ExpertNeurons at SciVQA-2025: Retrieval Augmented VQA with Vision Language Model (RAVQA-VLM)

Nagaraj Bhat
AI Researcher
nagbhat25@gmail.com

Joydeb Mondal
AI Researcher
joydeb28@gmail.com

Srijon Sarkar
AI Researcher
srijonsarkar41@gmail.com

Abstract

We introduce **RAVQA-VLM**, a Retrieval-Augmented Generation (RAG) architecture with Vision Language Model for the SciVQA challenge, which targets closed-ended visual and non-visual questions over scientific figures drawn from ACL Anthology and arXiv papers. Our system first encodes each input figure and its accompanying metadata (caption, figure ID, type) into dense embeddings, then retrieves context passages from the full PDF of the source paper via a Dense Passage Retriever. The extracted contexts are concatenated with the question and passed to a vision-capable generative backbone (e.g., Qwen-2.5, Pixtral-12B, Mistral-24B-small, InterVL-3-14B) fine-tuned on the 15.1K SciVQA training examples. We jointly optimize retrieval and generation end-to-end to minimize answer loss and mitigate hallucinations. On the SciVQA test set, RAVQA-VLM achieves significant improvements over parametric only baselines, with relative gains of +5% ROUGE1 and +5% ROUGE-L, demonstrating the efficacy of RAG for multimodal scientific QA. In this shared task, our **RAVQA-VLM** approach secured the top rank in the leaderboard with an F1 score of 0.8049 (ROUGE-1), 0.8043 (ROUGE-L), and 0.9849 (BERTScore).

1 Introduction

Scientific literature often conveys core findings through figures such as bar charts, line graphs, scatter plots, and compound diagrams. Understanding these figures requires interpreting both visual cues (e.g., color, shape, and size) and associated textual elements (e.g., captions, methodology descriptions, result interpretations) (Karishma et al., 2023; Li et al., 2024). This multimodal nature presents challenges for automated systems aiming to answer questions about scientific figures.

Traditional vision-only architectures such as standard convolutional neural networks (CNNs) and object detection models like Faster R-CNN (Ren et al., 2015) are limited to spatial and visual patterns and typically fail to reason over abstract visual encodings used in scientific plots. On the other hand, language-only models cannot perceive visual structure or layout, making them unsuitable for figure-centric reasoning tasks (Radford et al., 2021).

Recent advances in large vision-language models (LVLMs), such as InterVL-3-14B (Zhu et al., 2025), Qwen-2.5-VL (Bai et al., 2025), Phi-3.5 (Abdin et al., 2024), and Mistral-Small-24B (Mistral AI, 2025), have enabled more robust multimodal understanding. However, these models often produce hallucinated answers when key context is missing or ambiguous (Brown et al., 2020). Retrieval-Augmented Generation (RAG) offers a potential remedy by enriching model inputs with contextually relevant external passages at inference time (Lewis et al., 2020).

To accelerate research in this area, the SciVQA shared task (Borisova et al., 2025) provides a benchmark dataset of 3,000 figures from scientific documents, each accompanied by seven question-answer pairs and includes metadata such as caption, figure ID, figure type (e.g., compound, line graph, bar chart, scatter plot), QA pair type. The task emphasizes both visual and non-visual question types, facilitating comprehensive evaluation across multimodal reasoning skills (Borisova et al., 2025).

We build on these insights to propose Retrieval-Augmented Generation architecture with Vision Language Model **(RAVQA-VLM)**, a unified framework that:

1. retrieves paragraph-level context from the source PDF,
2. fuses visual features with retrieved textual evidence, and
3. generates accurate, closed-ended answers.

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 221–229

Our code and implementation details are publicly available at GitHub[1] for reproducibility and further research.

