Project Report: Visual Instruction Tuning

Chun-Yu Fang[1*] **Sheza Munir**[1*] **Collin Schultz**[1*] **Matt Schumacher**[1*] **Ritom Sen**[1*]
[1]University of Michigan
{bnfangok, shezamnr, cjschul, ummatt, ritomsen}@umich.edu

Abstract

Multimodal models that integrate vision and language modalities, such as LLaVA, show immense potential in addressing real-world problems. Further, their application to novel domains like aerial image analysis remains underexplored. This study adapts LLaVA, a model combining a pre-trained CLIP vision encoder and a Vicuna language model using a projection matrix, to analyze aerial imagery for tasks like emergency response, urban planning, and environmental monitoring. We propose an alternative LLaVA architecture utilizing the computationally efficient LLaMA-1B as the language model, balancing performance and resource constraints. Training was conducted in two stages: 1) aligning visual features to word embeddings and 2) fine-tuning for conversational tasks. This study also utilizes GPT-4 for evaluation and synthesizing of training data. Evaluations on LLaVA-Bench and the SkyView dataset show that our implementation achieves 25–37% of GPT-4's performance under limited computational resources, highlighting its ability to reason about complex visual contexts despite limitations in dataset size and model capacity. This work illustrates the adaptability of multimodal architectures to novel domains and emphasizes the trade-offs between model complexity, dataset quality, and computational efficiency.

1 Introduction and Problem statement

Multimodal models, which combine visual and textual understanding, have shown tremendous potential addressing complex real-world problems. Despite their versatility, extending these models to novel domains often remains a challenge due to limitations in datasets, computational resources, and adaptability. This work explores how multimodal architectures such as LLaVA can be extended to aerial image analysis, a critical task for emergency response, urban planning, and environmental monitoring. Conceptually, the problem involves adapting a multimodal model to perform detailed analysis of aerial images, requiring reasoning about unique spatial and visual contexts. Mathematically, the goal is to train a model such that the output descriptively answers a query about an aerial image with minimal error. The foundation of this work builds on the paper "Visual Instruction Tuning" by Liu et al. [8], which integrates the CLIP vision encoder with a Vicuna-based language model to create a Large Language and Vision Assistant (LLaVA). Previous studies emphasize the alignment of visual and language modalities, applications in aerial image analysis remain underexplored. This work adapts LLaVA to this domain, demonstrating its potential for novel applications and identifying limitations in resource-constrained environments.

2 Related Work

The development of LLaVA leverages foundational models and frameworks in the vision-language domain. Radford et al. [11] introduced the CLIP model, which serves as LLaVA's vision encoder by aligning visual representations with natural language descriptions, a critical milestone for cross-modal understanding. Based on this, Chiang et al. [4] present the Vicuna model, a finely tuned 13B LLaMA

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

model that achieves conversational quality comparable to GPT-4 and serves as the language decoder of LLaVA. These components are foundational for multimodal instruction tuning, enabling robust visual-linguistic interactions.

Awadalla et al. [2] introduced OpenFlamingo, an open-source framework that benchmarks vision-language models and establishes a comparative baseline for LLaVA's performance. Similarly, datasets such as Microsoft COCO [5] enhance object recognition capabilities with diverse non-iconic images, while Lu et al. [9] introduce chain-of-thought multimodal datasets that improve reasoning benchmarks by providing sequential annotations for complex tasks. OpenAI [10] further advanced multimodal alignment by outlining methodologies for leveraging image-caption pairs in training, which contributed significantly to LLaVA's instruction-following capabilities.

Recent innovations focus on refining the multimodal model architectures and training paradigms. Liu et al. [7] proposed improvements to vision-language connectors, achieving state-of-the-art performance across academic tasks oriented benchmarks while reducing computational requirements. Their work underscores the importance of connector efficiency in multimodal systems. Liu et al. [6] extend this concept in MG-LLaVA by introducing hybrid vision encoders that balance low- and high-resolution image processing. In addition, they incorporate object-level features to enhance the comprehension of intricate visual contexts, making significant strides in multimodal adaptability.

