11-777 Final Report

Akshay Badagabettu* **Nikolaj Hindsbo*** **Aayush Shah*** **Sai Yarlagadda***
{abadagab, nhindsbo, aayushsh, saisravy}@andrew.cmu.edu

Abstract

Multimodal Large Language Models (MLLMs) excel at integrating textual and visual modalities but face challenges in web-based tasks due to limited benchmarks. The VisualWebBench benchmark consists of diverse tasks which test MLLMs' understanding and grounding capabilities in web scenarios.

In our project, we aim to analyze and enhance MLLMs' performance on VisualWebBench through model fine-tuning, prompt engineering, and detailed evaluations of baseline models such as LLaVA, Phantom and TroL. Key contributions include fine-tuning LLaVA-v1.5-7b on the MultiUI dataset and analyzing model behavior to address alignment and spatial reasoning challenges. These efforts aim to advance robust multimodal learning for web-based environments. Codebase can be found in this GitHub Link

1 Introduction and Problem Definition

In recent years, Multimodal Large Language Models (MLLMs) have demonstrated significant potential in performing tasks that require understanding and integrating both textual and visual modalities. However, their evaluation in web-based environments has remained a challenge due to the lack of granular benchmarks that measure fine-grained skills such as optical character recognition (OCR), grounding, and semantic comprehension. To address these gaps, the VisualWebBench dataset has emerged as a comprehensive benchmark tailored to web-based tasks.

This project focuses on improving the performance of MLLMs on VisualWebBench by leveraging architectural innovations and training strategies. VisualWebBench includes diverse tasks such as **Webpage Question Answering (WebQA)**, **Action Grounding**, **Action Prediction**, **Element Grounding**, **Element OCR**, **Heading OCR**, and **Webpage Captioning**. These tasks require models to process and integrate visual and textual elements effectively, showcasing the fundamental multimodality of this dataset.

Our work builds on previous analyses and baselines, highlighting the strengths and limitations of current MLLMs like Phantom, LLaVA, and TroL. *While these models exhibit promising capabilities in tasks like OCR and captioning, they struggle with alignment and spatial reasoning in grounding tasks.* Our goal is to propose and evaluate novel modifications to existing models, aiming to address these challenges and enhance their ability to generalize across VisualWebBench tasks.

This report outlines our contributions:

1. A comprehensive evaluation of baseline models, including unimodal and multimodal variants, to identify their strengths and weaknesses across VisualWebBench tasks.

2. Insights into the effects of hyperparameters, input resolutions, and loss functions on model performance and a detailed analysis of failure cases and intrinsic metrics to guide subsequent improvements.

3. Fine-tuned LLaVA-v1.5-7b on MultiUI dataset to improve its web-page understanding capabilities.

4. Devised and implemented prompt enhancement techniques tailored for Element OCR and Action Prediction tasks to guide the model in generating its output.

Through this project, we aim to advance the state of multimodal learning and contribute to the development of robust, interpretable models for web-based environments.

*Everyone Contributed Equally – Alphabetical order

2 Related Work and Background

In recent years, there has been significant interest in Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs) capable of understanding user interfaces (UIs) (Rahman et al., 2024; Baechler et al., 2024; Gur et al., 2018; Li and Li, 2023). These advancements have paved the way for autonomous agents that navigate UIs using natural language commands provided by users.

2.1 Autonomous Agents and Navigation

Our focus is on autonomous agents designed for navigating UIs based on natural language instructions. Examples include AppAgent (Zhang et al., 2023), MobileAgent (Wang et al., 2024), CogAgent (Hong et al., 2024), Auto-UI (Zhang and Zhang, 2024), and V-Zen (Rahman, 2024). These agents rely on multimodal models to understand and act within GUI environments. A related problem is grounding, which involves identifying UI elements corresponding to a specific phrase. Models like Ferret (You et al., 2024) and Ferret v2 (Zhang et al., 2024) tackle this challenge effectively.

2.2 Language-Only Agents for Web Interaction

Another domain of interest is the use of language-only large language models (LLMs) as agents for web interaction, primarily using HTML representations. For example, (Deng et al., 2024b; Gur et al., 2023) have explored HTML-based navigation. Pix2Struct (Lee et al., 2023) was pre-trained to reproduce simplified HTML representations of screenshots, including masked parts of the UI. Such models enable grounding-style pretraining and screen content interpretation.

2.3 Grounding and GUI Understanding

Grounding is a critical component for multimodal agents. SeeClick, one of the most effective agent models, uses HTML data for grounding-style pretraining (Cheng et al., 2024). It demonstrates strong performance in identifying UI elements based on descriptions. Additionally, multitask training approaches, such as those by (Gao et al., 2024), integrate real-world task simulations to enhance model robustness.

One factor that significantly influences the visual perception and understanding capabilities of MLLMs is image resolution. The input resolution is determined by the vision encoder of the model, with low-resolution images often cause printed text to become blurred or unreadable. Leapord (Jia et al., 2024) proposes dividing high resolution images into multiple smaller subimages by image-splitting idea and adaptive high-resolution multi-image encoding strategy. We utilized dividing the images in the cropping strategy employed for our prompt enhancements.

2.4 MultiUI Dataset

The MultiUI dataset significantly advances text-rich visual understanding by synthesizing multimodal instructions from over 1 million websites. MultiUI's 7.3 million samples enable models to excel across GUI-specific benchmarks such as VisualWebBench and Mind2Web while also generalizing to document and chart understanding. Its diverse taxonomy—spanning visual reasoning, OCR, and grounding tasks—demonstrates the dataset's capability to bridge web-based and general multimodal applications. Training on MultiUI underscores the dataset's role in enhancing multimodal generalization (Liu et al., 2024c).

2.5 Important baselines

There are several baselines that have been run on the VisualWebBench dataset. TroL (Lee et al., 2024b), Phantom (Lee et al., 2024a), and LLaVA (Liu et al., 2024a) are particularly important for our analysis. TroL utilizes a novel architecture which re-uses the transformer layers during forward propogation, reducing the memory requirements, but improving the performance. Phantom utilizes a Phantom dimension in the latent layer to improve performance. However all these models have a major disadvantage which is their inability to locate UI elements. Our proposed model aims to mainly tackle this problem.

Related Datasets Table 1 shows a comprehensive overview of four key web and multimodal agent evaluation datasets.

Dataset	Purpose	Tasks	Key Features	Performance Insights
WebArena (Zhou et al., 2023)	Evaluate web agents on realistic, multi-step tasks	812 tasks across e-commerce, social forums, collaborative development, and content management	Real-world data (90k product entries, 127k subreddit posts), functional website replicas, natural language intents	Highlights challenges in adapting autonomous agents to real-world web interactions.
VisualWebArena (Koh et al., 2024)	Evaluate multimodal agents on visually grounded web tasks	910 tasks across Classification, Shopping, and Reddit	25.2% image-text inputs, realistic visual data, tasks like image-based search, VQA, and contextual decision-making	Human success: 88.7%, Best model (GPT-4V): 16.37%; significant gap highlights difficulty in visual understanding and task integration.
Mind2Web (Deng et al., 2024a)	Test generalist agents on open-ended tasks across real-world dynamic websites	2,350 tasks from 137 websites spanning 31 domains	Complex, multi-step interactions, >1,000 elements per page, high-level instructions without step-by-step guidance	Familiar tasks: 52% step success; unseen domains: 40%; challenges in generalization across websites and domains.
LiveBench (White et al., 2024)	Real-time evaluation of LLMs, mitigating contamination and evaluation biases	Questions across math, coding, reasoning, language comprehension, instruction following, and data analysis	Regularly updated, sourced from current materials (e.g., arXiv, news), automatic scoring against objective ground-truth answers	Top models <65% accuracy; highlights ongoing challenges in advancing LLM capabilities across diverse and complex tasks.

Table 1: Summary of Key Datasets for Evaluating Web and Multimodal Agents, Highlighting Their Purpose, Task Characteristics, Features, and Performance Insights

3 Task Setup and Data

Summary of the dataset: The dataset consists of 1500 samples, each representing a webpage from 139 real world websites. These websites span a wide range of industries and sectors, contributing to the diversity of data. The samples are drawn from 12 different domains (sports, animals, science, and etc) and 87 different sub-domains. Each website has a unique user-interface and structure. For example, an e-commerce site focuses on product displays and filters, while blog consists of long-form text and navigation through dropdowns buttons. Images are high-resolution website screenshots (1280 pixels wide). The total size of the dataset is 1.18GB and is downloadable on HuggingFace at the following link.
VisualWebBench has divided the tasks into seven major categories. A brief summary of each task is given below.

