of the individual points is consistent with the internal logic of the expert's knowledge and the background information, and there are no logical contradictions or errors in reasoning. **4 points:** The summary logic is generally clear, the content is well-structured, and the language is fluent; some in-depth connections to background or expert knowledge may be missing in a few sections, but overall it is consistent with common sense. **3 points:** The summary's logic is faulty, parts of it unfold in a way that is not entirely consistent with background or expert knowledge, or there is a lack of fluency that affects comprehension. **2 points:** The summary lacks logic, the language organization is confusing, and some of the content contains significant deviations from background knowledge or contradictions in reasoning. **1 point:** The summary is illogical, the content is disorganized, the language is not fluent, and it completely contradicts background knowledge in key parts, resulting in an inability to properly understand the core information of the figure.
Analysis	**5 points:** The summary has a deep understanding of the figures and related descriptions, and the analysis and explanations are completely correct and reasonable. The analysis is insightful and comprehensive. **4 points:** The summary demonstrates a good understanding of the figure, with sound analysis and only minor deficiencies or inadequate explanations. **3 points:** The summary demonstrates a basic understanding of the figure, but there are clear misunderstandings or inadequate explanations. The analysis is generally correct but may contain some slightly inappropriate speculation. **2 points:** The summary has a significant bias in understanding the figure, and the analysis is superficial or includes noticeable hallucinations.. **1 point:** The summary fails to understand the information in the figure correctly, and the analysis and interpretation are completely wrong or irrelevant.

Figure 6: Detailed evaluation criteria.

one of the figures (target_figure), with the other figures serving as background knowledge.

3.3 The Annotation Process and Evaluation Criteria

The gold-standard summaries in our AnaFig dataset are meticulously crafted by ten human experts (graduate students or researchers with PhD degrees). We leverage an annotation-evaluation-refinement process to ensure the quality of generated summaries. The expert annotation process is illustrated in Fig. 5. Five experts serve as initial annotators, generating summaries based on our established annotation criteria (see Fig. 6). Next, the other five experts act as checkers, independently evaluating and scoring all 2,000 data samples. We use five-dimensional evaluation criteria, e.g., faithfulness, completeness, conciseness, logicality, and analysis. **Faithfulness:** The summary must strictly adhere to the information presented in the figures and descriptions. **Completeness:** The summary should encompass all key information and trends depicted in the figures. **Conciseness:** The summary should avoid redundant information or non-critical details. **Logicality:** The summarized content should be logically coherent and consistent with common

Table 2: Final scoring statistics of the AnaFig dataset.

Score	Faith.	Compl.	Conci.	Logic.	Analy.	%
Avg.	4.78	4.52	4.37	4.56	4.66	-
5	1558	1046	736	1121	1312	57.73
4	442	954	1264	879	688	42.27
3/2/1	0	0	0	0	0	0
# Samp.	2000	2000	2000	2000	2000	100

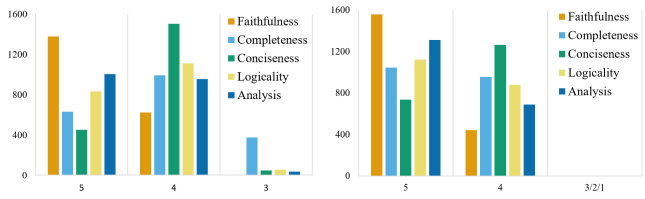

Figure 7: AnaFig dataset scoring results statistics. Results in the first annotation round (left, with no sample scored 2/1). Results in the final annotation round (right, with no sample scored 3/2/1).

sense and expert knowledge. **Analysis:** The summary's analysis should be insightful and demonstrate a thorough understanding of the data.

Checkers assign scores from 1 to 5 for each criterion. Summaries scoring 3 or lower on any dimension are iteratively revised until all samples achieve a score of 4 or higher across all five dimensions. The final scores are shown in Table 2. The improvements between the initial and final annotation tasks can be viewed in Fig. 7.

4 Experimental Setup

To cover multiple sizes and diverse MLLMs, representative models including Qwen2-VL-2B [32], Qwen2-VL-7B [32]), MiniCPM-V2.6 [40], InterVL-2.5-8B [8], GPT-4o [1], Claude-3-haiku [2], Claude-3.5-sonnet [3]) and Gemini-1.5-flash [31]) are selected for testing to establish benchmarks.