Zhao et al. [13] expand the scope of multimodal instruction tuning by enabling generative and image-editing tasks, demonstrating the versatility of vision-language systems. Complementing these efforts, Zhuang et al.[14] analyze the trade-offs between computational cost and model performance, emphasizing the importance of high-quality pretraining data and efficient tokenization strategies. These works collectively advance the field, paving the way for practical applications in domains such as aerial image analysis, where integrating robust vision and language understanding is paramount.

These developments highlight the rapid evolution of multimodal systems. With innovations in datasets, architectures, and training methodologies, models like LLaVA continue to push the boundaries of real-world applicability, enabling complex tasks that require seamless vision-language integration.

3 Method

3.1 Architecture

In the original paper, the architecture of LLaVA involved a visual encoder, a projection matrix, and an LLM. The pre-trained visual encoder extracts visual features Z_v from an input image X. A projection matrix W, then maps these features to language embedding tokens H_v that the LLM can understand. This way the features from the image are put into tokens with the same dimensionality as the word embeddings in the language model. With this sequence of visual tokens H_v, the LLM can then understand and reason with the image in context. Specifically, with prompt tokens P, and the visual tokens H_v, the LLM can then predict output tokens O. Mathematically this process is shown below.

$$
\begin{gathered}
Z_v = \text{visualEncoder}(X) \\
H_v = W \cdot Z_v \\
O = \text{LLM}([H_v, P])
\end{gathered}
$$

Figure 1: Flow of data through the visual encoder, projection matrix, and LLM.

For this project we recreated this architecture, coding it from scratch. Similar to the original paper, we utilized the pre-trained CLIP vision encoder ViT-L/14 from OpenAI [12] to extract visual features from images. To project these vision features onto the word embeddings space of the LLM, we training a linear embedding layer. Lastly, we utilized Llama-1B [1] as our LLM diverging from the original paper, which experimented with 13B and 7B Vicuna models. Our choice of Llama-1B, a smaller model, was driven by its computational efficiency, simplifying both training and inference processes. Consequently, this adaptation represents an alternative variant of LLaVA, enabling us to examine the trade-offs associated with employing a smaller-scale model.

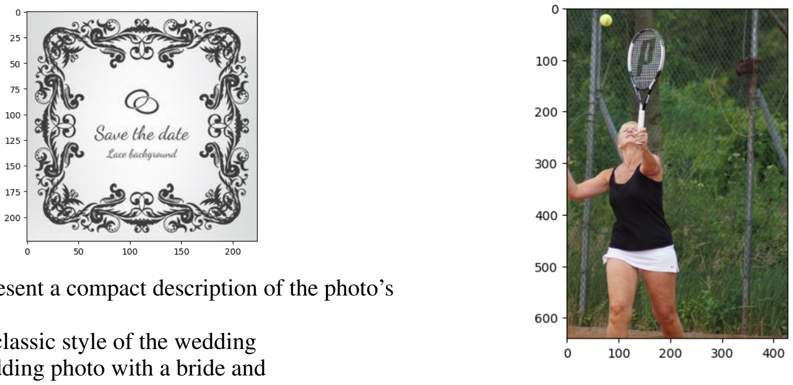
3.2 Training

Training LLaVA was split into two different stages: the first stage for aligning the projection matrix with the existing Llama tokenizer in simple image description tasks and the second stage to fine-tune the entire model for conversational tasks. The CLIP image encoder and Llama's tokenizer were left unchanged during both training stages.