- **Action Prediction:** This task requires MLLM's to predict the title of the webpage after clicking a specific element in a bounding box.
- **Action Grounding:** This task asks MLLMs to determine which element to click in a webpage to fulfill a specific human instruction
- **Element Grounding:**This helps in understanding MLLMs' ability to align image and text data by locating an HTML element in the webpage screenshot based on its description. MLLMs select the correct bounding box from eight candidates, using the extracted description as a guide.
- **Element OCR:** This task provides a screenshot with a bounding box indicating the test to be recognized.
- **Webpage QA:** In this task, MLLM's are required to answer open-ended questions based on the webpage's visual layout.
- **Heading OCR:** This task involves getting the heading text from the screenshot of the website.
- **Captioning:** This task evaluates MLLM's ability to generate high-quality meta descriptions for screenshots of webpages.

Table 2 shows the evaluation metrics used for the 7 different tasks in VisualWebBench.

Task	Evaluation Metric
Captioning	ROUGE-L
WebQA	F1 score
Heading OCR	ROUGE-L
Element OCR	ROUGE-L
Element Grounding	Accuracy
Action Prediction	Accuracy
Action Grounding	Accuracy

Table 2: The benchmark uses different metrics for different tasks.

We have used a subset of the MultiUI Dataset (Liu et al., 2024c) to fine-tune LLaVA-1.5 7B model. More details on the dataset, training strategies, and setup is in Section 5.

4 Baselines

4.1 Unimodal and Simple Multimodal Baselines

We choose five baseline models. Out of the five models, three models, viz., LLaVA, TroL, and Phantom, have both multimodal and unimodal variants. The unimodal analysis has been done by masking the encoders of one of the modalities. For example, if we want to create a text-only unimodal model, the vision encoder and, in some cases, the vision projector have been masked out completely.

4.2 Simple Multimodal Baseline

We have chosen InternLM2 (Cai et al., 2024a) (1.8B parameter) as a simple baseline model. The reason we chose this model is that it has a very basic architecture that is similar to that of LLaMa (Touvron et al., 2023), and it has not been trained on data similar to website images. Some key architectural features in InternLM include:

1. Like most modern large language models, InternLM relies on the transformer architecture. It only has a decoder component. This architecture includes an attention mechanism. The use of self-attention layers helps the model capture dependencies between words regardless of their distance from one another.

2. The model's architecture is very modular. That means it can be fine-tuned very easily for various downstream tasks.

3. The models were released in various sizes (1.8B, 7B, 20B) and performed very strongly on various evaluation metrics. The authors have also provided models at multiple stages of training to support research post-SFT and RLHF.

4. InternLM2 was designed with a context window of 200k and performed really well in the Needle in Haystack test.

5. They introduced the COOL RLHF technique, and it has improved the model's performance in conversation-related tasks.

The model has performed poorly on most of the tasks that it was tested out on. The results can be viewed in Table 3. But one observation was that the model was able to accurately reason about the question, but it was not able to reach to the final answer.

4.3 Unimodal, Multimodal and Competitive Baselines

4.3.1 LLaVA-7B - multimodal and unimodal (text)

Our first baseline model is LLaVA (Liu et al., 2024b). LLaVA was selected for a few reasons. Firstly, it was used in VisualWebBench for model scaling analysis (7B, 13B, and 34B models) and is used as the base model in their analysis on GitHub. Secondly, the simplicity and efficiency of LLaVA make it an ideal starting point for us. Also, it has both unimodal and multimodal options. Lastly, it outperforms many more complex models and has a great overall performance on 12 different benchmarks (Liu et al., 2024a).

The unimodal version of LLaVA takes text as its method of input. The multimodal version of LLaVA uses a simple projection matrix to connect its vision and language components. The vision encoder is a pre-trained CLIP ViT, and the language model is Vicuna (LLaMA variant). LLaVA undergoes a two-stage tuning process:

1. Pre-training on image-text pairs.

2. Fine-tuning on multimodal instruction-following data.

4.3.2 TroL-1.8B - multimodal and unimodal (text)

TroL stands for Traversal of Layers. The main idea behind this paper (Lee et al., 2024b) was to artificially increase the number of layers, i.e., increase the number of layers without increasing the number of parameters of the model.

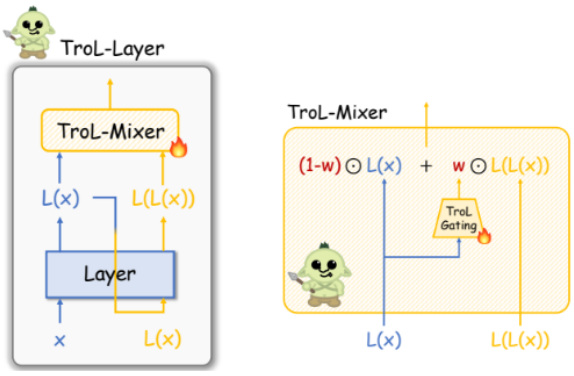


Figure 1: Visualized TroL-Layer and Trol-Mixer Architecture

Figure 1 shows the architecture of the TroL model. The TroL layers have been integrated into a backbone multimodal machine learning model. In

a single layer of this model, the input x is passed to a layer L; this produces an output L(x). Typically, in models, this output is given to the next layer, but in TroL, L(x) is passed back to the same layer, yielding L(L(x)). The final output from the layer is determined by a mixing ratio w. The values L(x) and L(L(x)) are weighted and added. The mixing ratio w is learned by a TroL gating, which is a two-layer MLP with GeLU activation.

There are three model sizes of TroL viz., 1.8B, 3.8B, and 7B. We have run the 1.8B parameter model as part of the competitive baseline. For training, the image is passed through a vision encoder (CLIP-L or InternViT) and a vision projector, which is an MLP, to match the hidden dimension of the encoding to the hidden dimension of the backbone MLLM. The model has been trained on 2.3M samples. For the unimodal analysis, the vision encoder and vision projectors were masked out, and only the prompt was passed to the model. The results and their analysis have been elaborated in Section 2.

4.3.3 BLIP-470M - unimodal (vision)

We performed unimodal analysis on images using the BLIP2-470M model, focusing solely on the visual modality while masking out the textual modality entirely. This was done to assess whether the model could extract core information from the images without relying on accompanying text. BLIP2 was introduced in the paper BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (Li et al., 2022). Some of the key contributions of this paper are:

1. BLIP uses a unified vision-language model architecture designed to handle both vision-language understanding (e.g., image-text retrieval) and generation (e.g., image captioning). This versatility makes the model applicable to a wide range of tasks.

2. The model architecture consists of a Vision Transformer (ViT) for image encoding and a transformer-based text encoder. These are fused together through cross-attention layers, enabling the model to jointly process visual and textual data.

3. A key innovation is the bootstrapping process, which involves using self-generated captions to iteratively improve the model. This method allows BLIP to learn from noisy image-text pairs without requiring high-quality, manually annotated data.

4. BLIP was pre-trained on several tasks, including image-text matching, masked language modeling, and image captioning. These tasks help the model capture both vision-language alignment and generative capabilities.

5. The model has demonstrated strong performance across a range of benchmarks, including image captioning, visual question answering (VQA), and image-text retrieval. BLIP also outperformed many previous models on these tasks.

4.3.4 Phantom-7B - multimodal and unimodal (text)

Phantom (Lee et al., 2024a) is the most recent work to use VisualWebBench, which claimed to outperform closed source models by making use of a method they termed as 'phantom dimension.' The phantom dimension refers to the expansion of latent hidden dimension in the attention mechanism to enhance the alignment between vision and language components in multimodal transformers. The key idea is to use the query, key, and value vectors at the location of the start-of-sequence (SOS) token and append them to the original query, key, and value matrices to expand the dimensions. To match the dimensions of these matrices, first, we inject the query, key, and value vectors of the SOS token into a multi-head cross-attention (MHCA) module. Then, they augment the query, key, and value components of self-attention with this additional latent hidden dimension, denoted as d. The outputs of the augmented query, key, and value are then processed through multi-head self-attention (MHSA) in order to maintain the latent dimension's contribution throughout the layers of the model. Finally, the outputs are compressed using a weighted-average mechanism, minimizing the information loss during the transformation. This is done at each layer of the network, where the outputs are expanded to match the number of tokens in the input sequence. The purpose of this is to allow the model to retain richer information in the cross-attention phase, helping the model better integrate the vision-language features.

