Visual Question Answering on Scientific Charts Using Fine-Tuned Vision-Language Models

Florian Schleid, Jan Strich, Chris Biemann

Language Technology Group, Universität Hamburg, Germany
florian.schleid@uni-hamburg.de

Abstract

Scientific charts often encapsulate the core findings of research papers, making the ability to answer questions about these charts highly valuable. This paper explores recent advancements in scientific chart visual question answering (VQA) enabled by large Vision Language Models (VLMs) and newly curated datasets. As part of the SciVQA shared task from the 5th Workshop on Scholarly Document Processing, we develop and evaluate multimodal systems capable of answering diverse question types - including multiple-choice, yes/no, unanswerable, and infinite answer set questions - based on chart images extracted from scientific literature. We investigate the effects of zero-shot and one-shot prompting, as well as supervised fine-tuning (SFT), on the performance of Qwen2.5-VL models (7B and 32B variants). We also tried to include more training data from domain-specific datasets (SpiQA and ArXivQA). Our fine-tuned Qwen2.5-VL 32B model achieves a substantial improvement over the GPT-4o mini baseline and reaches the 4th place in the shared task, highlighting the effectiveness of domain-specific fine-tuning. We published the code for the experiments[1].

1 Introduction

Figures are often the first thing that readers of scientific papers look at (Rolandi et al., 2011). Also, they frequently communicate the main results. Therefore, the ability to extract information from scientific chart images would be of great value. However, automatically interpreting charts poses challenges due to their detailed visual components and the complex spatial arrangements of elements. The process requires spatial reasoning and numerical understanding (Meng et al., 2024). New SOTA VLMs like Qwen2.5-VL (Bai et al., 2025) enable better results in the domain of chart VQA (Masry et al., 2025). Furthermore, recent datasets, like

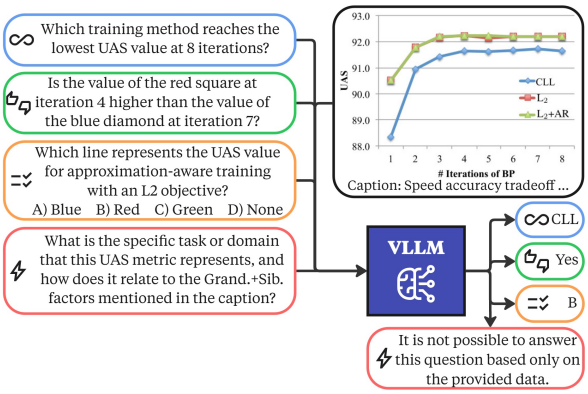


Figure 1: Overview of the system with the four question types: infinite answer set, yes/no, multiple-choice, and unanswerable.

SpiQA (Pramanick et al., 2024) and ArXivQA (Li et al., 2024), provide large amounts of data on scientific chart VQA. This paper intends to explore these new possibilities in the context of the SciVQA shared task (Borisova et al., 2025). It challenges participants to answer questions about scientific charts. An example of such questions can be seen in Figure 1.
The contributions of this paper are:

- **Fine-tuning Qwen2.5-VL models (Bai et al., 2025) for chart VQA:** The model size and the hyperparameters used for the fine-tuning have a strong impact on the results. This paper explores different configurations.
- **Testing prompt templates and one-shot prompting:** Prompt engineering is important to get the desired output format and can improve the results.
- **Exploring other datasets:** We investigate the influence of adding training data from similar-domain datasets.

