2025 Conference on Artificial Intelligence x Multimedia (AIxMM) | 979-8-3503-9097-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/AIxMM62960.2025.00018

Grounding Image Understanding to Oil and Gas Product Manuals: Refining LLaVA through Contextual Instruction Tuning

Hui Wang
SLB
Menlo Park, USA
hwang100@slb.com

Salma Benslimane
SLB
Menlo Park, USA
sbenslimane2@slb.com

***Abstract*—The significance of multimodal models lies in their enhanced capabilities to parse, understand, and reason on complex documents that contain a mixture of text, images, tables, and other components. These models have shown remarkable proficiency in providing coherent and contextually accurate descriptions and analyses, thus significantly advancing document processing tasks. However, to maximize their utility in specific industries, it is crucial to adapt these multimodal models to domain-specific tasks, where specialized knowledge and data are required for precise interpretation and application. In this paper, our approach to multimodal document processing bridges the gap between proprietary large-scale models and smaller open-source models. By combining the robust embedding capabilities of proprietary models with the modular and resource-efficient architecture of open-source models, the proposed method aims to enhance domain-specific image description tasks. Our approach focuses on fine tuning multimodal models on domain-specific data, particularly in the oil and gas industry, by generating data grounded in domain knowledge through the use of proprietary multimodal models. By adapting and fine tuning open-source smaller-scale models on our proprietary datasets, we were able to demonstrate the potential for significant advancements in generating precise image descriptions and answering detailed questions related to image content, thus providing meaningful insights and facilitating downstream applications.**

***Index Terms*—multimodal models, fine tuning, knowledge distillation, captioning**

I. Introduction

The processing of documents and multimodal data is increasingly facilitated by advancements in large language models (LLMs) and large multimodal models (LMMs). Documents such as technical papers, software manuals and training materials in specialized domains (such as oil and gas) typically contain a mixture of textual information, images, tables, equations, and content in other modalities. Thus, it is crucial to develop systems capable of comprehensively understanding and processing the diverse data types and generating meaningful insights.

Proprietary software solutions often encompass domain-specific knowledge and data that open-source models may not fully understand due to their limited availability in the public domain for model training. These specialized documents demand a deeper understanding and domain-specific knowledge to parse and interpret the data accurately. Consequently, there is a growing need to develop systems and models capable of processing such documents and offering various downstream applications, including chatbots, virtual assistants, and copilots, tailored to specific domain requirements.

In recent years, proprietary LMMs such as Gemini [1] and GPT-4 Vision (GPT-4v) [2] have showcased exceptional capabilities in image understanding. These models, trained on extensive multimodal datasets, excel in providing precise image descriptions and answering detailed questions related to image content. However, despite their robust performance, these models come with significant constraints, particularly in terms of accessibility and the substantial computational resources required for their operation.

Conversely, open-source, smaller-sized models (such as LLaVA [3]) have begun to demonstrate promising results in handling images, although they still lag behind their larger counterparts in overall performance. The key advantages of these smaller models lie in their open-source nature, allowing for modularity and adaptability. Users can modify components such as the image encoder, projector, or LLM, providing flexibility absent in larger, proprietary models. Additionally, these smaller models demand considerably less GPU power and computational resources, making them more accessible for fine tuning and retraining on specific domain datasets.

Typically, oil and gas software documentation contains both textual content and pictorial or visual content (images, graphs, screenshots, etc.). While LLMs capture domain textual knowledge quite accurately, open-source data-trained models mentioned above lack some domain understanding when analyzing oil and gas documentation related images. Beyond handling spatial elements, color, and shape well, we need models that grasp domain, taxonomy, and context from images. Therefore, our paper addresses bridging the gap between these models and domain expertise for better contextual understanding.