The equation for passing inputs into MHCA and expanding the dimensions:

Model	Method	Website			Element		Action		Average
		Caption	WebQA	HeadOCR	OCR	Ground	Prediction	Ground	
LLaVA-7B	Multimodal	19.11	**39.64**	19.11	**52.01**	**31.23**	1.07	10.68	**27.85**
LLaVA-7B	Unimodal (Text)	8.53	8.25	2.72	0	9.93	6.76	8.74	6.85
TroL-1.8B	Multimodal	23.06	5.51	59.59	30.48	24.69	14.59	**25.24**	26.17
TroL-1.8B	Unimodal (Text)	10.69	4.55	5.19	4.12	16.95	12.81	19.42	10.53
Phantom-7B	Multimodal	**24.5**	3.8	**67.5**	12.2	18.4	**35.9**	19.4	23.2
Phantom-7B	Unimodal (Text)	12.1	2.3	5.8	6.7	12.6	16.7	17.5	9.5
InternLM-1.8B	Multimodal	7.03	3.8	1.96	9.12	7.26	4.63	5.83	5.66
Blip-470M	Unimodal (Images)	8.58	9.51	3.42	7.67	12.6	0	0	5.97

Table 3: Our baseline findings - computed performance comparison of various MLLM models on VisualWebBench tasks

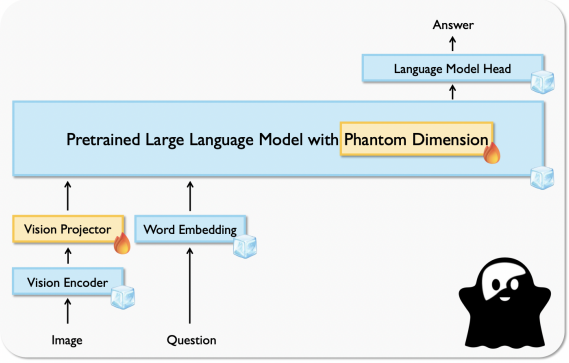


Figure 2: Phantom model architecture and components.

$$Q_l^* \leftarrow \text{MHCA}(q = Q_l, k/v = Q_l^*),$$
$$K_l^* \leftarrow \text{MHCA}(q = K_l, k/v = K_l^*),$$
$$V_l^* \leftarrow \text{MHCA}(q = V_l, k/v = V_l^*)$$

where the dimensions are expanded to:

$$[Q_l, Q_l^*] \in R^{N \times h_q \times \frac{2d_q}{h_q}}, \quad [K_l, K_l^*] \in R^{N \times h_k v \times \frac{2d_{kv}}{h_{kv}}},$$
$$[V_l, V_l^*] \in R^{N \times h_k v \times \frac{2d_{kv}}{h_{kv}}}.$$

Then, multi-head self-attention (MHSA) is applied to the concatenated matrices in the multimodal LLM:

$$O_l = \text{Softmax}\left(\lambda \left(\frac{2d_q}{h_q}\right)^{-\frac{1}{2}} [Q_l Q_l^*][K_l K_l^*]^T\right)[V_l V_l^*]$$

where λ is a regularization parameter and O_l is the output of MHSA.

Fig 2 represents the architecture used by Phantom, where the open-source MLLM used by them is enhanced with the phantom dimension as formulated before. They also utilize a vision encoder, which is the InternViT-300M model (Chen et al., 2024), and the vision projector is two fully connected layers with GELU activation. The large language model consists of a language head and word embeddings. The best-performing model reported by the authors is the Phantom-7B, which is based on the InternLM2.5-7B multimodal LLM (Cai et al., 2024b). For our multimodal and unimodal (text) baseline, we passed images and only text prompts, respectively, from VisualWebBench to the Phantom-7B model and evaluated the performance. The results have been tabulated in Table 3.

5 Proposed Model

Our approach leverages LLaVA 1.5, a lightweight, open-source MLLM with 1.5B-7B parameters, fine-tuned on the MultiUI dataset to address the nuanced challenges of VisualWebBench. The proposal is structured into two stages: Version 1, focusing on fine-tuning for foundational multimodal capabilities, and Version 2, introducing prompt engineering and task-specific cropping techniques for targeted improvements.

5.1 Motivation for Using MultiUI

The MultiUI dataset represents a groundbreaking benchmark for improving multimodal model capabilities in web environments. It demonstrates remarkable utility in enhancing model performance not only on web UI tasks but also across domains like document understanding, OCR, and chart interpretation. Models trained on MultiUI achieve:

- **48% improvement on VisualWebBench tasks**, a testament to its direct relevance for web-based multimodal tasks.
- **19.1% boost in element accuracy on Mind2Web**, showcasing its generalization capabilities.
- Strong transferability to non-web UI domains, making it a versatile dataset for multimodal learning.

Given these compelling advantages, we fine-tune LLaVA 1.5 on MultiUI to enhance its ability to process, interpret, and reason about visual and textual elements in complex web environments. By grounding the model in the rich diversity of MultiUI, we aim to significantly improve its performance on VisualWebBench tasks while preserving generalizability to related non-web domains.

5.2 Model Pipeline

Figure 3 outlines the pipeline, where Version 1 of our model is the stage where fine-tuning of the base LLaVA model on MultiUI happens. Version 2 is where our applied enhancements in prompt engineering and data pre-processing happen.

5.2.1 Version 1: Fine-Tuning with MultiUI

The goal of Version 1 is to establish a robust foundation for multimodal reasoning by fine-tuning the LLaVA 1.5 model on MultiUI. Key highlights of this stage include:

- **Dataset Features**: MultiUI provides high-resolution webpage screenshots (1280px width) from 139 diverse websites, spanning 12 domains and 87 subdomains. It includes challenging UI components such as product displays, dropdown menus, and dense text layouts.
- **Training Objectives and strategies**:

  The fine-tuning process leverages Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically using Low-Rank Adaptation (LoRA), to adapt a pre-trained large vision-language model efficiently. PEFT methods like LoRA introduce a small number of trainable parameters to the model, focusing on low-rank decompositions of the weight updates, which significantly reduces computational and memory requirements. The MultiUI dataset has 7.3M samples, but we trained it on a random subset of that. For the first checkpoint, we trained it on 12.5k data points for 5 epochs, and then for the final fine-tuned model, we continued training for 2 epochs with an additional 60k data points.

  - **Action Prediction**: Predicting the outcome of clicking an element in a bounding box.
  - **Action Grounding**: Identifying the clickable element to fulfill a user instruction.
  - **Element OCR and Heading OCR**: Transcribing text from webpage elements, often in complex or noisy layouts.
  - **Captioning and Webpage QA**: Generating high-quality descriptions and answering questions based on webpage layouts.
- **Loss Functions**: We used the instruction-following loss of LLaVA when fine-tuning it with the MultiUI Dataset. The instruction-following loss used for fine-tuning LLaVA is a token-level cross-entropy loss, designed to align the model's outputs with human-annotated responses for multimodal tasks. During fine-tuning, the model processes a combination of visual embeddings (from a pretrained vision encoder, such as CLIP or BLIP-2) and textual prompts (natural language instructions - which are the seven tasks in our case). The predicted tokens generated

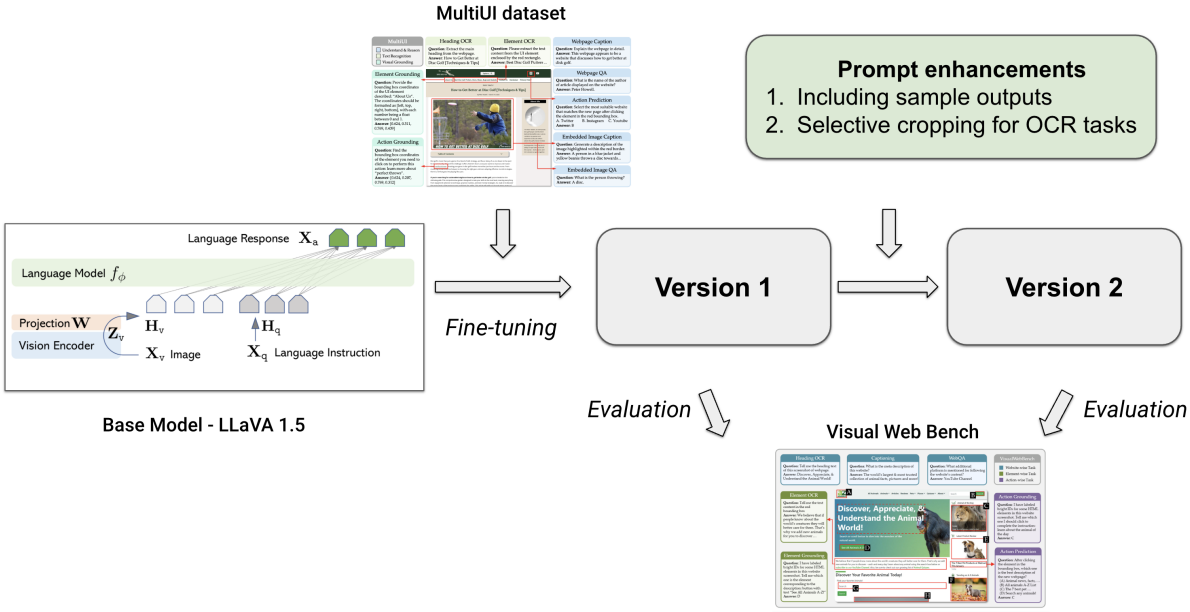


Figure 3: Overview of our approach: In Version 1, we are fine-tuning LLaVA 1.5 on MultiUI dataset to enhance its web understanding capabilities. For Version 2 we are applying prompting techniques to improve performance for a subset of tasks in Visual Web Bench.

by the decoder are compared to the ground truth response tokens, and the cross-entropy loss penalizes deviations from the expected sequence. This loss encourages the model to produce contextually accurate, and instruction-following outputs.