2 Related Work

Multimodal Scientific Figure Understanding. Scientific visual question answering (VQA) and captioning require models to interpret domain-specific plots, charts, and diagrams that differ significantly from natural images. Early datasets such as ACL-Fig introduced a taxonomy for figure types from the ACL Anthology, enabling classification and captioning research on structured scientific visual content (Karishma et al., 2023). SciGraphQA (Shengzhi Li, 2023), a foundational dataset for SciVQA, focused on QA over scientific graphs by pairing structured visual content with underlying textual and symbolic metadata. SciCap+ demonstrated that incorporating contextual mention-paragraphs improves caption quality for scientific figures (Yang et al., 2023), while Multimodal ArXiv showed that domain-specific fine-tuning on scientific plots closes the generalization gap of large vision-language models (LVLMs) (Li et al., 2024). SPIQA introduced one of the first QA benchmarks over interleaved figures and texts from scientific papers, emphasizing the importance of cross-modal reasoning in retrieval-based QA systems (Pramanick et al., 2024).

Retrieval-Augmented Generation in QA. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) combines dense retrieval with sequence-to-sequence generation to improve factual correctness and grounding in QA tasks. While originally introduced for open-domain QA, subsequent works have adapted RAG to handle domain-specific documents, including scientific literature, by embedding long-form PDFs (Rujun Han and Castelli, 2024) and utilizing contrastive retrieval strategies such as Dense Passage Retrieval (DPR) (Karpukhin et al., 2020). Recent multimodal QA studies have integrated RAG with LVLMs to support visual reasoning over complex figures and tables.

Large Vision-Language Models. Early LVLMs like CLIP and ViLT excelled on natural image benchmarks but struggled with abstract scientific diagrams due to limited domain grounding (Radford et al., 2021; Kim et al., 2021). Recent advances, including InterVL3-14B and other 14B+ parameter models, demonstrate better cross-modal understanding through pretraining on multimodal documents and structured figures (Li et al., 2024). However, these models still benefit significantly from RAG pipelines, which inject external domain knowledge and context—especially for nuanced figure-based QA tasks, as explored in our work.

3 Dataset

The SciVQA dataset[2] comprises scientific figures extracted from papers in the ACL Anthology and arXiv, each annotated with question–answer (QA) pairs and associated metadata. The dataset is organized into three splits: a training set with approximately 15k instances, a validation set with 1.7k instances, and a test set containing 4.2k instances. An instance is one datapoint consisting of figure and its respective question answer pair.

Each QA pair in SciVQA is categorized along two key dimensions: answerability and visual grounding. Based on answerability, QA pairs are labeled as either closed-ended (answerable solely from the image or image+caption), unanswerable (not inferable from the given source), finite answer set (with binary or multiple-choice answers), or infinite answer set (requiring open-form answers, such as numerical sums). Based on visual grounding, QA pairs are classified as either visual—requiring interpretation of figure elements like shape, size, position, height, direction, or colour—or non-visual, which do not involve these aspects.

The dataset also provides annotations for figure types, distinguishing between compound figures—those composed of multiple subfigures—and non-compound figures, which depict a single visual element. Figure types span common scientific visualizations such as line charts, bar charts, box plots, confusion matrices, and pie charts. We perform all the evaluation on test set only.

Figure 1 presents a sample QA pair along with its corresponding figure from the test set. In the SciVQA dataset, each figure is accompanied by a caption and is paired with seven distinct QA pairs, each corresponding to a different QA pair type. The example shown illustrates one such QA pair, demonstrating the format of the image, caption, and

[1] https://github.com/joydeb28/ExpertNeurons-SciVQA_2025

[2] https://huggingface.co/datasets/katebor/SciVQA

associated multiple-choice question and answer.

> *"Figure 6: Number of documents with an 'attacking' country per 3-month period, and coreference posterior uncertainty for that quantity. The dark line is the posterior mean, and the shaded region is the 95% posterior credible interval. See appendix for more examples."*

and the associated figure, a representative question is:

> *"Which line represents the quantity of documents with an 'attacking' country for Serbia/Yugoslavia?"*

with answer choices such as: **A.** The blue line, **B.** The red line, **C.** The gray line, **D.** All of the above.