Stage 1

As mentioned above, this stage was used to align the linear projection matrix applied on the CLIP encoded images with the existing tokenizer. To do this, the author's provided a filtered version of the CC-3M image-text dataset[1]. The authors of LLaVA filtered CC-3M to 595,000 image-description pairs. These pairs were then used with GPT-4 to generate different prompts asking LLaVA to describe the image. The training parameters used in this stage largely copy the original authors' parameters with a few changes to account for technical limitations. For this stage, we trained on 175,000 samples with a batch size of 16, 8 gradient accumulation steps, and Adam optimizer with a learning rate of $2e-3$ and 3% warm-up ratio. This was then trained for 1 epoch using a single spgpu Great Lakes Compute Cluster job with 16 cores and 128GB of memory for eight hours.

Figure 2 displays an example of LLaVA's performance after the first training stage both on the training dataset and on a complex reasoning task. Note that the model clearly can denote the main contents of images and describe them but will struggle with the more complex questions during this stage. Figure 2a is also providing a member of the training dataset to compare results to a known ground truth provided during training.



(a) **USER:** Present a compact description of the photo's key features.
LLAVA: the classic style of the wedding photo.###wedding photo with a bride and groom.<|end_of_text|>
GROUNDTRUTH: wedding invitation card with vintage baroque rococo lace , ornate background .

(b) **USER:** How can the woman's positioning affect the outcome of her shot?
LLAVA: up for the tennis tournament<|end_of_text|>

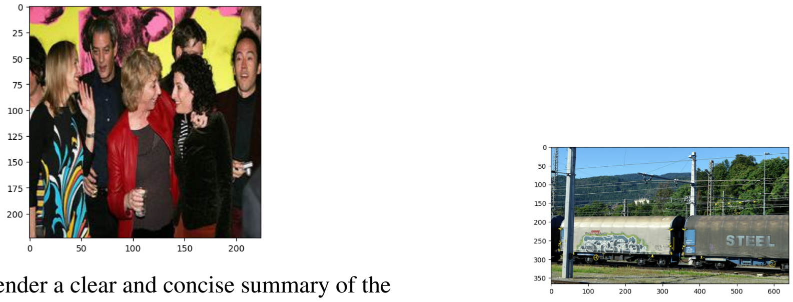
Figure 2: Stage 1 Output Examples

Stage 2

Using *LLaVA-Instruct-158k*[1], all weights of the model (excluding CLIP and the tokenizer) were now optimized for conversational outputs during Stage 2. *LLaVA-Instruct-158k*, built by the authors of LLaVA, used image-text pairs from the existing COCO 2017 dataset[2] and queried GPT-4 to generate multi-turn conversations between a human user and a chatbot. Using the same specifications as Stage 1 for our Great Lakes Compute Cluster, we finetuned our version of LLaVA on 150,000 image-conversation pairs from *LLaVA-Instruct-158k* with batch size of 4, 8 gradient accumulation steps, and the same Adam optimizer but now with a learning rate of $2e-3$. See Figure 3 for example outputs following Stage 2.

[1] https://github.com/haotian-liu/LLaVA/blob/v1.1.3/docs/Data.md

[2] COCO Dataset



(a) **USER:** Render a clear and concise summary of the photo.
LLAVA: ight the image, the person is standing in a room with a group of people, and they are all looking at her. The person is wearing a black dress and a black[a]

(b) **USER:** What are some possible reasons for the graffiti on the train cars?
LLAVA: .jpg">The graffiti on the train cars is likely a result of a recent incident or a ongoing problem. It could be a sign of a vandalism or a protest against the train[a]

[a] The rest of LLaVA's output not shown

Figure 3: LLaVA Stage 2 Output Examples

4 Evaluation

In the original study, the quantitative evaluation involved creating triplets of an image, its raw "ground-truth" textual description, and an associated question. The LLaVA model was then applied to this data set to generate responses to the questions. A larger model, GPT-4 was also used to predict responses based on the raw textual descriptions, serving as a comparison. The evaluator then assessed the LLaVA and GPT outputs based on four criteria: helpfulness, relevance, accuracy, and detail. Following this approach, we also employ GPT-4 as our evaluator model on a randomized set of images, descriptions, and questions. Our evaluation set includes 30 images from the COCO-Val-2014 dataset and 75 images from the SkyView dataset [3], allowing us to align with the original methodology, and evaluate with a novel dataset. Judging performance on the SkyView dataset shows how LLaVA can understand and describe landscapes given images. In the future, these abilities could help in many contexts including emergency landing scenarios.