[1] https://github.com/Flo0620/Scientific-Chart-QA

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 211–220

2 Related Work

The SciVQA shared task invites participants to develop multimodal systems for VQA on scientific charts (Borisova et al., 2025). To support this, we use the SciVQA dataset[2], which contains 3,000 real-world chart images from scientific papers, each paired with seven questions. The dataset features four question types: multiple-choice, yes/no, unanswerable, and infinite answer set questions. These are further categorized into visual and non-visual questions, where visual questions refer to attributes such as size, height, color, direction, shape, or position. This task aligns with growing research interest in chart-based VQA, where existing benchmarks such as ChartQA have seen performance plateaus among large VLMs, largely due to limited data diversity (Masry et al., 2025). In response, new benchmarks such as SpiQA (Pramanick et al., 2024) and ArXivQA (Li et al., 2024) have been introduced to address this limitation by using more diverse scientific charts from the real world.

These new datasets provide additional training data to fine-tune VLMs. Li and Tajbakhsh (2023) found a positive correlation between the size of the training set and the model performance when fine-tuning a VLM for chart VQA. Furthermore, Wu et al. (2024) showed that the used prompt has a significant influence on the results of the task of VQA on charts, underlining the importance of prompt engineering.

Recent progress in the field of VLMs includes models such as Qwen2.5-VL (Bai et al., 2025), a successor to Qwen2-VL that has achieved SOTA results in chart VQA tasks (Li et al., 2025; Masry et al., 2025). There are also models specifically developed for chart-related tasks, such as ChartLlama, which performed fine-tuning on a curated dataset and reached good results on the ChartQA benchmark (Han et al., 2023). ChartAssistant (Meng et al., 2024) and ChartVLM (Xia et al., 2024) fine-tuned models to perform chart-to-table translation. The output of these models is then used as input to specialized models fine-tuned for VQA.

The fine-tuning of such models is possible through Low Rank Adaptation (LoRA) (Hu et al., 2022), which drastically reduces the memory and computation requirements for the training.

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

3 Experiments

To explore the influence of the prompting strategy, the effectiveness of domain-specific fine-tuning, and the importance of the training dataset size, we conducted four different experiments to tackle the task. Firstly, we tried zero-shot inference in combination with prompt engineering. Secondly, we performed one-shot prompting with one example. Third, the VLMs were fine-tuned on the dataset provided by the task[2]. Lastly, we expanded the fine-tuning by including the SpiQA (Pramanick et al., 2024) and ArXivQA (Li et al., 2024) datasets in the training data. All four approaches were tested on the 7B and 32B variants of the Qwen2.5-VL model (Bai et al., 2025) and the first two on GPT-4o mini[3] (OpenAI et al., 2024).

3.1 Zero-Shot

In the zero-shot prompt, the model is given clear instructions on how to respond to different types of questions, reducing hallucinations by providing a desired output if the question cannot be answered from the given information. Furthermore, it is provided with the caption of the chart as an additional information source. The prompt templates are given in the Appendix A.1.

3.2 One-Shot

The user prompt is expanded with an example. If the target question is a multiple-choice question, we align the example to the target by using a multiple-choice question as an example. Otherwise, an infinite answer set question is used. The complete one-shot prompt can be seen in the Appendix in Figure 4. The multiple-choice example question and the infinite answer set example question were selected from the training split of the SciVQA dataset to have one visual and one non-visual question.

3.3 LoRA Fine-Tuning

The SFT of the 7B and 32B variants of the Qwen2.5-VL model (Bai et al., 2025) was performed using LoRA (Hu et al., 2022) with the zero-shot prompt template (see Appendix A.1) on the SciVQA training data[2] (15K questions). The base models are loaded in 8-bit, the learning rate was set to 2×10^{-4} with a linear learning rate scheduler. Hyperparameter tuning determined the

[3]https://openai.com/index/GPT-4o-mini-advancing-cost-efficient-intelligence/

Models	Zero-Shot	One-Shot	LoRA Fine-tuning	Fine-tuning + other datasets
Qwen 7B	**0.5968**	0.5972	0.8128	0.7989
Qwen 32B	0.5188	0.5243	**0.8361**	**0.8176**
GPT-4o mini	0.5424	**0.6326**	-	-

Table 1: Performance comparison of the Qwen 7B, Qwen 32B, and GPT-4o mini models on the SciVQA test set (4,200 questions) across different learning paradigms: zero-shot, one-shot, LoRA fine-tuning, and fine-tuning with additional datasets. Reported values are the average of the F1-scores of ROUGE-1, ROUGE-L, and BERTScore. The best score in each setting is highlighted in bold.