We conduct zero-shot testing for the above models. For open-source models, we obtain the models and their pre-training weights and use the default hyper-parameter configurations during testing to ensure the reproducibility of the testing process and the comparability of the results. For proprietary models, we used API for testing by requesting them via their official websites to ensure that the performance of these models is evaluated in a fair environment. The instructions used to direct MLLMs to generate summaries are shown in Fig. 8. We report five evaluation metrics, widely used in generation tasks, namely ROUGE (R1=ROUGE1, R2=ROUGE2, RL=ROUGEL) [20], BLEU [28], METEOR (MET.) [5], BERTScore [48], and MLLM-Score. MLLM Score is designed to align human experts for scoring. We Instruct Gemini-1.5-flash to perform reference-free scoring by inputting *[Figure, Descriptive Text, Pre_summary]* (Pre_summary denote the model generated summary), with human-aligned scoring criteria (shown in Fig. 6) as part of the instruction.

Instruction

Given the following information:1) Figure, 2) Caption, 3) Context.
Based on the above information, analyse and summarize the figure in detail, paying attention to the following requirements: 1) The summary should be centred on the information in the figure. "Caption" and "Context" can be used as background information, but the summary should focus on the information in the figure. 2) There may be multiple figure inputs, and only the "Target_figure" needs to be summarized, with the other figures serving only as background information. 3) The generated summary should concisely and clearly summarize the information conveyed by the figure in no more than 200 words.
Please note the following five-dimensional evaluation criteria, which is the key point for evaluating the quality of the summary.
Faithfulness: The summary must strictly adhere to the information presented in the figures and descriptions.
Completeness: The summary should encompass all key information and trends depicted in the figures.
Conciseness: The summary should avoid redundant information or non-critical details.
Logicality: The summary should be logically coherent and consistent with common sense and expert knowledge.
Analysis: The summary's analysis should be insightful and demonstrate a thorough understanding of the data.

Figure 8: The instruction used to generate the summaries. Target_figure is marked during data processing.

Table 3: Results of various evaluation methods in sample-level.

Model	BLEU	MET.	BERT Score	ROUGE			MLLM Score
				R1	R2	RL	
Qwen2-2B	0.0954	0.2878	0.1750	0.4605	0.2103	0.3135	3.36
MiniCPM	0.0991	0.3621	0.2550	0.5026	0.2180	0.3165	3.82
InterVL2.5	0.0645	0.3126	0.2154	0.4618	0.1728	0.2792	3.73
Qwen2-7B	**0.1214**	**0.3846**	0.2585	0.5051	**0.2509**	**0.3423**	3.80
Claude-3	0.1003	0.3810	0.2654	0.4792	0.2252	0.3106	3.89
GPT-4o	0.0893	0.3204	0.2931	0.5148	0.2067	0.3218	3.90
Gemini-1.5	0.0993	0.3330	**0.2960**	**0.5222**	0.2159	0.3228	3.95
Claude-3.5	0.1024	0.3645	0.2903	0.5114	0.2274	0.3153	**3.98**

5 Results

5.1 Main Results

We observe a clear positive correlation between model size and performance in Table 3, with the larger Qwen2-(VL)-7B model consistently outperforming the smaller Qwen2-(VL)-2B across all metrics. Qwen2-7B also exceeds other MLLMs in the majority of metrics. Second, Almost all models perform poorly on the reference-based metrics (ROUGE, BLEU, METEOR, BERTScore), which suggests that the current MLLMs still have a large gap with humans in SFA tasks. It also suggests that the semantic similarity-based computation has limitations in aligning with humans. Third, the Gemini and Claude models demonstrate competitive performance, with Claude-3.5 achieving the highest MLLM-based evaluation score, suggesting its potential for generating more human-like and insightful summaries. However, the differences among these top-performing models are relatively small, indicating that further improvements are needed to bridge the gap to human-level analytical capabilities in scientific figure analysis.

Table 4 presents a granular evaluation of the MLLM-generated summaries using a five-dimensional framework. While all MLLMs achieve relatively high scores in "Logicality", their performance varies significantly across other dimensions. Notably, the "Analysis" proves particularly challenging, with scores considerably lower than those for "Faithfulness" or "Completeness". This suggests that

Table 4: MLLM Score in five-dimensional evaluation.

	Fai.	Com.	Con.	Log.	Ana.	Avg.
Human	4.78	4.52	4.37	4.56	4.66	4.58
Qwen2-2B	3.48	2.95	3.97	3.66	2.72	3.36
MiniCPM	3.80	3.77	3.84	4.18	3.51	3.82
InternVL2.5	3.64	3.72	3.80	4.05	3.47	3.74
Qwen2-7B	3.84	3.60	3.94	4.26	3.35	3.80
Claude-3	3.85	3.76	3.98	4.27	3.57	3.89
GPT-4o	3.88	3.63	4.16	4.26	3.58	3.90
Gemini-1.5	3.87	3.65	**4.35**	4.38	3.49	3.95
Claude-3.5	**3.91**	**3.79**	4.05	**4.40**	**3.74**	**3.98**
Average	3.78	3.61	4.01	4.18	3.43	3.80

Table 5: Ablation study with machine evaluation.