4.1 LLaVA Bench

To test the effectiveness of our LLaVA model, and how it performed compared to the LLaVA model in the original paper, we leveraged the LLaVA-Bench (COCO) benchmark. LLaVA-Bench was created by randomly selecting 30 images from the COCO-VAL-2014 dataset, and generating 3 questions by prompting GPT-4 to curate this for each image. The questions include a conversational question (ex. How many cars are in the picture?), a question requesting a detailed and comprehensive description, and a complex reasoning question (ex. Is this a good day to go outside?). The full GPT-assisted visual instruction data generation pipeline can be found in section 3 of the Visual Instruction Tuning paper [8]. This benchmark is suitable to test the models' capabilities with consistent images.

Given these questions, the authors of the paper prompted GPT-4 with the same images and questions. This created a theoretical upper bound given that GPT-4 is much larger than our LLaVA model. We then obtained responses from our LLaVA model with the same images and questions, and again leveraged GPT-4 as a judge to measure the quality of the responses given the image. Specifically, GPT-4 evaluated the descriptiveness, clarity, accuracy, and level of detail of the responses, giving scores from 1-10 for each category, and an overall score from 1-10, where a higher score indicates better performance. Lastly, we then compared LLaVA's score to the theoretical upper bound (GPT's response) to get a percentage that shows its relative score. Contrary to the original paper, we leveraged GPT-4 API's new ability to take in images in its prompt when judging between the responses, rather than relying on just "ground truth" descriptions. The results are shown below in Figure 4.

The median in this evaluation was 0.37, meaning that our LLaVA model was around 37% as good as GPT-4's response given the scoring system. In the original paper, LLaVA was around 85% as good as GPT-4. This means our model was a little worse than half as good as LLaVA in the paper. Considering our shortened training schedule, and smaller base LLM (LLaMA-1B), this makes a lot of

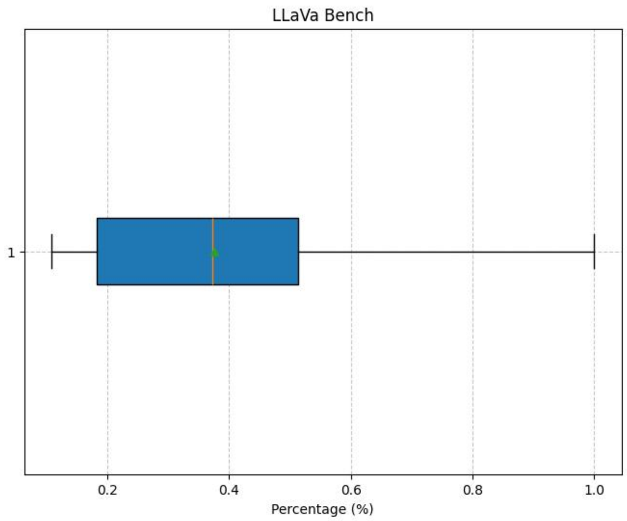


Figure 4: LLaVA performance on LLaVA-Bench (COCO) relative to GPT-4

sense. For the majority of questions, the figure shows that LLaVA performed from a range of 20% to 50% as good as GPT, although there was instances where LLaVA performed just as good as GPT-4.