- **Hyperparameters and loss curve:**

  Table 4 shows the hyperparameters used when fine-tuning LLaVA-1.5 7B and Figure 4 shows the training loss curve.

Hyperparam	Value
LoRA Rank	256
LoRA Alpha	512
Projector LR	2e-5
Epochs	5 + 2
Batch size	32
Learning rate	2e-4
Warmup ratio	0.03

Table 4: Hyperparameters when training.

- **Expected Impact:** Fine-tuning on MultiUI equips LLaVA with stronger visual-text alignment, robust OCR capabilities, and improved grounding accuracy. This stage prepares the model to handle the complexity of VisualWebBench tasks while generalizing well to non-web domains like document processing and chart interpretation.

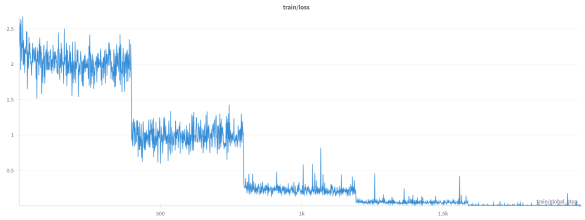


Figure 4: Training loss curve

5.2.2 Version 2: Prompt Engineering and Selective Cropping

Building on the improvements from fine-tuning, Version 2 focuses on task-specific adaptations to further optimize performance on VisualWebBench.

1. **Prompt Engineering**
   - **Problem**: Default prompts often fail to provide clear instructions, leading to ambiguity in tasks like action grounding and webpage QA.
   - **Solution**: We design task-specific prompts with:
     - Contextual examples (few-shot prompting).
     - Explicit task instructions (e.g., step-by-step reasoning for action grounding).

  - Structured formats to standardize outputs and reduce hallucinations.
  - Guidelines for OCR tasks to prevent the model from making common mistakes, like expanding acronyms or summarizing big numbers.
- **Example**: For action grounding, the prompt explicitly asks, *"Which element should I click to complete the task?"* while highlighting relevant bounding boxes and instructions. Additionally, in action prediction, the default prompts did not clearly specify how to select an option. So an ordered list was introduced with explicit labels such as 'A', 'B', etc. This structured alignment between the presented options and the expected output format significantly improved the model's performance by reducing ambiguity.

2. **Selective Cropping**

- **Problem**: Full-page screenshots introduce noise. When analyzing attention maps, it was found that irrelevant visual elements were taking model focus and reducing accuracy.
- **Solution**: Bounding box annotations are used to dynamically crop regions of interest, isolating key elements for tasks like element OCR and action prediction. Some examples can be seen in Table 11.
- **Implementation**:
  - Bounding box ratios are calculated to crop and resize the visual input dynamically.
  - Cropped images are fed into the vision encoder, reducing distractions and enhancing task-specific precision.

Expected Impact:

- **OCR Tasks**: Cropping improves transcription accuracy by eliminating extraneous visual noise.
- **Grounding Tasks**: Prompts with exemplars and cropped inputs improve grounding performance by clarifying task requirements and focusing the model's attention.

Feature	**Version 1: Fine-Tuning**	**Version 2: Prompt Engineering and Cropping**
Dataset	MultiUI as training dataset	VisualWebBench-specific enhancements; no training
Focus	General multimodal capabilities	Task-specific optimization
Techniques	Fine-tuning, hybrid loss functions	Few-shot prompts, structured templates, data processing during inference
Tasks Targeted	All VisualWebBench tasks	Element OCR, action prediction, grounding
Expected Improvements	Broad performance gains	Targeted accuracy boosts in OCR and grounding

Table 5: Comparison of Fine-Tuning and Prompt Engineering Approaches

6 Results

Overall, the results shown in Figure 5 highlight the substantial impact on LLaVA's performance across the VisualWebBench dataset after our fine-tuning in Version 1 and prompt enhancement in Version 2. The base model, without any training on domain-specific data, achieves an average score of 13.07%, as it struggles to accuractely interpret and integrate the visual and textual components of complex webpages coupled with its inability to spatially locate UI elements. From the first checkpoint of domain specific training, the model shows slight improvements in certain areas, such as WebQA, Element OCR, Element Grounding and Action Prediction. However, its overall average remains 11.59%. There was a stark decrease in the performance of this model on Web Captioning task. While some tasks benefit from the early stages of fine-tuning, these results indicate that a more comprehensive approach is necessary to unlock LLaVA's full potential.

A significant performance leap occurs with full tuning and prompt enhancements, reflecting the importance of adapting both the model's weights and its input instructions sent to the model. After full tuning alone, the model's average score increases to 14.62, with noteworthy gains in Element OCR (22.86%) and Action Prediction (20.52%). The integration of the newly crafted prompts and instructions produces an average score of 27.89%, surpassing previous setups by a wide margin. This improvement is especially evident in Element OCR (54.82%) and action prediction (77.94%), demonstrating that enhanced prompting strategies, combined with targeted fine-tuning, significantly improve the model's ability to extract information from complex webpage layouts and predict appropriate user actions.

7 Analysis

7.1 Intrinsic Metrics

Intrinsic metrics are not the exact metrics that are used to measure the downstream task success; rather, it is the inherent skills that the model possesses. We have identified a few intrinsic metrics, such as the inference time of a model, which gives insight as to how the model would perform in real-time, the effect of increasing the number of parameters of the model and performing chain of thought reasoning before giving an answer. The intrinsic metrics are explained in more detail below.

Model parameter size The parameter size of a model is a fundamental intrinsic characteristic that impacts several aspects, including generalization ability, memory footprint, and computational efficiency. Generally, increasing the number of parameters enhances the model's capacity, allowing it to learn more complex patterns and representations. However, beyond a certain threshold, the marginal improvement might diminish, especially if the dataset does not provide enough diverse training examples. In Figures 9 and 10 we can see the effect of model performance when the sizes of Phantom (Lee et al., 2024a) and TroL (Lee et al., 2024c) models are varied. For Phantom, the 7B model has the highest scores across all the tasks, and similar trend can be seen for TroL. However, for tasks such as Heading OCR and Element Grounding, the effect of increasing size is marginal in Phantom, suggesting that model size is not a critical parameter for these tasks.

Inference time analysis Inference time is a critical metric, particularly when evaluating models for real-time applications. It measures how long the model takes to process an input and generate an output. In the context of interactive web tasks, lower inference time is preferred for better user experience. In Figures 6 and 7 we can see the effect of varying model size on inference speed. Generally, increasing the model size increases the time taken for inference as expected. There is some slight variation in Phantom between 0.5B and 1.8B, however that can be attributed to the fact that these models are based on different backbone architectures, causing deviation in the trend due to the way inputs are processed by them. In Figure 8 we can see variations between the remaining models, where LLaVA (Liu et al., 2024b) is the fastest for most tasks except Action Ground, where it takes the longest. From the Table 8 we can see that the fine-tuned LLaVA model exhibits roughly the same inference times as the original LLaVA. However, we observe an increase in inference times after applying prompt tuning, which may be attributed to the increased number of input characters.

Effect of image size On analysing the effect of varing sizes for LLaVA-7B model, we saw that the model performance improves from 224x224 resolution to 336x336 resolution for the tasks of WebQA, Web Caption and Element Grounding.