LoRA parameters rank $= 64$, alpha $= 128$, and dropout $= 0.2$, since they led to the best average of the ROUGE-1, ROUGE-L, and BERTScore F1-scores (see Table 4 in Appendix). This average also determines the ranking in the competition. Fine-tuning the Qwen2.5-VL 32B model for four epochs with the described parameters led to the best results after two epochs (see Table 5 in Appendix). Therefore, the 7B and 32B models used in the evaluation were fine-tuned for two epochs. The GPUs used for the fine-tuning were one NVIDIA RTX A6000 (48GB VRAM) for the 7B model and one NVIDIA A100 (80GB VRAM, PCIe) for the 32B model.

3.4 LoRA Fine-tuning with other Datasets

To explore the effects of using more domain-specific data for fine-tuning, we sought datasets similar to SciVQA[2]. We therefore incorporated the SpiQA (Pramanick et al., 2024) and ArXivQA (Li et al., 2024) datasets as additional data sources, as they also use real-world scientific charts and primarily contain infinite answer set and multiple-choice questions, respectively. To avoid overlap, any questions from papers which were also scraped in the SciVQA dataset were excluded.

To align the filtered SpiQA questions closer to the SciVQA questions, only questions with answers that have at most 50 characters were retained, leaving 39K mostly infinite answer set questions. The resulting average answer length in the filtered SpiQA questions is with 15.3 characters, relatively close to the average answer length of 14.4 characters of the SciVQA train dataset.

From ArXivQA, only multiple-choice questions with 4 options were kept, as the multiple-choice questions in the SciVQA dataset also have 4 options. This yielded 61K questions. Images from both datasets were resized to a maximum of 500K pixels while preserving the aspect ratio. These filtered datasets, along with the SciVQA train dataset, were combined to 115K questions and used to fine-tune both the 7B and 32B Qwen2.5-VL models for one epoch each. Due to time constraints, no hyperparameter tuning could be performed for the training with this combined dataset. Therefore, except for the described changes in the data and the number of epochs, the other training parameters, as well as the GPUs used, were the same as described in Subsection 3.3.

4 Evaluation

This section presents the main experimental results and provides a detailed comparison between our models and the GPT-4o mini model, including a manual error analysis in the ablation study.

4.1 Main Results

The experiments were evaluated on the test split of the SciVQA dataset, which contains 4200 questions. As the main evaluation metric, the average of the F1-scores of ROUGE-1, ROUGE-L, and BERTScore was used. The results of the experiments are presented in Table 1.

For the zero-shot experiment, the fact that the 7B Qwen model received a significantly better score (0.597) than the 32B model (0.519) and GPT-4o mini (0.543) was unexpected. However, taking a closer look at the provided answers revealed that the answers given by the 32B model had an average length of 351.8 characters, while the 7B model had an average answer length of 57.3 characters, and the GPT-4o mini model of 64.9 characters. Though the test set answers are not public, the validation set has an average answer length of 14.4 characters. Since most of the ground-truth answers only contain a few words and often only one word, the precision of the ROUGE-1 and ROUGE-L metrics reduces for long answers, and the recall is capped at one. This explains why the F1-scores and therefore their average for the 32B model are poor.