	BLEU	MET.	BERT Score	ROUGE			MLLM Score
				R1	R2	RL	
w/	0.1129	0.4091	0.2953	0.4825	0.2211	0.3041	3.95
w/o	0.0737	0.2860	0.2637	0.4081	0.1076	0.2357	3.51
Δ	↓0.0392	↓0.1231	↓0.0316	↓0.0744	↓0.1135	↓0.0684	↓0.44

Table 6: Ablation study with MLLM and human evaluation.

	Fai.	Com.	Con.	Log.	Ana.	Avg.
MLLM Score						
w/ discript.	3.86	3.78	4.17	4.25	3.68	3.95
w/o discript.	3.19	2.95	3.96	4.01	3.42	3.51
Human Eval.						
w/ discript.	3.63	3.58	3.89	4.01	3.42	3.71
w/o discript.	2.53	2.01	3.59	2.15	2.18	2.49

current MLLMs struggle to provide insightful interpretations and go beyond surface-level descriptions of the figures. Overall, all MLLMs can generate concrete summaries, given their comparatively higher scores on average.

5.2 Effectiveness of Descriptive Text

Our central hypothesis with the AnaFig dataset is that the accompanying descriptive text is crucial for generating insightful figure summaries, as it provides essential scientific background. To validate this, we perform an ablation study using 100 summaries generated by Claude-3, evaluating the impact of removing the descriptive text. We employ both machine-based metrics and human evaluation.

Table 5 demonstrates that excluding the descriptive text leads to significant reductions across all machine evaluation metrics. It shows that the tested model (Claude-3) lacks knowledge in the field of scientific analysis and is unable to analyze complex figure information without scientific background knowledge. Specifically, the largest decreases are observed in "Faithfulness" and "Completeness" within the MLLM-based scores in Table 6. Interestingly, despite the sharp drop in completeness, the "Analysis" score remains relatively high even without the descriptive text, which seems counterintuitive. This discrepancy prompts us to conduct the human evaluation. The human evaluation reveals that the most substantial performance drops occur in "Completeness" and, importantly, "Analysis". This difference between machine and human evaluation in assessing analytical depth raises concerns about the reliability of MLLM-based evaluators for nuanced analytical tasks and suggests potential biases in their assessments [27].

Table 7: Averaged MLLM scores over the five evaluation dimensions, based on evaluators with different MLLM.

	Qwen2	Gemini	Claude	Avg.
Qwen2 Evaluator	3.71	3.66	3.63	3.67
Gemini Evaluator	3.80	3.95	3.89	3.88
Claude Evaluator	3.84	3.99	3.94	**3.92**

5.3 The Bias of MLLM Evaluators

The observed discrepancies between machine and human evaluations raise a crucial question: are MLLM-based evaluators inherently biased towards summaries generated by models sharing their architectural foundation?

To investigate this, we employ a controlled evaluation setup. Specifically, we use Qwen2-VL-7B, Gemini-1.5-flash, and Claude-3-haiku as MLLM evaluators, each assessing summaries produced by models with the same underlying architecture. The results, presented in Table 7, reveal a clear trend: MLLM evaluators generally assign higher scores to summaries generated by the same foundation MLLM. This suggests a potential self-referential bias in MLLM-based evaluation. Furthermore, our findings indicate that the Claude evaluator, in particular, exhibits a tendency towards leniency, yielding the highest average MLLM scores across all evaluated models. This observation highlights the importance of carefully considering potential biases [24] when relying on MLLM-based evaluation for challenging SFA tasks, especially when comparing models with varying architectures. The implication is that direct comparisons of MLLM-generated summaries using MLLM-based evaluators may be skewed, and human evaluation remains crucial for obtaining a more objective and comprehensive assessment of summary quality.

6 Conclusion

This paper introduces AnaFig, a new benchmark dataset designed to advance the challenging task of SFA, focusing on generating insightful figure summaries. Evaluation of widely-used MLLMs on AnaFig reveals that while these models can produce coherent and logical text from figures, they exhibit weaknesses in both faithfulness and completeness. Critically, their ability to provide in-depth analysis falls considerably short of human expert performance. Our ablation study further highlights the importance of incorporating relevant scientific background knowledge to enable MLLMs to generate more complete and insightful summaries from figures. However, we also observe that MLLM-based evaluators may be susceptible to biases when assessing summaries in SFA.

References

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

[2] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

[3] Anthropic. 2024. Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet.

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

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

[6] 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*. 1537–1545.

[7] 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 arXiv:1906.02850* (2019).

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

[9] Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657* (2023).