Model	Source	Website			Element		Action		Average
		Caption	WebQA	HeadOCR	OCR	Ground	Prediction	Ground	
Heavyweight Closed-Source Models									
Gemini 1.0 Pro	VisualWebBench results	25.0	55.5	**75.1**	65.4	44.3	26.7	43.7	48.0
Gemini 1.5 Pro	VisualWebBench results	31.6	69.0	54.5	76.6	**70.0**	74.4	**77.7**	64.8
Claude Sonnet	VisualWebBench results	28.9	**81.8**	70.3	**89.2**	68.8	63.4	58.3	65.8
Claude Opus	VisualWebBench results	26.7	75.4	63.7	87.1	57.7	60.4	38.8	58.5
GPT-4 Vision	VisualWebBench results	**34.5**	75.0	68.8	62.8	67.5	67.6	75.7	64.6
Light, Open-Source and Fine-Tuned Models									
InternLM	Ours	7.03	3.8	1.96	9.12	7.26	4.63	5.83	5.66
LEOPARD-LLaVA	Leopard Paper	-	-	-	-	-	-	-	24.91
LEOPARD-Idefics2	Leopard Paper	-	-	-	-	-	-	-	25.60
LLaVA 1.5-7B - No Training	Ours	15.08	2.03	33.16	6.32	10.41	9.96	14.56	13.07
LLaVA 1.5-7B - Checkpoint 1	Ours	0.85	11.01	17.55	7.32	17.43	15.30	11.65	11.59
LLaVA 1.5-7B - Version 1	Ours	8.67	3.03	13.79	22.86	19.85	20.52	13.59	14.62
LLaVA 1.5-7B - Version 2	Ours	8.67	3.03	13.79	54.82	19.85	**77.94**	13.59	27.38
TroL-1.8B	Ours	23.06	5.51	59.59	30.48	24.69	14.59	25.24	26.17
Phantom-0.5B	Ours	23.14	1.73	66.65	10.00	16.95	13.52	13.59	20.79
Phantom-1.8B	Ours	22.52	1.99	64.71	10.79	18.16	15.66	14.56	21.19
Phantom-7B	Ours	24.5	3.8	67.5	12.2	18.4	35.9	19.4	25.95
TroL-1.8B	TroL paper results	14.2	38.3	58.5	29.5	24.7	14.2	29.1	29.8
TroL-3.8B	TroL paper results	22.5	65.3	70.2	63.0	69.7	20.3	39.8	50.1
TroL-7B	TroL paper results	23.6	44.7	74.4	38.6	40.9	16.0	32.3	38.6
TroL-7B	Ours	25.44	7.02	68.27	38.45	38.49	16.37	30.09	32.01
State of the Art (Bolded Values)	All Above	**34.5**	**81.8**	**75.1**	**89.2**	**70.0**	**77.94**	**77.7**	**65.8**

Table 6: The VisualWebBench (VWB) benchmark evaluates multimodal large language models (MLLMs) across seven tasks: captioning, webpage QA, heading OCR, element OCR, element grounding, action prediction, and action grounding. **Our LLaVA fine-tuned model (*Full Tuning + Prompt Enhancements*) achieves state-of-the-art performance in the Action Prediction task with a score of 77.94 and second-best among open-source models in HeadOCR with a score of 54.82 – showing the power of our fine-tuning and prompt engineering techniques.**

Model	Caption	Website (WebQA)	HeadOCR	Element (OCR)	Element (Ground)	Action (Prediction)	Action (Ground)	Average Score
LLaVA - no training	15.08	2.03	33.16	6.32	10.41	9.96	14.56	13.07
LLaVA - checkpoint 1	0.85	11.01	17.55	7.32	17.43	15.30	11.65	11.59
LLaVA - full tuning	8.67	3.03	13.79	22.86	19.85	20.52	13.59	14.62
LLaVA - full tuning + prompt enhancements	8.67	3.03	13.79	54.82	19.85	77.94	13.59	27.38

Table 7: Performance comparison of LLaVA models under various training regimes.

Further, from 336x336 to 448x448, performance improves for Element OCR and Action grounding. After this, increasing the image resolution does not really improve the performance. From this, we can infer that very low resolution of images lower the performance, and increasing the resolution improves performance on tasks requiring fine-grained interpretation of images, like OCR and Grounding. However, after a resolution of 448x448, the information gained is marginal. We can conclude that 448x448 is the ideal image size, balancing performance and inference speed, and we consider this input size for comparing the performance of LLaVA against other models.

Capability of locating UI elements In our previous report, we have stated that the MLLM was not able to identify critical regions of an image and was only focussing on the big and the bold parts of the image. Upon finetuning a LLaVA model on the MultiUI dataset, we have seen that the MLLM is now able to identify the right regions and is not always focussing on the bold and the big text at the center. We have performed qualitative analysis on some examples. (Figures: 15, 16, and 17)

7.1.1 Qualitative Analysis – Failure Modes and Improvements Across Model Versions

Table 11 shows failure and successes among different versions of LLaVA explored and the rationale followed to improve the model iteratively. Understanding success and justifying our design was important, so a further summary is given here:

The evaluation of the LLaVA model across multiple versions reveals its evolving ability to handle multi-modal inputs in tasks involving bounding box interpretation and webpage context. This analysis aims to capture both the quantitative improvements and qualitative limitations across the versions.

Base LLaVA: The base model lacked structured reasoning about bounding box content and often deviated from the required output format. Key issues included:

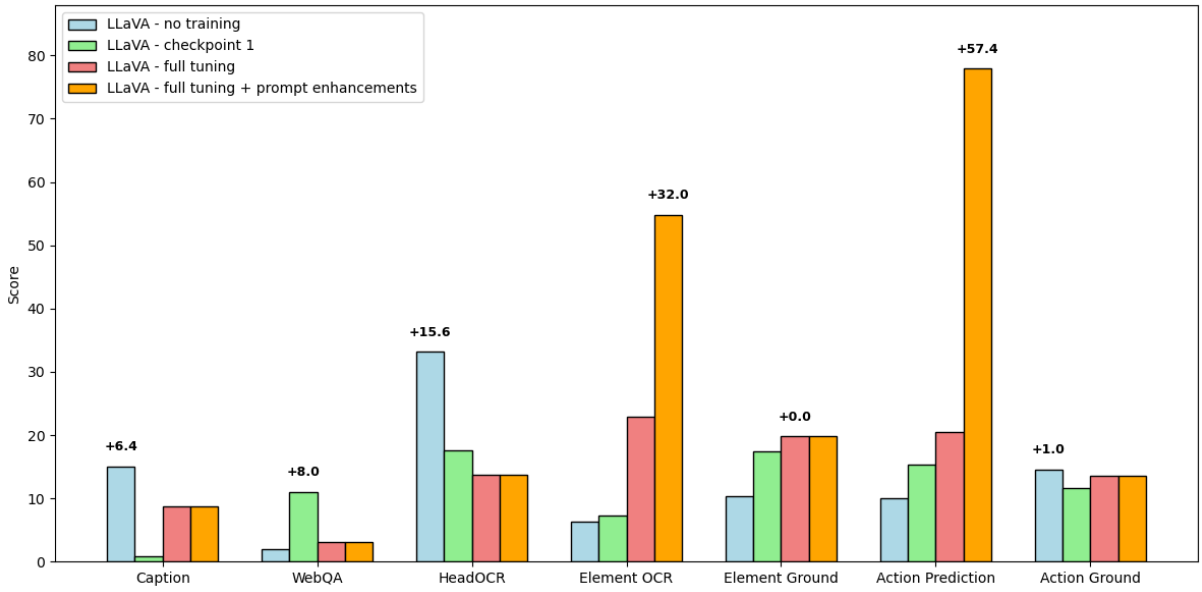


Figure 5: Performance comparison of LLaVA models under various training regimes

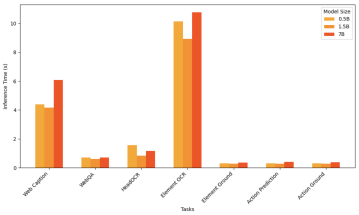


Figure 6: Phantom - Varying Sizes

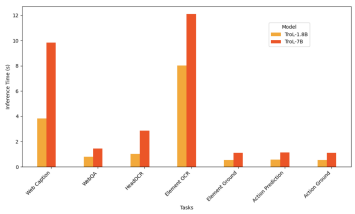


Figure 7: TroL - Varying Sizes

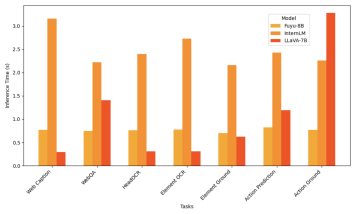


Figure 8: Fuyu, LLaVA, InternLM

- **Formatting Failures:** Despite explicit instructions, the model frequently returned verbose textual answers instead of single-character responses.

- **Global Bias Over Local Context:** Predictions favored bold or prominent text elements rather than focusing on the bounding box's content.