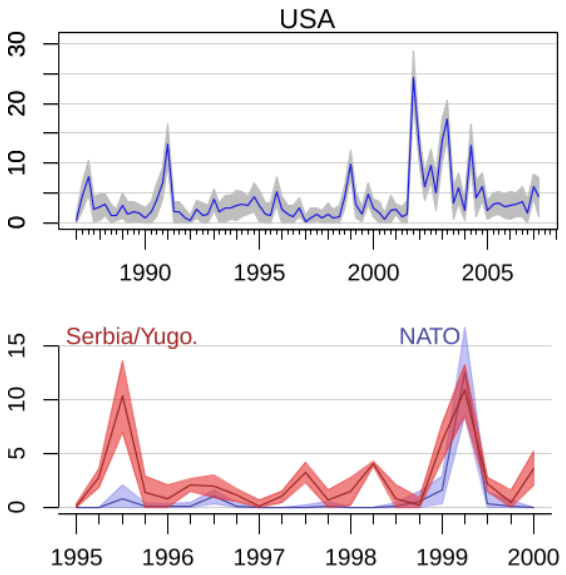

Figure 1: Example scientific figure from the SciVQA dataset showing temporal trends for different countries.

We also do preprocessing of images in the later step as mentioned in Setting C. Captions and questions are tokenized using the BERT tokenizer, with a maximum sequence length of 512 subword tokens. This preprocessing ensures a consistent input structure for our RAG-based architecture. The term subword tokens refers to the output units produced by the BERT tokenizer after applying WordPiece tokenization to the input text.

4 Methodology

The overall flow of our proposed approach is illustrated in Figure 2. We conducted experiments across multiple distinct configurations, each incrementally improving upon the last to evaluate model capabilities comprehensively. Below are details of input meta information across settings.

Setting A: For inference, the prompt includes `image_file`, `caption`, and `question`.

Setting B: During fine-tuning, only `image_file` is used. For inference, the prompt includes `image_file`, `caption`, and `question`.

Setting C: Fine-tuning uses `image_file` and the corresponding `PDF`. Inference is performed using `image_file`, `caption`, and `question`.

Setting D (Final Approach): Identical to Setting C, fine-tuning utilizes `image_file` and `PDF`, and inference uses `image_file`, `caption`, and `question`.

4.1 Setting A: Baseline Evaluation with Image-Only Inputs

In this preliminary evaluation, we assessed several state-of-the-art multimodal models based on Open VLM leaderboard (opencompass) to establish baseline performance on the SciVQA chart image question-answering task without additional training or context. Due to resource constraints for further finetuning, we limited our experiments to models up to 32 billion parameters only. Models evaluated included Pixtral-12B (Agrawal et al., 2024), Mistral-Small-24B (Mistral AI, 2025), InternVL3-14B (OpenGVLab), and Qwen-2.5-VL (Alibaba Group, 2024). The InternVL3-14B model demonstrated notably superior initial performance, as summarized in Table 1. Consequently, InternVL3-14B was selected as the foundational model for all subsequent experimental settings.

4.2 Setting B: Image-Only Finetuning (SciVQA Data)

Building upon our baseline, we finetuned the InternVL3 model using the official SciVQA training dataset. Finetuning employed a Low-Rank Adaptation (LoRA) (Hu et al., 2021) strategy with the following hyperparameters: rank = 64, epochs = 4, and a learning rate of 4×10^{-4}. The purpose was to specialize the model explicitly toward chart-based visual question-answering tasks. The models were finetuned on a single A100 GPU of 80 GB RAM.

4.3 Setting C: Enhanced Contextual Finetuning (Image Sharpening and RAG)

Analysis of results from Setting B via manual verification of 100+ random samples highlighted two prevalent challenges:

