Enhancing Multimodal Large Language Models for Dynamic GUI Understanding

Sreekant Baheti, Srijan Gupta, Jayavibhav Niranjan Kogundi, Rahil Parikh, Mehul Sethi
University of Southern California
Group 25
`(bahetis, srijangu, jniranja, rahilpar, mehulset)@usc.edu`

1 Tasks That Have Been Performed by Us

Recent advancements in multimodal LLMs have led to key datasets and methodologies for enhancing GUI interaction. Dongping Chen et al. (1) (2024) introduced the GUI World dataset, while Zhuosheng Zhang et al. (2) (2023) proposed multimodal chain of thought reasoning to improve LLM reasoning. Additionally, Jingxuan Wei et al. (3) (2023) developed the COCO Multi Modal Reasoning dataset, and Yangyi Chen et al. (4) (2024) explored adaptations of multimodal models for broader applications.

For the purposes of our project we began by sourcing a dataset from HuggingFace (5), which initially included around 7 million rows. Given its large size, we decided to scale it down to approximately 120,000 rows to make it more manageable for our needs. After a detailed review, we noticed some inconsistencies in the images and JSON structures, such as improper formatting, so we removed those problematic entries, leaving us with a refined dataset of 87,400 rows.

From this refined dataset, we randomly sampled 12,000 examples for fine-tuning, using a random seed of 2020. We chose random sampling because gathering a balanced set of different image types was quite challenging. We then reformatted the dataset to align with our model's requirements, ensuring that user queries and assistant responses, including both text and images, were properly paired for training. For this project, we used the Llama-3.2 11B Vision Instruct model (quantized to 4-bit) (6) and fine-tuned its language layers, MLP, and attention layers using LoRA (Low-Rank Adaptation) (7). We set the LoRA parameters with a rank of 16 and an alpha of 16.

For the training process, we configured the model with 125 warmup steps, 1250 total training steps, a total batch size of 8 (computed as $2 \times 4 \times 1$), and a learning rate of 2e-4. This setup resulted in 52,428,800 trainable parameters out of 11 billion, which is about 0.48% of the total. We used the Adam optimizer with weight decay to ensure effective training. The results of our fine-tuned model, compared to the base model, are presented in Table (1).

Table 1: Evaluation of Base Model vs Fine-tuned Model

Metric	Base Model	Fine-tuned Model
BLEU (NLTK)	0.1988	**0.6098**
BLEU (SacreBLEU)	21.44	**63.92**
ROUGE-1	0.5240	**0.8392**
ROUGE-2	0.4647	**0.7333**
ROUGE-L	0.5045	**0.8183**

We evaluated our fine-tuned model against the base model using a test set of 200 samples. The table shows improvements across all metrics. For example, the BLEU (NLTK) score increased from 0.1988 to 0.6098, and the BLEU (SacreBLEU) score rose from 21.44 to 63.92, suggesting better accuracy and fluency in the generated responses. The ROUGE scores also improved, with ROUGE-1 going from 0.5240 to 0.8392, indicating enhanced alignment with reference outputs in terms of precision and recall. These results demonstrate the positive impact of our fine-tuning approach using LoRA. The training loss over the steps is shown in Figure (1). Table (2) provides a comparison of the outputs of the base model with the fine-tuned model.

Figure 1: Training loss vs. steps

Table 2: Comparison of the base model with the fine-tuned model

Question	Base Model Output	Fine-Tuned Model Output
1. Reply to the question below using a single word or brief phrase: What is the date of Bridge Pedal?	'05	Sunday, August 23, 2009
2. What is the name of the company? Based on the screenshot, provide your answer in one word or phrase.	event essentials	Event Essentials Pty Ltd
3. Identify and provide the bounding box for the element described by: "Facebook"	I don't have access to that information.	[0.043, 0.962, 0.133, 0.995]
4. What activities can be done on the Estacade beach? Look at the image and construct a detailed response to the question.	On the Estacade beach, you can enjoy the following water-based activities: **Shell-fishing**, **Surfing**, **Kitesurfing**	Based on the webpage content, I found a section describing activities that can be done on the Estacade beach, which includes **Swimming**, **shell-fishing**, and other sports such as **kitesurfing** and **sailing**.

2 Risks and Challenges That We Think We Need to Address by the Project Deadline

One of the main challenges we faced was finding a dataset that fit our needs, as it had to be in a specific format for our model training. This is something we need to address to ensure our workflow runs smoothly. Additionally, we were limited by our computational resources, which made it difficult to work with larger datasets that could potentially improve our model's performance.

Another key issue was selecting the right training parameters. Choosing incorrect parameters could lead to problems like overfitting or underfitting, which would negatively impact our results. We also initially aimed to make our model compatible with multiple platforms, such as Windows, Linux, and macOS, but this proved challenging since not all users have access to the same operating systems. Finally, our model was trained only on English data, which limits its use for GUIs in other languages, as not everyone is fluent in English.

3 Our Plan to Mitigate the Risks and Address the Challenges

To tackle the challenge of dataset selection, we spent time researching various options and ultimately chose the HuggingFace dataset because it best matched our project's needs in terms of structure and content. To address the issue of limited computational resources, we further reduced our dataset from 87,400 to 12,000 rows for fine-tuning, using random sampling since collecting diverse image types was difficult.

For the challenge of selecting training parameters, we experimented with different settings through a trial-and-error approach until we found a configuration that gave us satisfactory results. As for cross-platform compatibility, we realized that supporting multiple operating systems was not feasible within our current scope, so we decided to focus solely on Windows-based applications. Lastly, to address the language limitation, we looked for datasets in other languages but couldn't find suitable ones, so we adjusted our goal to focus only on English-based data.

4 Individual Contributions

Our team worked collaboratively, with each member contributing to different aspects of the project. Jayavibhav Niranjan Kogundi, Sreekant Baheti, and Mehul Sethi took on the task of researching and identifying suitable datasets, ensuring we had the right data for our goals. The same group—Jayavibhav, Mehul, and Sreekant—also handled the sampling and fine-tuning of the dataset, preparing it for model training.

Jayavibhav Niranjan Kogundi, Srijan Gupta, and Rahil Parikh managed the model training, overseeing the fine-tuning and optimization processes. Srijan Gupta and Rahil Parikh then evaluated the model, analyzing its performance and identifying areas for improvement. Sreekant Baheti, Mehul Sethi, Srijan Gupta, and Rahil Parikh worked together to research potential applications for our model, exploring how it could be used in real-world scenarios. Finally, all of us—Jayavibhav, Srijan, Rahil, Sreekant, and

Mehul—contributed to preparing this status report, ensuring it captures our progress clearly.

References

[1] Dongping Chen, et al. GUI World A Dataset for GUI oriented Multimodal LLM based Agents. Technical Report, 2024

[2] Zhuosheng Zhang, et al. Multimodal Chain of Thought Reasoning in Language Models. Preprint, 2023.

[3] Jingxuan Wei, et al. Enhancing Human like Multi Modal Reasoning A New Challenging Dataset and Comprehensive Framework. Preprint, 2023.

[4] Yangyi Chen, et al. Reformulating Vision Language Foundation Models and Datasets Towards Universal Multimodal Assistants. Technical Report, 2024.

[5] Neulab. MultiUI: A Dataset for GUI-oriented Multimodal LLM-based Agents. HuggingFace, 2024. https://huggingface.co/datasets/neulab/MultiUI.

[6] Meta AI. Llama-3.2 11B Vision Instruct Model. Meta AI, 2024. https://meta.ai/llama.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. https://doi.org/10.48550/arXiv.2106.09685.