Providing the model with a one-shot example works best on the GPT-4o mini model. It reached a

Model	Infinite		Yes/No		Multiple-Choice		Unans.	Overall
	v	n-v	v	n-v	v	n-v		
GPT-4o mini 0-shot	0.4500	0.6833	0.6625	0.7125	0.5042	0.5167	0.6250	**0.5935**
GPT-4o mini 1-shot	0.3875	0.6500	0.6458	0.7000	0.5042	0.5708	0.7708	**0.6042**
Ours	0.5458	0.7500	0.8000	0.8167	0.7583	0.6958	0.9792	**0.7637**

Table 2: Manual evaluation of results obtained for the Qwen2.5-VL 32B model, fine-tuned for two epochs (Ours), and GPT-4o mini. The table shows the fraction of correctly answered questions on the SciVQA validation dataset (1680 questions) per question type. Each question type contains 240 questions. The fine-tuning on the Qwen model was performed on the training split of the SciVQA dataset (15K questions). 'v' and 'n-v' indicate if the questions are visual or non-visual.

Model	Infinite		Yes/No		Multiple-Choice		Unans.	Overall
	v	n-v	v	n-v	v	n-v		
Combined	0.6342	0.7443	0.7971	0.8388	0.7694	0.7718	0.9546	**0.7872**
SciVQA	0.6878	0.7877	0.8527	0.8611	0.8227	0.7821	0.9669	**0.8230**

Table 3: Comparison of the average F1-scores of the ROUGE-1, ROUGE-L, and BERTScore metrics by question type between the Qwen2.5-VL 7B model that was fine-tuned on the combined dataset and the 7B model exclusively fine-tuned on the SciVQA dataset. 'v' and 'n-v' indicate if the questions are visual or non-visual. The evaluation was done on the SciVQA validation dataset (1680 questions). Each question type contains 240 questions.

score of 0.633, outperforming both the 7B and 32B Qwen models and improving greatly compared to the GPT-4o mini model with the zero-shot prompt. Surprisingly, adding a one-shot example does not lead to great improvements for the 7B and 32B models as compared to the zero-shot setting. An analysis of the 7B model's responses revealed that it marked over 2,200 out of 4,200 questions as unanswerable, despite only 600 questions being unanswerable. For the 32B model, the answers even got longer, with an average answer length of 433.3 characters. The average answer length of the GPT-4o mini model reduced to 40.6 characters. These results show that the GPT-4o mini model can leverage one-shot examples much better than the Qwen2.5-VL models and that one-shot prompting can be suitable for doing VQA on charts.

Fine-tuning led to the best result we could achieve across our experiments, with a score of 0.836 for the Qwen2.5-VL 32B model. The expected superiority of the 32B model is also evident here. It outperformed the fine-tuned 7B variant by 0.023, and the GPT-4o mini models, that did not receive fine-tuning, by 0.206. This shows the great potential of domain-specific fine-tuning.

Adding more training data from the SpiQA (Pramanick et al., 2024) and ArXivQA (Li et al., 2024) datasets, as described in Subsection 3.4 resulted in a score of 0.818 for the Qwen2.5-VL 32B model. It therefore reduced the performance in comparison to the model fine-tuned only on the SciVQA dataset. The reason for that is not clear, and further studies are needed to explain the performance drop. A starting point could be to perform dedicated hyperparameter tuning for the combined dataset, since the substantially larger number of training samples could require the hyperparameters to be adjusted. Also, the fact that especially ArXivQA covers a wider range of scientific fields than the SciVQA dataset[2] (Li et al., 2024; Li and Tajbakhsh, 2023; Karishma et al., 2023) should be further investigated as a possible problem source.

4.2 Ablation Studies

Although the F1-scores of ROUGE-1, ROUGE-L and BERTScore provide a useful estimate of the result quality, accurate evaluation, where the answer length does not influence the results, requires a more detailed analysis. Therefore, a manual error analysis was conducted on the SciVQA validation dataset (1680 questions) for the Qwen2.5-VL 32B model fine-tuned for two epochs on the SciVQA dataset, and the GPT-4o mini model using zero- and one-shot prompting. Each answer was manually checked by one annotator to determine whether it accurately answers the given question. The formu-

lation was not taken into account. Table 2 shows the fractions of correctly answered questions.