In this paper our approach leverages the strengths of both categories of multimodal models by integrating the robust embedding capabilities and comprehensive image understanding from proprietary models like GPT-4v with the modular, resource-efficient architecture of smaller open-source models

979-8-3503-9097-1/25/$31.00 ©2025 IEEE
DOI 10.1109/AIxMM62960.2025.00018

Authorized licensed use limited to: SLUB Dresden. Downloaded on August 11,2025 at 18:38:02 UTC from IEEE Xplore. Restrictions apply.

like LLaVA.

Our paper proposes the following contributions:

1) we created a dataset pipeline that merges textural domain context from oil and gas manuals and image descriptions from GPT-4v resulting in a diverse visual instruction-following dataset (including detailed description and categorical question and answering);
2) we fine tuned and benchmarked model generation from different LLaVA versions, backbone language models and compare their performance on specific oil and gas test dataset.

In Section II, we will discuss the related works as well as the current advancements in multimodal models and fine-tuning techniques. Following this, Section III will elaborate on our methodology, the process of dataset creation, and the training specifics. Section IV will provide a comprehensive summary of the experiments, results, and key findings before concluding and outlining future directions in Section V.

II. Related Works

In recent years, multimodal models have become popular due to their ability to process and integrate multiple types of data such as text, images, audio, and more. While these models have demonstrated remarkable capabilities in general-purpose applications, their performance in domain-specific tasks often necessitates an additional step: fine tuning. Fine tuning LMMs is a critical step in harnessing their full potential for specific applications.

A. LMMs

OpenAI's CLIP [4] released in January 2021, marked a significant milestone in the development of multimodal models, showcasing the capability to understand and generate text based on images and vice versa. Following CLIP, several other notable research projects have further advanced the field, including BLIP-2 [5], which enhances the understanding of multimodal content through large-scale datasets, and LLaVA [3], which refines the alignment between language and vision inputs. Moreover, the LLaVA family of models introduced the use of instruction tuning in the multimodal domain, training on a curated dataset collected with GPT-4, making this strategy one of the most promising approaches for building LMMs. OpenFlamingo [6] also emphasizes integrating multimodal data to create robust models. Recent advancements include OpenAI's GPT-4 Vision and GPT-4 Omni [2], extending the GPT series by incorporating vision/audio for comprehensive content understanding and generation, and Gemini [1] , which aims to integrate multiple modalities to develop more versatile AI systems. These developments highlight the rapid progress in multimodal AI, underscoring the importance of integrating diverse data types to create powerful and flexible models.

B. Domain-Specific LMM

LMMs are typically pretrained on vast datasets that encompass a broad range of general knowledge. However, domain-specific tasks often require specialized expertise that is not fully captured during pretraining. We found extensive work on LLaVA that adapts the vanilla LLaVA to a variety of multimodal tasks, making it versatile and useful across different domains. In the medical domain, for instance, [7] adapts LLaVA model to the biomedical domain, first by aligning biomedical vocabulary and then learning open-ended conversational semantics. In the agriculture industry, [8] examined the potential of LLaVA to automate and enhance fresh fruit bunch ripeness assessment. In the field of remote sensing image analysis, [9] fine tunes LLaVA for remote sensing image captioning and question answering, while [10] fine tunes LLaVA to create a remote sensing-domain vision language model—GeoChat. The rationale behind this preference is that, in contrast with proprietary models like GPT-4 Vision and Gemini, LLaVA is a prominent open-source alternative. LLaVA's cost-effectiveness, scalability, and notable performance in multimodal benchmarks, especially LLaVA-NeXT [11], make it an enticing choice for a variety of multimodal tasks, such as visual question answering, image captioning, and visual reasoning.

C. Fine Tuning LMMs

In general, fine tuning has been the predominant methodology for adapting pretrained models to novel tasks or domains. However, with the advent of LLMs and LMMs, this method has become highly resource-intensive, requiring significant computational power and memory. Additionally, full fine tuning risks overfitting and catastrophic forgetting, hindering generalization. To address these challenges, low-rank adaptation (LoRA) [12] has emerged as an efficient alternative. LoRA is a parameter-efficient fine-tuning technique that adapts large pretrained models to specific tasks by introducing low-rank matrices into the model's weight matrices. These low-rank matrices are trainable, while the original-weight matrices remain fixed. This approach significantly reduces the number of parameters that need to be trained, making the fine-tuning process more efficient. In this paper, we use LoRA fine tuning to efficiently adapt large-scale models to new tasks and domains.