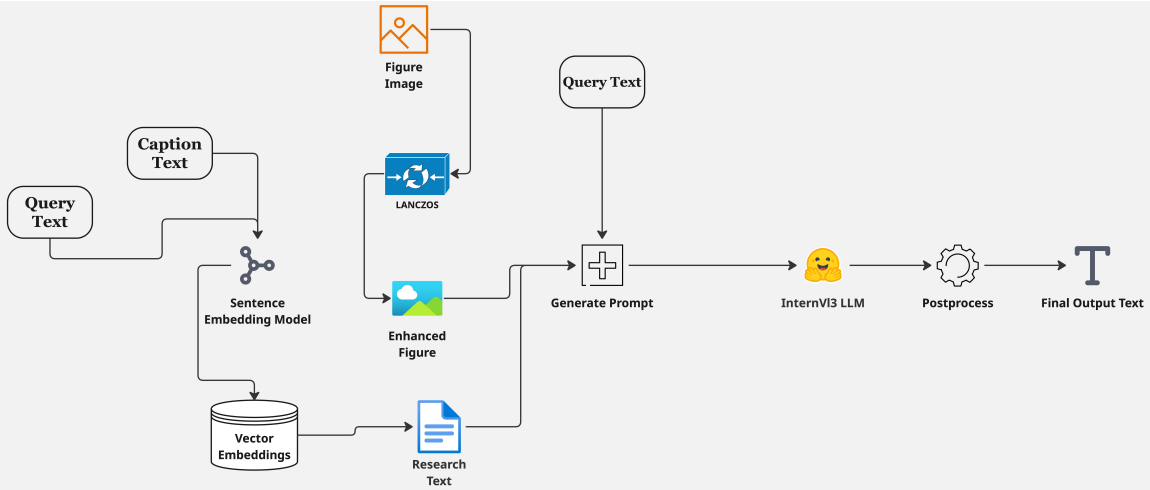

Figure 2: Overall Methodology

1. **Image Quality:** A significant portion of failed cases on the validation set were associated with poor image clarity, which hindered effective visual information extraction. To investigate this, we manually inspected 100 randomly sampled failure cases. Among these, approximately 20% (i.e., 20 samples) were found to exhibit image quality issues. These included low-resolution renders, blurry charts, and faint or unreadable axis labels and legends. The annotations were performed manually by the authors, who marked these images as visually noisy or difficult to interpret for tasks such as reading precise values or identifying attributes like bar height or line trends. Figure 3 in Appendix A shows one such sample image.

2. **Contextual Insufficiency:** Another source of model error stemmed from textual context. In certain cases, relying solely on the figure and its caption failed to provide sufficient cues such as variable definitions, experimental configurations, or axis descriptions needed to fully disambiguate the question. While the dataset formally categorizes most QA pairs as closed-ended (i.e., answerable from the image and caption), we found that in practice, additional context from the surrounding text could enhance answerability. During our manual analysis of 100 failed cases, we noticed around 7% of the samples which could have benefited by additional context provided in the caption or data from the paper. These were also verified by the authors through a qualitative assessment of whether access to more textual context (e.g., caption or the paragraph surrounding the figure in the paper) could plausibly improve performance. While the correct answer may not always be explicitly stated in the surrounding text, this additional context often reinforces key concepts, thereby supporting more accurate answer generation. Sample instances illustrating such contextual gaps are included in the Appendix A referred to in Figure 4 and Figure 5.

   Also note that these annotations were based on a limited, manually inspected subset (n=100) due to resource constraints. While the proportions reported here may not generalize to the entire dataset, our intent is to identify common failure modes rather than provide exact quantitative prevalence.

To address these issues, we adopted two significant improvements:

Image Upscaling and Sharpening: We applied a Lanczos resampling technique (Turkowski and Gabriel, 1990), which is renowned for effectively preserving edge sharpness, to enhance image clarity. Specifically, each image was resized by doubling its original dimensions uniformly to maintain aspect ratios while improving visual fidelity.

Retrieval-Augmented Generation (RAG): To incorporate broader textual context for scientific visual questions, we implemented a retrieval-augmented pipeline that extracts relevant text from the source papers associated with each figure in the SciVQA dataset.

For each figure instance, we first downloaded

the corresponding academic paper in PDF format using the metadata provided (e.g., arXiv or ACL Anthology identifiers). The full text of the PDF was segmented into semantically meaningful blocks—such as section titles, paragraphs, captions, and table/figure references—using PDF parsing tools like PDFMiner[3]. These blocks, separated based on structural whitespace in the document, were treated as retrieval units.

To locate the caption associated with each figure, we applied regular expressions to detect references such as "Figure X" in the parsed text. This enabled us to extract the specific caption block aligned with the figure metadata.