Fine-Tuned LLaVA: Fine-tuning introduced structured response handling (e.g., aligned to 'A, B, C' formats). However, the model's reasoning about bounding boxes remained inadequate:

- **Weak Bounding Box Integration:** While the model improved in following structured formats, its inability to link bounding box content with the task prompt persisted.

- **Webpage Comprehension Bias:** Attention maps indicated a preference for high-level webpage themes rather than granular bounding box localization.

Version 2.0 - Cropping Strategy: Cropping the bounding box improved the model's ability to focus on relevant content, but this raised questions about whether the task was being converted into an OCR exercise rather than remaining in the spirit of a multi-modal benchmark. Key observations:

- **Improved Success Rates:** Cropping succeeded in clear-cut cases, suggesting that bounding box attention alignment is critical.

- **Loss of Multi-Modal Context:** The removal of broader webpage context led to failures where bounding box content alone was ambiguous or insufficient.

- **OCR Alignment Concerns:** With bounding box cropping, the task risked becoming an OCR exercise rather than a true multi-modal challenge.

Version 2.1 - Appended Bounding Box Context: This version embraced the spirit of MMML by appending the bounding box below the full webpage image, preserving both local and global contexts. Key observations:

Model	Method	Image Size	Website			Element		Action	
			Caption	WebQA	HeadOCR	OCR	Ground	Prediction	Ground
Fuyu-8B	Normal	448x448	0.77	0.75	0.76	0.78	0.70	0.82	0.77
	CoT	448x448	-	-	-	-	0.85	0.93	1.08
InternLM	Normal	448x448	3.16	2.22	2.4	2.73	2.16	2.43	2.26
	CoT	448x448	-	-	-	-	3.5	3.73	3.39
LLaVA-7B	Normal	224x224	0.29	1.41	0.31	**0.31**	0.62	1.19	3.28
	Normal	336x336	0.29	1.4	0.31	**0.31**	0.61	1.18	3.28
	Normal	448x448	0.29	1.4	0.31	**0.31**	0.62	1.18	3.27
	Normal	672x672	**0.28**	1.4	0.3	**0.31**	0.617	1.18	3.27
	CoT	448x448	-	-	-	-	1.26	2.02	1.16
Phantom-0.5B	Normal	490x490	4.39	1.73	1.56	10.14	0.31	0.31	0.3
Phantom-1.8B	Normal	490x490	4.16	0.6	**0.83**	8.93	**0.27**	**0.28**	**0.27**
Phantom-7B	Normal	490x490	6.07	**0.72**	1.15	10.76	0.36	0.41	0.37
	CoT	490x490	-	-	-	-	7.13	7.06	7.39
TroL-1.8B	Normal	490x490	3.84	0.79	1.01	8.03	0.55	0.56	0.55
	CoT	490x490	-	-	-	-	2.87	5.28	8.72
TroL-7B	Normal	490x490	9.83	1.44	2.86	12.09	1.11	1.13	1.11
LLaVA	tuning	490x490	0.28	1.4	0.3	0.31	0.62	1.12	3.26
LLaVA	tuning and prompt enhancements)	490x490	0.46	1.44	0.94	0.89	0.62	1.13	1.16

Table 8: Inference Times

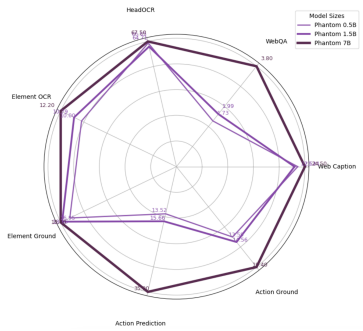


Figure 9: Phantom - Varying Sizes

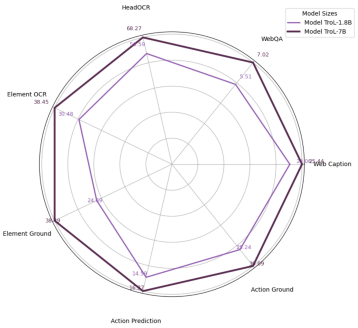


Figure 10: TroL - Varying Sizes

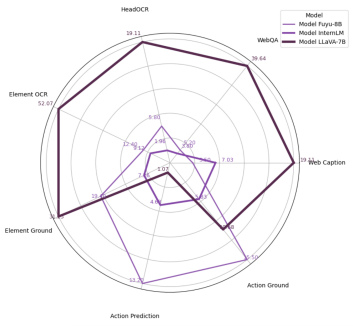


Figure 11: Fuyu, LLaVA, InternLM

- **Multi-Modal Synergy:** Combining bounding box content with full webpage context aligns with MMML's goals of holistic reasoning.

- **Improved Interpretability:** The appended bounding box acted as a natural preprocessing step, enabling the model to reason about the bounding box without explicit coordinate ratios.

- **Limitations on Ambiguity Resolution:** Ambiguous bounding boxes or poor webpage design still led to failures, emphasizing the need for enhanced attention mechanisms or robust data augmentation.

7.2 Qualitative Analysis and Examples

Tables 9, 10, and 11 show a comprehensive qualitative analysis with a few example inputs, outputs, and reasons for their failure. The base LLaVA model, v1 version of our proposed model along with v2 version of our proposed model is covered in these tables. Please refer to the Appendix to view the improvement in attention maps that we observed in our proposed model compared to the base model. We can see that the model is now focusing on the correct UI elements before answering and is able to perform relatively hard tasks as well.

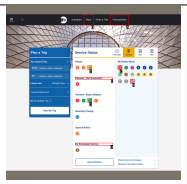
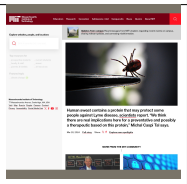
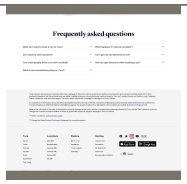
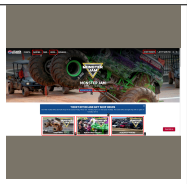
Question	Input Image	Predicted Answer	Ground Truth	Task	Reasons of Failure
In this website screenshot, I have labeled IDs for some HTML elements as candidates. Tell me which one best matches the description: No Scheduled Service. You should directly tell me your choice in a single uppercase letter, and do not output any explanation or any other contents.		A	H	Element Ground	Hallucination
What is the name of the event that Dustin Connell recently won? You should directly tell me your answer in the fewest words possible, and do not output any explanation or any other contents.		Ultra	Bass Pro Shops REDCREST 2024	WebQA	Though the model is looking at the right region, the answer is very biased as the model is paying more attention to the text that is bold in that particular region.
You are given a screenshot of a webpage. Please generate the main text within the screenshot, which can be regarded as the heading of the webpage. You should directly tell me the main content, and do not output any explanation or any other contents.		Exploring Valuable Technologies	Massachusetts Institute of Technology	HeadingOCR	In this particular datapoint, the answer is very ambiguious. The model was able to retrieve the right context of the web page, but as the answer is very different from the answer given by the model, the model failed on this datapoint.
You are given a screenshot of a webpage with a red rectangle bounding box. The [x1, y1, x2, y2] coordinates of the bounding box are [0.707 0.767 0.73 0.793]. (4 options are given along with the bounding box coordinates) and the model has to predict the best out of 4 options.		[0.707 0.767 0.73 0.793]	Turo (@turo) • Instagram photos and videos (Option 2)	Action Prediction	The model is not able to do action prediction tasks. It hallucinates and outputs irrelavent information.
You are given a screenshot of a webpage. Please generate the meta web description information of this webpage.		National Archives showcases historical coal mining	Explore the National Archives, the nation's record keeper for historical documents, photos, and records. Access veterans' service records, educator resources, and plan your visit to the museum. Stay informed with the latest Archives News and shop online for publications.	Web Caption	Focusing on the center of the image, rather than giving equal weight to all parts of the image.
In this website screenshot, I have labeled IDs for some HTML elements as candidates. Tell me which one I should click to complete the following task: view details about the Monster Jam event.		B	F	Action Ground	Hallucination
You are given a screenshot of a webpage with a red rectangle bounding box. The [x1, y1, x2, y2] coordinates of the bounding box are [0.022 0.844 0.486 0.892]. Please perform OCR in the bounding box and recognize the text content within the red bounding box.		Aerobrew Coffee Company	Take off with Aerobrew Coffee Company! Our unique air roasting method ensures the freshest and most flavorful cup of aviation coffee you'll ever have.	Element OCR	Model gave more emphasis to sthe bold text that was above the bounding box and did not follow the text prompt correctly.

Table 9: LLaVA-7B Error Analysis: In some tasks, we observed this model has given more emphasis to bold and large texts in the image and did not give a lot of importance to the text modality.