III. Method

In this section, we will introduce how to generate a domain-specific instruction following dataset with Teacher LMM and how to fine tune Student LLaVA, as shown in Fig. 1.

A. Dataset Creation

Each image X_v is situated within a document that includes context, captions, and domain-specific information. Since the captions are often missing or not sufficiently descriptive, we use the text that surrounds each image in the document to create context $X_{context}$ corresponding to image X_v.

The focus being detailed image description $X_{caption}$, we create an instruct following dataset $X_{instruct}$ containing both $X_{context}$ and $X_{question}$, such that:

$$X_{instruct} = X_{context} + X_{question} \tag{1}$$

Authorized licensed use limited to: SLUB Dresden. Downloaded on August 11,2025 at 18:38:02 UTC from IEEE Xplore. Restrictions apply.

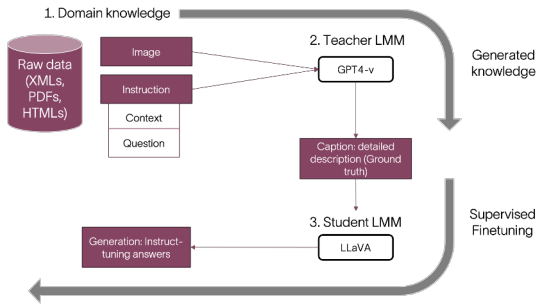

Fig. 1. Overview of the workflow of knowledge distillation: From knowledge generation by a teacher LMM, to supervised low-rank fine tuning of a smaller-scale student LMM.

such that:

$$\begin{aligned} Human : X_v X_{instruct} < STOP > \\ Assistant : X_{caption} < STOP > \end{aligned} \tag{2}$$

We use GPT-4v to generate the captions given the image and the context. By providing both inputs, the caption would contain a detailed description of the image context and spatial information in addition to meaningful terminology and taxonomy of domain-specific terms, plots, and representations.

We generate 25,000 image-caption pairs from oil and gas domain-specific software manuals. Although this dataset mainly represents captioning tasks as instructions for a single task, it covers a rich data variety (Table I) and different software manuals (subsurface, production, reservoir simulation, etc.).

TABLE I
LIST OF IMAGE TYPES COLLECTED FROM OIL AND GAS SOFTWARE MANUAL

Type	Descriprion
Figures	Flowchart, sketching, plotting
Maps	Topography 2D map, geomorphologic map
Software Interface	Screenshots of software window (complete or partial)
3D Models	3D subsurface models, geological representations
Others	Equation, code

B. Model Training and Knowledge Distillation

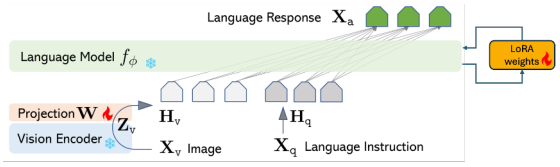

Fig. 2. Our model architecture, built upon the original form of LLaVA architecture in [3], the vision encoder and language model weights are frozen, we only update the projection matrix while LoRA fine tuning. The LoRA fine tuning add extra weights to the original language model.

We use the same architecture of LLaVA, which involves a linear projection layer that connects the visual and textual encoders, enabling the model to understand context and perform tasks that require both visual and textual understanding. The visual encoder, CLIP ViT-L/14, excels at extracting features from images. The language backbone of LLaVA can be the most popular LLMs; e.g., Vicuna or Mistral, even GPT series or LLaMA-3. In this paper, we experimented with Vicuna and Mistral backbone. We keep both the vision encoder and language model weights frozen, only updating the MLP projection matrix during LoRA fine tuning, as shown in Fig. 2. Using LoRA, we refine the parameters Wq and Wv through low-rank adaptation, with a designated rank r set to 128 in our implementation. Each training step incorporates specifically crafted multimodal instructional templates designed for detailed image description and question-answering tasks during the training process. We use AdamW optimizer with a cosine learning rate scheduler to train our model. We fine tune LLaVA with a single A100 GPU.