[10] Ruifeng Guo, Jingxuan Wei, Linzhuang Sun, Bihui Yu, Guiyong Chang, Dawei Liu, Sibo Zhang, Zhengbing Yao, Mingjun Xu, and Liping Bu. 2023. A survey on image-text multimodal models. *arXiv preprint arXiv:2309.15857* (2023).

[11] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. 2016. Deep learning for visual understanding: A review. *Neurocomputing* 187 (2016), 27–48.

[12] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. ChartLlama: A Multimodal LLM for Chart Understanding and Generation. arXiv:2311.16483 [cs.CV] https://arxiv.org/abs/2311.16483

[13] Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Poonam Poonam, Michael Glöckler, Alex Bäuerle, and Timo Ropinski. 2024. A Survey on Quality Metrics for Text-to-Image Generation. *arXiv preprint arXiv:2403.11821* (2024).

[14] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. 2021. SciCap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624* (2021).

[15] Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769* (2024).

[16] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486* (2022).

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017).

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

[19] Panfeng Li, Qikai Yang, Xieming Geng, Wenjing Zhou, Zhicheng Ding, and Yi Nian. 2024. Exploring diverse methods in visual question answering. *arXiv preprint arXiv:2404.13565* (2024).

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

[21] 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*. Springer, 740–755.

[22] Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2024. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems* 36 (2024).

[23] Tianrui Liu, Qi Cai, Changxin Xu, Bo Hong, Jize Xiong, Yuxin Qiao, and Tsungwei Yang. 2024. Image Captioning in news report scenario. *arXiv preprint arXiv:2403.16209* (2024).

[24] Rui Mao, Guanyi Chen, Xiao Li, Mengshi Ge, and Erik Cambria. 2025. A Comparative Analysis of Metaphorical Cognition in ChatGPT and Human Minds. *Cognitive Computation* 17, 35 (2025), 1–12.

[25] Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A Survey on Assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Torino, Italia, 7844–7866.

[26] Rui Mao, Kai He, Claudia Beth Ong, Qian Liu, and Erik Cambria. 2024. MetaPro 2.0: Computational Metaphor Processing on the Effectiveness of Anomalous Language Modeling. In *Findings of the Association for Computational Linguistics: ACL*. Association for Computational Linguistics, Bangkok, Thailand, 9891–9908.

[27] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection. *IEEE Transactions on Affective Computing* 14, 3 (2023), 1743–1753.

[28] 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*. 311–318.

[29] Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. ChartSumm: A Comprehensive Benchmark for Automatic Chart Summarization of Long and Short Summaries.. In *Canadian AI*.

[30] RANJAN Sapkota, SHAINA Raza, MAGED Shoman, A Paudel, and M Karkee. 2025. Multimodal large language models for image, text, and speech data augmentation: A survey. *arXiv preprint arXiv:2501.18648* (2025).

[31] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530 [cs.CL] https://arxiv.org/abs/2403.05530

[32] Qwen Team. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[33] Heng Wang, Tan Yue, Xiang Ye, Zihang He, Bohan Li, and Yong Li. 2022. Revisit finetuning strategy for few-shot learning to transfer the emdeddings. In *The Eleventh International Conference on Learning Representations*.

[34] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. 2022. Integrating scientific knowledge with machine learning for engineering and environmental systems. *Comput. Surveys* 55, 4 (2022), 1–37.

[35] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

[36] Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. 2025. AntiLeakBench: Preventing Data Contamination by Automatically Constructing Benchmarks with Updated Real-World Knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Vienna, Austria.

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

[38] Liming Xu, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. 2023. Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing* 546 (2023), 126287.

[39] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2024. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems* 36 (2024).

[40] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800* (2024).

[41] Tan Yue, Zihang He, Chang Li, Zonghai Hu, and Yong Li. 2022. Lightweight fine-grained classification for scientific paper. *Journal of Intelligent & Fuzzy Systems* 43, 5 (2022), 5709–5719.

[42] Tan Yue, Yong Li, and Zonghai Hu. 2021. Dwsa: An intelligent document structural analysis model for information extraction and data mining. *Electronics* 10, 19 (2021), 2443.

[43] Tan Yue, Yong Li, Xuzhao Shi, Jiedong Qin, Zijiao Fan, and Zonghai Hu. 2022. Papernet: A dataset and benchmark for fine-grained paper classification. *Applied Sciences* 12, 9 (2022), 4554.

[44] Tan Yue, Rui Mao, Xuzhao Shi, Shuo Zhan, Zuhao Yang, and Dongyan Zhao. 2025. QAEval: Mixture of Evaluators for Question-Answering Task Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14717–14730.

[45] Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. 2023. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion* 100 (2023), 101921.

[46] Tan Yue, Xuzhao Shi, Rui Mao, Zonghai Hu, and Erik Cambria. 2024. SarcNet: a multilingual multimodal sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 14325–14335.

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

[48] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).