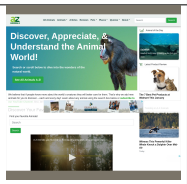
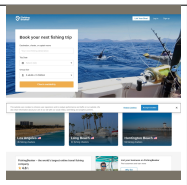
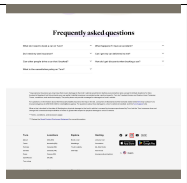
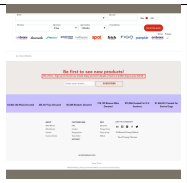
Question	Input Image	Predicted Answer	Ground Truth	Task	Reasons of Failure
In this website screenshot, I have labeled IDs for some HTML elements as candidates. Tell me which one best matches the description: A B Banjo Paterson		E	G	Element Ground	Hallucination
What is the subject of the trending article on A-Z Animals? You should directly tell me your answer in the fewest words possible, and do not output any explanation or any other contents.		The subject of the trending article on A-Z Animals is "101 Amazing Facts About the Animal World".	Killer Whale and Dolphin	WebQA	Both the answers, (Predicted as well as Ground Truth) are correct in this particular case.
You are given a screenshot of a webpage. Please generate the main text within the screenshot, which can be regarded as the heading of the webpage. You should directly tell me the main content, and do not output any explanation or any other contents.		The world's largest fishing booking platform. Over 100,000 trips worldwide. Book your next fishing trip in minutes.	Book your next fishing trip	HeadingOCR	From the attention map (Figure 12, that is attached below, it is clear that the model is able to look at the right location. This particular prompt can have multiple answers.
You are given a screenshot of a webpage with a red rectangle bounding box. The [x1, y1, x2, y2] coordinates of the bounding box is [0.707 0.767 0.73 0.793].Please select the best webpage description that matches the new webpage after clicking the selected element in the bounding box:		A	D	Action Prediction	From Figure 13, even though the model is looking at the right location, it is generating the wrong answer.
You are given a screenshot of a webpage. Please generate the meta web description information of this webpage, i.e., content attribute in ¡meta name="description" content=""¿ HTML element.		A	Explore the Internet Archive, a non-profit digital library offering free access to millions of books, movies, music, software, and over 624 billion web pages archived through the Wayback Machine.	Web Caption	Hallucination.
In this website screenshot, I have labeled IDs for some HTML elements as candidates. Tell me which one I should click to complete the following task: subscribe to the magazine.		A	G	Action Ground	Look at the Figure 14 to understand more about the reason of failure.
You are given a screenshot of a webpage with a red rectangle bounding box. The [x1, y1, x2, y2] coordinates of the bounding box is [0.205 0.388 0.795 0.415].		Be the first to see new products! Enter your email below to receive the latest news and product updates.	Be a Hero- Sign up to receive our emails today and we'll donate a meal to shelter dog on your behalf	Element OCR	Hallucination

Table 10: LLaVA-7B Finetuned Error Analysis: After finetuning, even though the model is able to look at the right locations (as seen in the attention maps), the answers generated by the models were wrong in certain examples

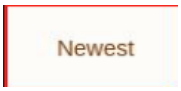
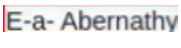
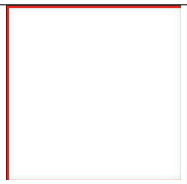
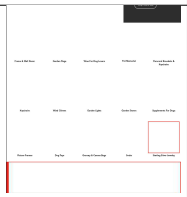
Question	Input Image	Predicted Answer	Ground Truth	Version	Analysis and Failure Reasons
You are given a screenshot of a webpage with a red rectangle bounding box. The coordinates of the bounding box are [0.384 0.613 0.476 0.666]. Please select the best webpage description.		Kneeling at the gate of hell - a poem by E-a-Abernathy - All Poetry	D (Newest Poems)	Base LLaVA	Failed to format the answer in an uppercase single letter format, and instead seemed to give a summary of the webpage.
You are given a screenshot of a webpage with a red rectangle bounding box. The coordinates of the bounding box are [0.224 0.706 0.307 0.726]. Please select the best webpage description.		A (All Poetry)	H (E-a-Abernathy - poet at allpoetry)	Base LLaVA	Failed to contextualize the bounding box content, instead defaulting to bold webpage text.
You are given a cropped screenshot of a webpage focusing on a rectangular region of interest. Please select the best webpage description.		D (Newest Poems)	D (Newest Poems)	Version 2.0	The bounding box contained relevant information, but may have turned this question closer into a heading / element OCR instead of action prediction.
You are given a cropped screenshot of a webpage focusing on a rectangular region of interest.		H (E-a-Abernathy - poet at allpoetry)	H (E-a-Abernathy - poet at allpoetry)	Version 2.0	The bounding box contained relevant information, but may have turned this question closer into a heading / element OCR instead of action prediction.
You are given a cropped screenshot of a webpage focusing on a rectangular region of interest.		G (Shop iHeartDogs, iHeartCats — Vinoshipper)	C (Dog Themed Sterling Silver Jewelry)	Version 2.0	The blank bounding box offered no visual information. Without the full webpage context, the model struggled to infer meaningful relationships.
You are given an image of a webpage with the bounding box appended to the bottom. Please select the best webpage description.		D (Newest Poems)	D (Newest Poems)	Version 2.1	Successfully reasoned about bounding box context by leveraging both local and global information.
You are given an image of a webpage with the bounding box appended to the bottom. Please select the best webpage description.		D (Newest Poems)	H (E-a-Abernathy - poet at allpoetry)	Version 2.1	The bounding box content was warped, and the guess of E-a Abernathy was unlikely without the bounding box information, as most models would not predict that is an interactable element. So, instead, it focused on the most likely element that existed on the webpage.
You are given an image of a webpage with the bounding box appended to the bottom. Please select the best webpage description.		G (Shop iHeartDogs, iHeartCats — Vinoshipper)	C (Dog Themed Sterling Silver Jewelry)	Version 2.1	Poor bounding box quality over a broken UI element resulted in failure. The model struggled to resolve ambiguity without sufficient external cues. All other poorly aligned buttons had clues related to dogs, so it likely struggled to align the box and found information about dogs on the website.

Table 11: Qualitative analysis of LLaVA model performance across multiple versions. Each example highlights the model's predicted answer, ground truth, and reasons for success or failure.

8 Future work and Limitations

Despite the progress made through targeted fine-tuning and prompt enhancements, our approach to improving MLLM performance on VisualWebBench remains constrained by several factors. First, while the incorporation of MultiUI fine-tuning and customized prompts demonstrated substantial gains in specific tasks such as Action Prediction and Element OCR, these improvements did not translate equally across all tasks. Some tasks, particularly WebQA and web captioning, continue to lag behind state-of-the-art results. This discrepancy suggests that current training regimes and prompt designs may not sufficiently capture all the nuances needed for robust semantic reasoning in complex web layouts. Future work should investigate more advanced data augmentation strategies, curriculum learning techniques, or meta-learning approaches to help models better generalize to these challenging tasks.

Additionally, our current approach relies heavily on manual prompt engineering and carefully tailored instructions. Although this method proved effective in boosting performance, it lacks scalability and may not adapt well to diverse and evolving web environments. A more automated approach to prompt optimization—potentially through reinforcement learning, evolutionary algorithms, or differentiable prompt search—could reduce human intervention and yield more consistent improvements. Another limitation lies in the limited understanding of the spatial and hierarchical structure of webpages. Although cropping and bounding box-focused strategies help, future work could explore explicit modeling of webpage layouts, semantic element hierarchies, or graph-based representations of web content. Such structures might enable the model to handle more dynamic and interactive web scenarios, from rich multimedia content to complex user interface components.

Finally, large multimodal models often remain computationally expensive and may not run efficiently in resource-constrained settings. While some architectural innovations like TroL and Phantom aim to reduce parameter footprints, further research into model compression, quantization, and efficient encoding strategies is needed. It remains crucial to balance model complexity with the ability to process high-resolution visual data and large textual contexts concurrently. Addressing these limitations collectively—through better generalization strategies, automated prompt tuning, richer structural representations, and improved efficiency—will bring us closer to building versatile, domain-adaptive MLLMs that excel across the breadth of tasks presented by VisualWebBench and beyond.

9 Ethical Concerns and Considerations

As we continue to develop MLLMs capable of navigating and interpreting the modern web, we must remain vigilant of ethical concerns and potential misuse. One primary concern is the inadvertent extraction and exposure of sensitive or personal information. Models trained on screenshots and textual data derived from real websites could be exploited to retrieve private user data, confidential corporate information, or personally identifiable content. Ensuring that future models incorporate robust de-identification techniques, strict filtering of sensitive data, and secure handling of proprietary content is imperative.