IV. Experiments and Results

Many domain-specific tasks suffer from a scarcity of labeled data, which poses a significant challenge for training robust models. Fine tuning addresses this issue by requiring relatively less labeled data compared to training a new model from scratch. By leveraging the extensive knowledge acquired during pretraining, fine tuning can effectively use smaller, domain-specific datasets to achieve high performance.

A. Experiments

As a baseline, we use vanilla LLaVA (model with original weights after pretraining) to evaluate its capabilities when dealing with domain-specific oil and gas images. The objective is to compare its performance with our fine-tuned models.

For this purpose, we randomly select 500 unseen images from our dataset and question-answer pairs as our test dataset. Prompts (detailed description questions) were randomly selected from a fixed set of questions from [3] to elicit detailed description responses. Note that no context was provided to the models as part of the prompt, given that this knowledge had been incorporated into the model weights during the fine-tuning process. As an output, the different models generate a detailed description of the input image. We conduct experiments to evaluate models' performance on question-answering tasks and detailed description tasks. We would like to evaluate two hypotheses: performance of LLaVA before and after fine tuning, and the choice of the best language backbone of LLaVA on our dataset.

B. Evaluation

We use the BLEU [13] score as a metric for evaluation along with GPT-4o as an evaluator to rate the responses against the ground truth. The BLEU score is well-established and widely used in the field of Generative AI, which means that it allows for easy comparison with other models and systems. This standardization facilitates benchmarking and helps us gauge the relative performance of our fine-tuned model against pretrained. GPT-4o, has a broad knowledge base and can provide insights on a wide array of topics, is expected to have

Authorized licensed use limited to: SLUB Dresden. Downloaded on August 11,2025 at 18:38:02 UTC from IEEE Xplore. Restrictions apply.

a profound understanding of language, which makes it well-suited for reviewing and critiquing generated text.

We use GPT-4o to assess the correctness of model answers with image context and ground truth. We compare responses to the same question from vanilla LLaVA and fine-tuned LLaVA. Using the responses, question, image caption, image context, and ground truth, GPT-4o scores the helpfulness, relevance, accuracy, and detail level, providing an overall score from 1 to 10. GPT-4o also gives a detailed explanation to help us understand the model responses. We then normalize the scores by computing the relative score using GPT-4o's reference.

C. Results

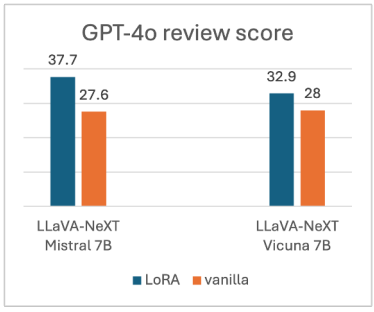

Fig. 3. Comparison between vanilla LLaVA and fine-tuned LLaVA with different backbone LLMs (Vicuna and Mistral). The fine-tuned model learns the domain better and has higher metrics compared to the vanilla model.

From Fig. 3, the GPT-4o review scores for the LLaVA model using both Mistral and Vicuna backbones indicate a substantial enhancement of up to 36% following LoRA fine tuning. Although the initial GPT-4o review score for the Vicuna backbone is slightly higher compared to the Mistral backbone, the Mistral backbone exhibits superior performance post-fine tuning. This suggests that the Mistral backbone demonstrates a higher degree of adaptability to domain-specific data when subjected to LoRA fine tuning. For applications where domain-specific performance is paramount, the Mistral backbone is recommended due to its demonstrated superior adaptability and improved performance metrics after LoRA fine tuning.