Next, we generated sentence embeddings (Reimers and Gurevych, 2019) for all textual blocks using a pre-trained Sentence-BERT model. An embedding was also computed for the extracted figure caption. To identify the most relevant textual context, we computed cosine similarity between the caption embedding and each block embedding within the same paper. The top two blocks with the highest similarity were selected—typically the caption itself and an adjacent explanatory section (e.g., description of results or methods).

To further enrich context, we also generated an embedding of the input question and used it to retrieve an additional textual block. This block often provided broader or complementary information from the paper, such as experimental setup, variable definitions, or related discussion, which might not be present near the figure.

Thus, each instance is paired with three retrieved text blocks: the caption block and two additional context blocks (one based on caption similarity, one on question similarity). These were concatenated and used as external context alongside the image during fine-tuning. We retained the LoRA-based fine-tuning strategy from Setting B.

4.4 Setting D: Augmented Dataset and Post-processing Refinement

To further enhance model robustness and generalization, we augmented the training data with additional samples from the ChartQA dataset (Masry et al., 2022), which features complex reasoning-based questions spanning diverse chart types. ChartQA was selected due to its structural and semantic alignment with SciVQA, particularly in its inclusion of real-world scientific plots, numeric reasoning, and visual attribute-based questions. From this dataset, we integrated approximately 2,500 samples into our training corpus. These samples were filtered to retain those that met two criteria: (i) the figure type was within the scope of our model (e.g., bar, line, or pie charts), and (ii) the questions were of high quality, which we ensured by selecting only those samples from the ChartQA dataset that were explicitly tagged as human authored. In ChartQA, each QA pair includes metadata indicating whether it was generated by a human or machine based method. We filtered out all machine generated questions and retained only those tagged as human annotated, as these are typically designed to be of higher semantic quality. The dataset filtering was done automatically solely based on the tags provided in the dataset without any manual inspection. The 2,500 sample limit was chosen to maintain a balanced distribution with the original SciVQA samples and to prevent the model from overfitting to the style or domain of a single dataset.

Despite accuracy improvements, we observed that the fine-tuned InternVL model occasionally generates a range of values (e.g., *"between 0.2 and 0.3"*) instead of a single numerical answer, particularly in cases where the model exhibits uncertainty. This behavior appears to stem from the model's tendency to express ambiguity when it is not confident about a precise value. We have included an example of such a case in the Appendix A. While such responses can be semantically reasonable (particularly when axis resolution is low or approximate visual estimation is needed), they pose challenges for automatic evaluation, which often relies on exact matching or scalar closeness to gold answers.

To address inconsistent range-based outputs in direct answer questions, we implemented a lightweight post-processing module using regular expressions and simple heuristics to detect numeric ranges and replace them with their arithmetic mean. This standardization improves alignment with expected ground truth formats and ensures more consistent scoring under numeric evaluation schemes. While this transformation may introduce minor inaccuracies when ranges are semantically justified, it generally enhances answer conformity and evaluation robustness.

This combined approach leveraging data set enhancement and output refinement further improved model precision and interpretability, as shown in Table 1.

[3] https://github.com/pdfminer/pdfminer.six

Setting	Model Settings	R-1 F1	R-L F1	BS F1
A	Pixtral-12B	0.6480	0.6480	0.9680
	Mistral-Small-24B	0.6787	0.6782	0.9742
	Qwen-2.5-VL	0.6780	0.6780	0.9610
	InternVL3-14B	0.7130	0.7130	0.9750
B	InternVL3-14B + Finetuning	0.7753	0.7750	0.9804
C	InternVL3-14B + Finetune + RAG	0.7986	0.7983	0.9846
D	InternVL3-14B + Finetune + RAG + Augmentation and Post Refinement	0.8049	0.8043	0.9849

Table 1: Evaluation metrics across multiple settings. Each row shows results using progressively advanced configurations for vision-language QA. Setting A includes baseline models; B-D represent stages of fine-tuning, RAG integration, and data augmentation. R-1: ROUGE-1, R-L: ROUGE-L, BS: BERTScore

Table 2 presents the leaderboard results for the SciVQA 2025 shared task. Our system, ExpertNeurons, achieved the highest performance across all evaluation metrics, demonstrating the effectiveness of our RAG-VLM architecture.