Overall, our LLaVA model wasn't consistent with its ability to understand the complexities of an image. For example, it struggled to understand a picture of a giraffe's head, but was able to comprehend an image of a woman riding a motorcycle reasonably well. Because of this, when given more complex reasoning questions, the model struggled. However, if LLaVA was given some context for the image in the instructions, it was able to respond to the prompt much better (ex: Why might these giraffes be gathering near the same tree?). These shortcomings are possibly due to the shortened training on our projection matrix. The projection matrix may not have translated the image features properly, and thus the model wasn't able to answer the questions correctly and may have misunderstood what the pictures contained entirely. For example, when it was given the image of the giraffe's head, it mistook it for a dog. Still, given certain images and instructions our LLaVA model was effective at understanding, reasoning with, and describing images, but overall wasn't nearly as effective as GPT-4 or the LLaVA model from the original paper.

4.2 SkyView

To extend the authors' work to a novel setting, we evaluate our model on the SkyView dataset. Specifically, we provide LLaVA an image from this dataset and ask it to determine if it is safe to land a plane in the designated area. This task serves as a means to understand the model's complex reasoning ability. Successfully completing this task demonstrates LLaVA's ability to both interpret visual scenes and evaluate their suitability as landing zones.

To evaluate our results on the SkyView dataset, we employed the strategy described above. In addition, we introduce a baseline response consisting of a random reasoning statement from GPT-4o-mini. This baseline serves as a means to understand how the evaluator judges responses that are not relevant to the provided image. Our results are shown in Figure 5.

From our results, we find that our LLaVA implementation achieves performance of roughly 25% of the reference model. We believe this result can be primarily attributed to the difference in parameters, compute, and data used in each model. Given the computational and temporal constraints our group faced, we used significantly less resources than OpenAI in the training of the reference model. As a result, our LLaVA implementation was unable to include the same level of detail and quality in reasoning statements as GPT-4o mini.

Additionally, the model's architecture limits its ability to conduct a comprehensive analysis of the image. Since the image is encoded by CLIP before the language element of the input is considered, LLaVA inherently lacks the ability to view the image in light of a specific prompt. In the context

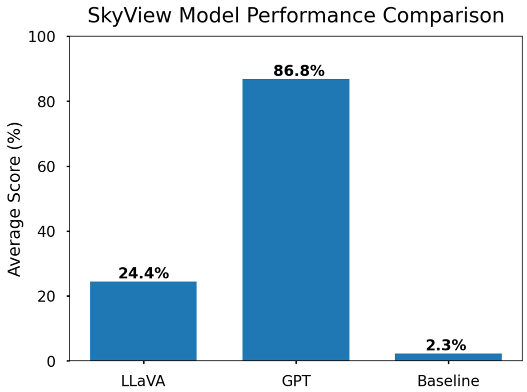


Figure 5: SkyView dataset results for our LLaVA model, GPT-4o mini, and the baseline

of this task, this limitation is significant as features such as bodies of water or rapid changes in elevation play an important role in determining the suitability of a landing site but may not be initially considered by the model.

Despite these limitations, we believe that our results underline the effectiveness of LLaVA's model architecture. With minimal training resources and a smaller model size, we were able to achieve impressive performance on a rather complex reasoning task. Scaling the model in addition to tweaking the architecture may result in impressive results with regard to this task.

5 Conclusion

Through this project, we gained valuable insight into the adaptation of multimodal models to novel datasets. Specifically, we learned how the combined vision and language modalities within LLaVA's architecture can be extended to tasks beyond its original scope, such as analyzing aerial images. A key strength was that LLaVA's architecture integrates CLIP and LLaMA, allowing for effective alignment of visual and language features. Another strength was the use of GPT-4 for generating and evaluating multimodal instruction data, providing a consistent evaluation framework. However, the main weakness we encountered was the limited resources. We chose to train the smaller LLaMA 1B model compared to the Vicuna 13B model used in the paper, and although LLaMA 1B was computationally efficient, our results were limited in reasoning depth. The project successfully demonstrated the feasibility of applying LLaVA to aerial image analysis, achieving meaningful qualitative results and approximately 1/4 of GPT-4's performance despite fewer resources. However, the quantitative performance on the SkyView dataset fell short of initial expectations due to smaller training sets and limited fine-tuning. Challenges included the model's struggles with clearly defining aerial features and the constraints preventing the use of larger models or datasets. In general, the project highlighted the trade-offs between model complexity, dataset size, and computational resources, and leveraging language models for evaluations to capture qualitative performance aspects in complex multimodal tasks.