These results show that our fine-tuned model outperformed the GPT-4o mini model by ~16%, demonstrating the effectiveness of domain-specific fine-tuning. A significant improvement was observed for multiple-choice and unanswerable questions, suggesting that fine-tuning may have reduced hallucinations and helped the model to estimate when not to answer. Additionally, there was a marked difference in performance between visual and non-visual infinite answer set questions for both models. Referencing visual elements appears to be considerably more challenging in the infinite answer set context. Interestingly, this difficulty was not as pronounced in other question types. Surprisingly, providing a one-shot prompt with a visual infinite answer set question led to even worse results for that question type.

To further investigate the poorer results of the fine-tuning on the combined dataset, we compare them with the scores of the 7B model fine-tuned solely on SciVQA in Table 3. Using only SciVQA as training data led to a better score across all question types. Even on multiple-choice questions the results of the model trained on the combined dataset are notably worse than those of the model fine-tuned exclusively on SciVQA. This is unexpected since more than half of the samples in the combined dataset were multiple-choice questions.

5 Conclusion

This paper explored the application of VLMs for scientific chart VQA in the context of the SciVQA shared task. We evaluated zero-shot and one-shot prompting alongside domain-specific fine-tuning using LoRA on Qwen2.5-VL models. Our experiments showed that fine-tuning, especially on the SciVQA dataset alone, led to the most significant performance gains, outperforming the GPT-4o mini baseline and reducing hallucinations. In contrast, incorporating external datasets offered limited benefits, possibly due to suboptimal training conditions or data mismatch. Overall, the results emphasize the value of targeted fine-tuning and careful dataset curation for improving VQA on scientific charts. Future work could include a more sophisticated, possibly manual, selection of training data to further improve the fine-tuning. Testing different hyperparameters for the fine-tuning with more data might also improve the results.

6 Limitations

A key limitation of the system lies in its performance on visual questions with an infinite answer set. For such questions the manual evaluation showed that the model frequently fails to return the exact value of a target datapoint, often producing approximate answers that are close, but fall outside the acceptable error margin to be considered correct. Moreover, the observation that fine-tuning with additional datasets from closely related domains led to a decline in performance suggests limited generalization capabilities. Despite the apparent similarity between the datasets, subtle domain shifts such as differences in the underlying research area or question phrasing, may hinder the model's ability to transfer learned concepts effectively. Potentially, the larger training dataset might also require more trainable parameters than our fine-tuned models had. This highlights potential challenges in developing robust, generalizable models for scientific chart understanding across diverse real-world sources.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 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.

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 preprint*. ArXiv:2311.16483.

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

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. 2023. Acl-fig: A Dataset for Scientific Figure Classification. In *Proceedings of the Workshop on Scientific Document Understanding co-located with 37th AAAI Conference on Artificial Inteligence (AAAI 2023)*, volume

3656 of *CEUR Workshop Proceedings*, Washington DC, USA. AAAI.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Shengzhi Li and Nima Tajbakhsh. 2023. SciGraphQA: A Large-Scale Synthetic Multi-Turn Question-Answering Dataset for Scientific Graphs. *arXiv preprint*. ArXiv:2308.03349.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. *arXiv preprint*. ArXiv:2501.02189.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025. ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering. *arXiv preprint*. ArXiv:2504.05506.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. ChartAssistant: A Universal Chart Multimodal Language Model via Chart-to-Table Pre-training and Multitask Instruction Tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7775–7803, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024*. NeurIPS Vancouver, BC, Canada, December 10 - 15, 2024.

Marco Rolandi, Karen Cheng, and Sarah Pérez-Kriz. 2011. A Brief Guide to Designing Effective Figures for the Scientific Paper. *Advanced Materials volume 23*, pages 4343 – 4346.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. ChartInsights: Evaluating Multimodal Large Language Models for Low-Level Chart Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12174–12200, Miami, Florida, USA. Association for Computational Linguistics.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2024. ChartX & ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning. *arXiv preprint*. ArXiv:2402.12185.