#	Team	R-1 F1	R-L F1	BS F1
1	ExpertNeurons	0.8049	0.8043	0.9849
2	THAii_LAB	0.7899	0.7892	0.9839
3	Coling_UniA	0.7862	0.7856	0.9817
4	florian	0.7631	0.7621	0.9831
5	Infyn	0.7350	0.7345	0.9787

Table 2: Leaderboard on SciVQA 2025 test set. R-1: ROUGE-1, R-L: ROUGE-L, BS: BERTScore. Baseline not ranked.

5 Discussion

Table 1 summarizes the performance of F1 for ROUGE-1, ROUGE-L and BERTScore in the four setting of methodology (A to D) on test set. Each stage demonstrates incremental improvements with better contextual modeling and data augmentation. Setting D secured the top rank in the leaderboard with 0.8049 (ROUGE-1 F1-score), 0.8043 (ROUGE-L F1-score), and 0.9849 (BERTScore F1-score).

Our experiments highlight several key insights into the performance and limitations of retrieval-augmented VQA systems in scientific domains.

Baseline models evaluated under Setting A (A1–A4) demonstrated limited ability to handle scientific chart-based questions. Among them, InternVL3-14B (A4) performed the best with ROUGE-1 and ROUGE-L scores of 0.7130, and a BERTScore F1 of 0.9750, indicating that even strong vision-language models struggle without task-specific adaptation. This highlights the inherent complexity of scientific figures, which often lack standalone semantics and require specialized training or contextual information.

With fine-tuning on the SciVQA dataset (Setting B), InternVL3-14B achieved a substantial performance boost—ROUGE-1 improved from 0.7130 to 0.7753 (+6.23%), and BERTScore rose to 0.9804. However, we observed a plateau on questions demanding deeper reasoning beyond surface-level visual cues, underscoring the need for additional context.

Setting C addressed these limitations by integrating high-resolution image sharpening and contextual grounding via our RAG pipeline. This led to a further increase in ROUGE-1 to 0.7986 and BERTScore to 0.9846, suggesting enhanced capacity for visual-textual reasoning through targeted retrieval from source PDFs.

Finally, Setting D yielded the highest performance: ROUGE-1 reached 0.8049, and BERTScore climbed to 0.9849. The 0.63% gain in ROUGE-1 and marginal BERTScore improvement over Setting C reflect the complementary benefits of including 2,500 reasoning-centric samples from ChartQA and the application of post-processing techniques to resolve answer ambiguity

6 Limitations

Although our approach demonstrates promising results, it still has several limitations stemming from two primary factors.

Firstly, certain challenges arise from the data itself. These include poor image quality, lack of contextual information, or missing visual elements. Additionally, in some instances, the correct answer is visually ambiguous or difficult to distinguish from the figure for example, differentiating between values such as 0.54 and 0.56 in a bar graph.

Secondly, while our method incorporates additional contextual information to support answer prediction, this context is not always sufficient or

fully relevant. Although the inclusion of retrieved context generally improves performance, there are edge cases where questions originally labeled as unanswerable could become answerable with the added context potentially leading to inconsistencies in evaluation and lower performance. A systematic analysis of these cases is currently lacking and would require additional strategies to robustly identify and handle such cases. Furthermore, although the auxiliary dataset used to enhance model performance contributes positively, it does not comprehensively capture the full complexity and diversity of question types presented in the shared task.

7 Conclusion

We present a Retrieval-Augmented VQA pipeline that combines vision-language modeling with document-aware context retrieval to improve scientific chart understanding. Through progressive experimentation and enhancement, our method achieved significant gains in accuracy, reasoning depth, and answer quality.

By integrating image sharpening, textual retrieval, and dataset augmentation, the system successfully bridges the gap between purely visual inputs and the rich semantic context needed for effective scientific QA. Our approach demonstrates the potential of LLM enhanced vision-language systems in handling complex academic visual data.