6 Contributions

6.1 Chun-Yu Fang

I contributed to the project by leading the collection and organization of data and graphs from the training and evaluation stages. I played a key role in designing and creating the presentation poster, taking responsibility for finalizing its content and layout, and ordering its printing for presentation day. Additionally, I co-authored the introduction and conclusion sections of the final paper, and synthesized the team's findings and aligning the report towards broader implications and directions for future studies.

6.2 Sheza Munir

I collaborated with the team to research and identify potential research papers for reproduction and pitched the selected study. Together, we explored novel datasets to extend the original paper's scope. For the project proposal, I co-authored the methods section, contributing to the clarity and feasibility of our approach. During the poster presentation, I assisted the team in refining the poster for effective communication. For the final report, I focused on the related works section, providing a comprehensive context for our research and situating it within the existing literature.

6.3 Collin Schultz

Collin handled training Stage 1 and 2 of the model. This included writing the code for llava_init_wip.ipynb and vicuna_llava.py. The notebook was used as a working coding area for testing the LLaVA model's methods written in the python script.

6.4 Matt Schumacher

Matt worked from Ritom's code to complete the initial version of the Stage 1 training implementation. However, as Ritom mentions, this code was ultimately not used in the final training process. Additionally, Matt was responsible for extending the model to the new dataset by conducting the SkyView evaluation. Regarding the paper and poster, Matt wrote the SkyView evaluation sections and conducted general review. In the project proposal, he wrote the Related Works section. Additionally, he worked with the team to identify potential papers to replicate and datasets to extend the authors' work.

6.5 Ritom Sen

I worked with the team to research potential novel datasets to extend the original paper. For the project proposal I wrote the Novel Component section. Then I wrote code to implement the model architecture and prepared the model for pre-training (Stage 1 of Training), however this code wasn't robust enough and wasn't used. After training, I selected 75 images from the Skyview Dataset [3] to create a dataset for evaluation. I created the prompt and prompted GPT to get it's evaluation for each image and saved this data as a JSON. Then, I took the LLaVA-Bench from the original paper and recreated it with our LLaVA Model, and made a graph from the results. For presentation, I helped our team refine the poster and wrote the LLaVA Bench and Architecture sections. In addition, I helped make edits throughout the paper and ensured that our paper followed the guidelines for our project.

References

[1] Meta AI. Introducing quantized llama models with increased speed and a reduced memory footprint. `https://ai.meta.com/blog/meta-llama-quantized-lightweight-models/`, October 2024. Accessed on December 14, 2024.

[2] Anas Awadalla, Irena Gao, Josh Gardner, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[3] Ankit Bhardwaj and Yessica Tuteja. Skyview: An aerial landscape dataset, 2021. Accessed: 2024-12-14.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Joseph E Gonzalez, Ion Stoica, and Eric P Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. *arXiv preprint arXiv:2303.03025*, 2023.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2015.

[6] Haotian Liu et al. Mg-llava: Towards multi-granularity visual instruction tuning. *arXiv preprint arXiv:2406.17770*, 2023.

[7] Haotian Liu, Chunyuan Li, et al. Improved baselines with visual instruction tuning. *CVPR 2024 Proceedings*, 2024.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[9] Pan Lu, Swaroop Mishra, Tony Xia, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. *arXiv preprint arXiv:2209.09513*, 2022.

[10] Josh Achiam OpenAI et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[13] Yuxuan Zhao et al. Generative visual instruction tuning. *arXiv preprint arXiv:2406.11262*, 2024.

[14] Siyuan Zhuang et al. Llava-next: What else influences visual instruction tuning beyond data? *GitHub project*, 2024.