A Appendix

A.1 Prompt Templates

This section contains the final system prompt in Figure 2 and the user prompt used in the zero-shot experiment and to fine-tune the models in Figure 3. Figure 4 shows the user prompt with an example used for the one-shot experiment.

A.2 Hyperparameter Tuning

The correct hyperparameters are essential for good results of the fine-tuned model. This especially applies to the LoRA parameters rank and alpha as well as the dropout. Multiple combinations were tested by fine-tuning the Qwen2.5-7B model on the train split of the SciVQA dataset with 15K questions and evaluating the fine-tuned models on the validation split with 1680 questions. The learning rate used in the experiments was 2×10^{-4} together with a linear learning rate scheduler. Based on the results visible in Table 4 we used a rank of 64, an alpha of 128, and a dropout of 0.2 since it led to the best average of F1-scores, which determines the ranking in the SciVQA competition.

Another important hyperparameter is the number of training epochs. The Qwen2.5-VL 32B model was trained for 1 to 4 epochs on the training data of the SciVQA dataset with LoRA rank $= 64$, alpha $= 128$, and dropout $= 0.2$. To evaluate the training runs the validation split of the SciVQA dataset was again used. As visible in Table 5 the model performs best across all metrics after two training epochs. Therefore, we used two training epochs for the training on the SciVQA data (see Subsection 3.3).

System Prompt

You are an expert data analyst. You will be given an image of a chart and a question.
You will answer the question based on the image of the chart.
If you are sure that you do not have enough information to answer the question answer with: 'It is not possible to answer this question based only on the provided data.'

Figure 2: System Prompt used for fine-tuning the Qwen2.5-VL models, as well as for the zero-shot and one-shot inference.

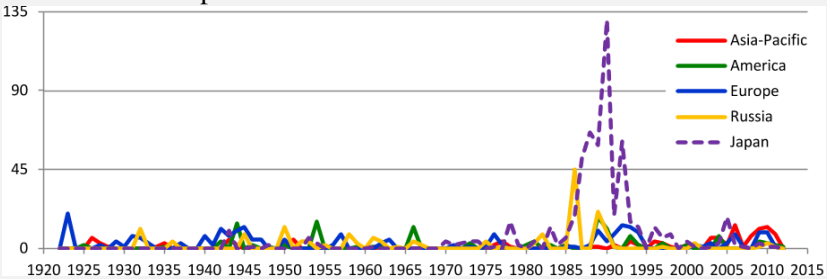
User Prompt

Here is the caption of the image:
{{ caption }}
This is the Question:
{{ question }}
{% if answer_options %}
You have the following answer options to choose from. Multiple answers may be correct. List only the letter of the correct answers in the order they are given without spaces between them.
Answer Options:
{{answer_options}}
{%endif%}
Give a short and precise answer:

Figure 3: User Prompt for fine-tuning the Qwen2.5-VL models and for the zero-shot inference.

One Shot User Prompt

Here is an example:
The caption of the image is:
Figure 3. Annual frequency of USA being mentioned with Russia, Japan, and G20 countries
This is the Question:
{% if answer_options %}
The line of which color had highest annual mention frequency before 1925?
You have the following answer options to choose from. Multiple answers may be correct. List only the letters of the correct answers in the order they are given without spaces between them.
Answer Options:
A: Red line
B: Green line
C: Blue line
D: Yellow line
{% else %}
Which country, besides the USA, is mentioned the most frequently in the year 1990?
{%endif%}
Give a short and precise answer:



{% if answer_options %}
Answer: C
{% else %}
Answer: Japan
{%endif%}

This is the real query you should answer:
Here is the caption of the image:
{{ caption }}
This is the Question:
{{ question }}
{% if answer_options %}
You have the following answer options to choose from. Multiple answers may be correct. List only the letter of the correct answers in the order they are given without spaces between them.
Answer Options:
{{answer_options}}
{%endif%}
Give a short and precise answer:
<image>

Figure 4: User Prompt with one-shot example for the one-shot inference.