Future work will explore multi-modal attention mechanisms across figure-caption-text triplets and generalize the framework to broader scientific domains, enabling more diverse and open-ended question answering capabilities.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Alibaba Group. 2024. Qwen2.5-VL. https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct. Accessed: 2025-06-18.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. 2025. SciVQA 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria. Accepted.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *NeurIPS*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zeba Karishma, Shaurya Rohatgi, Kavya Puranik, Jian Wu, and C. Lee Giles. 2023. Acl-fig: A dataset for scientific figure classification. *arXiv preprint arXiv:2301.12293*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Wen-tau Yih, Tim Rockt"aschel, and Sebastian Riedel. 2020. Dense passage retrieval for open-domain question answering. *Proceedings of EMNLP*.

Wonjae Kim, Bokyung Cho, Sungdong Yoo, Jaewoo Choi, Jinyeong Kim, Taeuk Lee, Jaewook Kang, and Jaewhan Choi. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *ICML*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Vladimir Karpukhin, Naman Goyal, Abdelrahman Mohamed, Tim Rockt"aschel, and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*.

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

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

Mistral AI. 2025. Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1. Accessed: 2025-05-23.

opencompass. Open-vlm-leaderboard. https://huggingface.co/spaces/opencompass/open_vlm_leaderboard. Accessed: 2025-06-18.

OpenGVLab. InternVL3-14B. https://huggingface.co/OpenGVLab/InternVL3-14B. Accessed: 2025-06-18.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *arXiv preprint arXiv:2407.09413*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *ICML*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Peng Qi Yumo Xu Jenyuan Wang Lan Liu William Yang Wang Bonan Min Rujun Han, Yuhao Zhang and Vittorio Castelli. 2024. Rag-qa arena: Evaluating domain robustness for long-form retrieval augmented question answering. *EMNLP*.

Nima Tajbakhsh Shengzhi Li. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*.

Ken Turkowski and Steve Gabriel. 1990. Filters for common resampling tasks. In Andrew Glassner, editor, *Graphics Gems I*, pages 147–165. Academic Press. Includes descriptions of Lanczos interpolation.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Error Analysis Examples

All the samples shown below are from validation set.

Case 1: Low Quality Image

Example 1: Figure 3 shows an image with low visual resolution. Such figures may hinder model

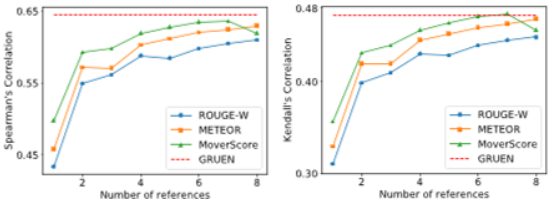

Figure 3: image_file : 2010.02498v1-Figure3-1.png

comprehension of fine-grained visual details, including axis labels and line plots, impacting the model's accuracy in visual question answering.

Case 2: Need for Additional Context Beyond Caption

Example 1: Figure 4 depicts a Q-network structure. The question requires reasoning beyond the figure and its caption. Without context, the model misinterprets the output of the Q-network.

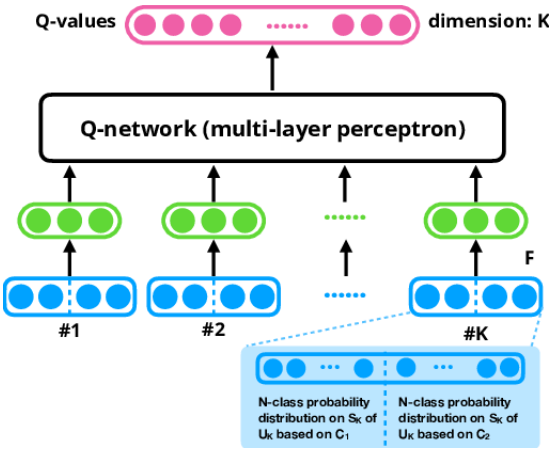

Figure 4: image_file : N18-1113.pdf-Figure3.png

Instance ID: `07a642e0d2e24761496b7e0a3b41d5fd`

Question: *Is 'Q-keys' the output of 'Q-network'?*

Caption: *Figure 3: The structure of Q-network. It chooses a unlabeled subset from U1, U2, ..., UK at each time step. The state representation is computed according to the two classifiers N -class probability distribution on the representative example Si of each subset Ui.*

Context Extracted from PDF: *The Q-value Q(st,a) is determined by a neural network as illustrated in Figure 3.*

Gold Answer: No

Model prediction without Context: Yes

Model prediction with Context: No

Explanation: In this example, the model initially struggled to produce the correct answer when relying solely on the image and its caption. The

term "Q-keys" does not appear in the figure or caption, making it difficult to verify whether it is part of the Q-network's output. However, upon incorporating the additional textual context which explicitly states that "the Q-value Q(st, a) is determined by a neural network" the model is able to correctly infer that the Q-network's output is the Q-value, not "Q-keys". This additional information provides supporting clarification and helps leads to the correct answer.