D. Discussion

In the comparative analysis of LLaVA-NeXT models, we observed distinct performance differences between the Mistral 7B and Vicuna 7B backbones, both in their vanilla and fine-tuned states. The fine-tuned LLaVA-NeXT Mistral 7B, despite some inaccuracies and irrelevancies, provided more contextually appropriate responses compared to the entirely off-topic and unclear responses of the vanilla version. Similarly, the fine-tuned LLaVA-NeXT Vicuna 7B produced detailed explanations, but with inaccuracies and unsupported assumptions, whereas the vanilla version was entirely irrelevant. When comparing the fine-tuned versions, the Mistral 7B backbone demonstrated superior adaptability to domain-specific tasks, achieving higher GPT-4o review scores despite Vicuna's slightly higher MMMU [14] leaderboard score. Vicuna's responses frequently contained inaccuracies and irrelevant details, leading to lower contextual relevance and accuracy. Conversely, Mistral's responses were more accurate, relevant, and detailed, aligning well with the provided context and resulting in higher performance scores. This indicates that the Mistral backbone is better suited for our domain-specific applications requiring precise and relevant responses.

Our evaluation metrics are BLEU and GPT-4o, while these metrics provide a useful baseline for evaluating model performance, they often fail to capture the full spectrum of contextual nuances essential for advanced natural language understanding, especially in industrial domain. To overcome these limitations, human domain-expert annotators would be an ideal choice.

Additionally, as we push the boundaries of technologies, it is crucial to address privacy and data security concerns. Ensuring robust anonymization techniques and strict compliance with data protection regulations will be essential to maintain user trust and protect customer data.

V. Conclusion

The experiment results proved that our proposed approach successfully bridges the gap between proprietary large-scale models and smaller open-source models for multimodal document processing. By combining the robust embedding capabilities of proprietary models in dataset generation step, with the modular and resource-efficient architecture of open-source models in the LLaVA fine-tuning step, the method enhances domain-specific image-text understanding tasks. The balance achieved by this approach provides meaningful insights and facilitates downstream applications, demonstrating the potential for improved performance and adaptability in domain-specific multimodal tasks.

Although we believe that fine-tuning LLaVA with oil and gas domain data represents a significant step towards making LMMs applicable to specialized industry scenarios, we note that fine-tuned model is limited by hallucinations and weak complex reasoning capabilities that common to many LMMs. As next steps, we can incorporate these models with the retrieval-augmented generation (RAG) process to enhance the model's ability to access and integrate relevant external information, thereby improving the accuracy and richness of generated content.

Acknowledgement

We would like to express our sincere gratitude to all those who have supported us throughout this project. First and foremost, we extend our heartfelt thanks to our leadership, Dr. Swaroop Kalasapur and Dr. Jose Celaya for their continuous support, guidance, and invaluable insights. Their expertise and encouragement have been crucial to the completion of this work.

Authorized licensed use limited to: SLUB Dresden. Downloaded on August 11,2025 at 18:38:02 UTC from IEEE Xplore. Restrictions apply.

We would like to recognize the individual efforts of our group members:

- Aakarshan Dhakal for his dedication to data collection and analysis.
- Dr. Indranil Roychoudhury and Dr. Anatoly Aseev for the exceptional work on the final review and meticulous proofreading.

Thank you to everyone who has been a part of this journey.

REFERENCES

[1] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[5] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[6] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[7] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] J. Y. Goh, Y. M. Yunos, U. U. Sheikh, and M. S. Mohamed Ali, "Vision language models for oil palm fresh fruit bunch ripeness classification," in *2024 IEEE 8th International Conference on Signal and Image Processing Applications (ICSIPA)*, 2024, pp. 1–6.

[9] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, and F. Melgani, "Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery," *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.

[10] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 831–27 840.

[11] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[14] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9556–9567.

Authorized licensed use limited to: SLUB Dresden. Downloaded on August 11,2025 at 18:38:02 UTC from IEEE Xplore. Restrictions apply.