			Train Epochs: 1				Train Epochs: 2			
r	α	**d**	**BERT**	**R-1**	**R-L**	**Avg.**	**BERT**	**R-1**	**R-L**	**Avg.**
16	16	0.1	0.9814	0.7250	0.7242	0.8102	0.9810	0.7230	0.7219	0.8087
16	16	0.2	0.9810	0.7278	0.7268	0.8119	0.9806	0.7270	0.7262	0.8113
16	32	0.1	0.9815	0.7259	0.7250	0.8108	0.9825	0.7323	0.7313	0.8154
16	32	0.2	0.9811	0.7188	0.7178	0.8059	0.9817	0.7190	0.7183	0.8063
32	32	0.1	0.9811	0.7288	0.7281	0.8127	0.9810	0.7330	0.7323	0.8155
32	32	0.2	0.9821	0.7271	0.7267	0.8120	0.9822	0.7279	0.7271	0.8124
32	64	0.1	0.9816	0.7381	0.7374	0.8190	0.9819	0.7355	0.7346	0.8173
32	64	0.2	0.9808	0.7301	0.7288	0.8133	0.9810	0.7347	0.7337	0.8165
64	64	0.1	0.9817	0.7318	0.7310	0.8149	0.9828	0.7435	0.7425	0.8230
64	64	0.2	0.9823	0.7400	0.7391	0.8205	0.9819	0.7427	0.7420	0.8228
64	128	0.1	0.9805	0.7315	0.7307	0.8156	0.9811	0.7316	0.7304	0.8144
64	128	0.2	0.9826	0.7388	0.7378	0.8197	0.9822	0.7452	0.7437	0.8237
128	128	0.1	0.9823	0.7401	0.7392	0.8205	0.9827	0.7434	0.7425	0.8228
128	128	0.2	0.9821	0.7376	0.7367	0.8188	0.9819	0.7437	0.7427	0.8228
128	256	0.1	0.9803	0.7270	0.7260	0.8111	0.9815	0.7404	0.7395	0.8205
128	256	0.2	0.9821	0.7329	0.7318	0.8156	0.9823	0.7361	0.7352	0.8179
256	256	0.1	0.9799	0.7292	0.7279	0.8123	0.9808	0.7332	0.7320	0.8153
256	256	0.2	0.9804	0.7302	0.7291	0.8133	0.9808	0.7334	0.7320	0.8154
256	512	0.1	0.9809	0.7263	0.7251	0.8108	0.9813	0.7319	0.7306	0.8146
256	512	0.2	0.9825	0.7249	0.7236	0.8103	0.9818	0.7359	0.7348	0.8175

Table 4: Evaluation with the SciVQA validation dataset (1680 questions) on the fine-tuned Qwen2.5-VL 7B model for the different hyperparameters LoRA rank, alpha, and dropout. The learning rate was always 2×10^{-4}, and the learning rate scheduler was linear. The metrics are the F1-scores of BERTScore, ROUGE-1, ROUGE-L, and their average.

#epochs	**BERT**	**ROUGE-1**	**ROUGE-L**	**Average**
1	0.9836	0.7606	0.7591	0.8345
2	**0.9849**	**0.7723**	**0.7709**	**0.8427**
3	0.9848	0.7698	0.7683	0.8410
4	0.9844	0.7652	0.7637	0.8378

Table 5: F1-scores of BERTScore, ROUGE-1 and ROUGE-L with their average across one to four training epochs for fine-tuning Qwen2.5-VL 32B with LoRA rank $= 64$, LoRA alpha $= 128$, dropout $= 0.2$ and 8-bit quantization. The fine-tuning was performed on the SciVQA train split, and the evaluation was done on the SciVQA validation dataset (1680 questions).