Example 2: In Figure 5, the question requires semantic inference of correlation between social score and Airbnb penetration. The additional context helps the model to better comprehend this information.

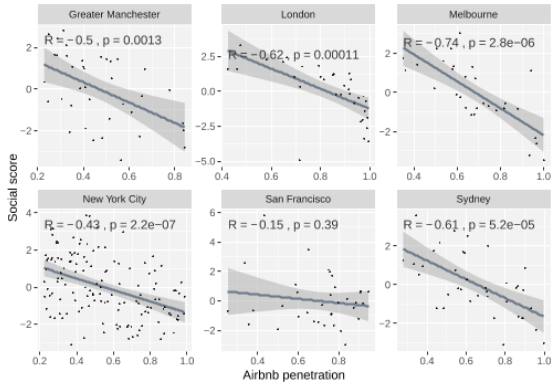

Figure 5: image_file : 2004.11604v1-Figure8-1.png

Instance ID: `db801444b0b421e86bc07199fa465997`

Question: *Is the social score negatively correlated with Airbnb penetration rate in every city?*

Caption: *Fig. 8: Social score against area Airbnb penetration rate (on a per city basis)*

Context Extracted from PDF: *Figure 8 shows the scatter plot (along with Pearson Correlation) between the Airbnb penetration rate and the social score for neighbourhoods in each city in our dataset. We observe that neighbourhoods with very high Airbnb adoption rates show lower social scores than those with lower penetration rates (Pearson correlation up to -0.74). Results are valid across all cities considered*

Gold Answer: Yes

Model prediction without Context: No

Model prediction with Context: Yes

Explanation: In this case, the model failed to produce the correct answer when limited to just the image and caption. The additional context, however, clearly asserts that the results are "valid across all cities considered" and quantifies the negative correlation (Pearson correlation up to -0.74). This reinforces the claim that high Airbnb penetration consistently corresponds to lower social scores in every city analyzed. With this information, the model is able to identify the presence of a negative correlation across all cities, showing that additional textual context can help the model in answering complex questions.

Case 3: Sample highlighting postprocessing module refinement

Example 1

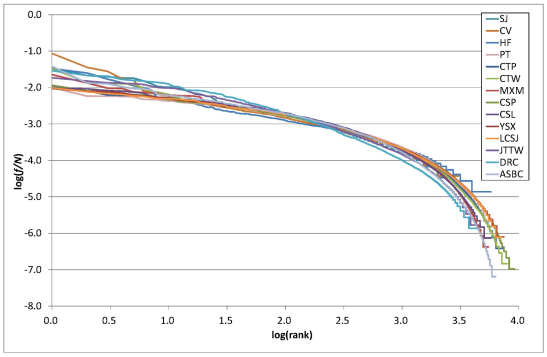

Figure 6: image_file : 1709.05587v1-Figure1-1.png

Instance ID: `6b81a93e1cce9b999b05564beda9ba52`

Question: *What is the approximate value of log(f/N) for the blue line labeled 'DCR' at a log(rank) value of 3.5?*

reference figure: Figure 6

Gold Answer: -5

Model prediction before postprocessing step: between -4 and -6

Model prediction after postprocessing step: -5

Explanation: In this case, the model exhibited uncertainty regarding the exact answer and returned a range as output. Our heuristic based post processing module identified this pattern and replaced the range with a single scalar value, computing the mean of -4 and -6 to produce -5. The rationale behind this step is to standardize outputs and thereby improve the reliability and consistency of the evaluation process, which might benefit from precise answers for comparison against ground truth.