Another key ethical issue involves bias and representation. Web content can often reflect biases in language, culture, or demographics present in the broader digital ecosystem. MLLMs that learn from unmoderated web data may inadvertently replicate these biases, potentially leading to discriminatory outputs in tasks such as recommendation generation or user-interface interpretation. Continued research is needed to develop bias detection and mitigation strategies, fairness metrics, and evaluation frameworks that highlight and rectify skewed model behavior.

Moreover, the dual-use nature of advanced MLLMs raises additional concerns. Enhanced grounding, OCR capabilities, and improved navigation through websites could facilitate malicious activities, such as automated phishing, fake reviews, or tailored misinformation campaigns. As models become increasingly adept at mimicking human behavior in digital environments, preventative measures—including stricter access controls, cryptographic verification of trusted datasets, and ongoing monitoring of model behavior—should be implemented. Collaboration with policymakers, ethicists, and platform maintainers can help define responsible usage guidelines.

Finally, transparency and explainability remain crucial. Users interacting with MLLMs deserve to understand the model's decision-making process, know the provenance of its data, and hold

developers accountable for mistakes or harmful outputs. Implementing explainable AI (XAI) techniques, clear documentation, and user-friendly interfaces to display model reasoning can promote trust and responsible adoption. In sum, ethical considerations must guide every stage of MLLM development—data collection, model training, evaluation, and deployment—ensuring that these powerful tools serve society beneficially and fairly.

10 Team member contributions

Akshay Badagabettu worked on the v1 version of the proposed model.

Sai Sravan Yarlagadda worked on the v1 version of the proposed model.

Aayush Shah worked on the v2 version of the proposed model.

Nikolaj Hindsbo worked on the v2 version of the proposed model.

References

T. Baechler et al. 2024. Multimodal large language models for complex environments. Proceedings of CVPR.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024a. Internlm2 technical report. arXiv preprint arXiv:2403.17297.

Zheng Cai, Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, and et al. 2024b. Internlm2 technical report. CoRR, abs/2403.17297.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks.

M. Cheng et al. 2024. Seeclick: Grounding ui elements with html data. ECCV.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024a. Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems, 36.

Z. Deng et al. 2024b. Html-based navigation in language models. Web AI Journal.

L. Gao et al. 2024. Multitask training for gui understanding. ICML.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. A real-world webagent with planning, long context understanding, and program synthesis. arXiv preprint arXiv:2307.12856.

M. Gur et al. 2018. Early approaches in vision-language models. Vision AI Journal.

S. Hong et al. 2024. Cogagent: Cognitive models for gui navigation. Nature Machine Intelligence.

Mengzhao Jia, Wenhao Yu, Kaixin Ma, Tianqing Fang, Zhihan Zhang, Siru Ouyang, Hongming Zhang, Meng Jiang, and Dong Yu. 2024. Leopard: A vision language model for text-rich multi-image tasks.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. arXiv preprint arXiv:2401.13649.

B. Lee et al. 2023. Pix2struct: Pretraining html representations. ICLR.

Byung-Kwan Lee, Sangyun Chung, Chae Won Kim, Beomchan Park, and Yong Man Ro. 2024a. Phantom of latent for large language and vision models.

Byung-Kwan Lee, Sangyun Chung, Chae Won Kim, Beomchan Park, and Yong Man Ro. 2024b. Trol: Traversal of layers for large language and vision models. arXiv preprint arXiv:2406.12246.

Byung-Kwan Lee, Sangyun Chung, Chae Won Kim, Beomchan Park, and Yong Man Ro. 2024c. Trol: Traversal of layers for large language and vision models.

Junnan Li et al. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

X. Li and Y. Li. 2023. Multimodal large models for web interaction. ACM Transactions on Multimedia.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. Advances in neural information processing systems, 36.

J. Liu et al. 2024c. Multiui: A dataset for text-rich visual understanding. ArXiv Preprint.

A. Rahman. 2024. V-zen: Versatile agents for gui navigation. Journal of AI Research.

A. Rahman et al. 2024. Advances in vision-language models for ui understanding. Journal of Multimodal AI.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

J. Wang et al. 2024. Mobileagent: On-the-go gui automation. Mobile Computing.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. 2024. Livebench: A challenging, contamination-free llm benchmark. arXiv preprint arXiv:2406.19314.

X. You et al. 2024. Ferret: Grounding in multimodal agents. Multimodal AI Journal.

F. Zhang and H. Zhang. 2024. Auto-ui: A framework for multimodal interaction. Advances in Neural Information Processing Systems.

Y. Zhang et al. 2023. Appagent: Multimodal agents for ui interaction. IEEE Transactions on Neural Networks.

Y. Zhang et al. 2024. Ferret v2: Enhanced grounding for gui models. CVPR.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. 2023. Webarena: A realistic web environment for building autonomous agents. arXiv preprint arXiv:2307.13854.

A Appendix

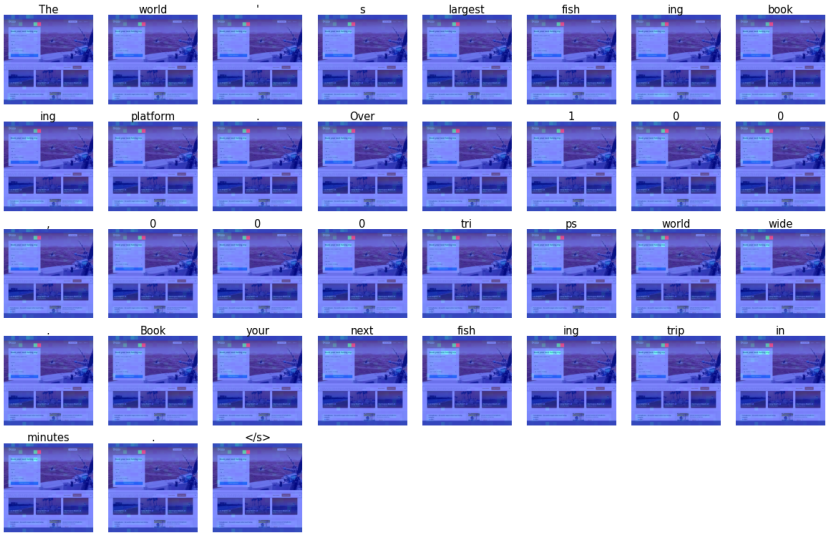


Figure 12: Table 10, row number 3: Attention maps indicating that the model is looking at the right location (HeadingOCR task)

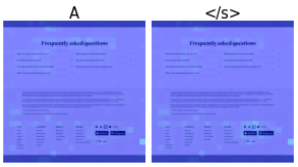


Figure 13: Table 10, row number 4: Attention maps indicating that the model is looking at the right location (Action Prediction task). The correct answer to the prompt is Instagram, and from the attention maps, we can understand that the model is paying attention to the instagram symbol.

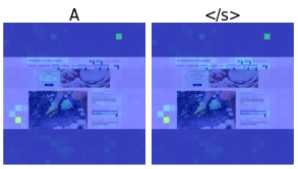


Figure 14: Table 10, row number 6: Attention maps indicating that the model is getting confused with having multiple bounding boxes on the image (Action Grounding task). Because of this reason, the model is not able to locate the correct bounding box.

B Some difficult examples for which the fine-tuned model correctly predicted the answer:

B.1 Example 1

Prompt: What additional platform is mentioned for following the website's content? You should directly tell me your answer in the fewest words possible, and do not output any explanation or any other contents.
Output: Youtube

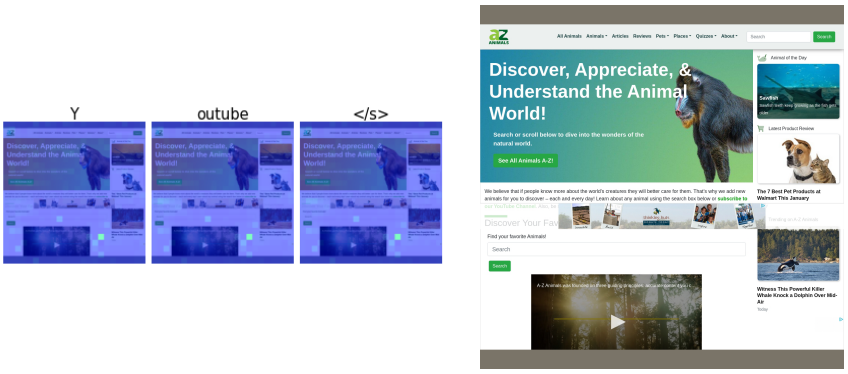


Figure 15: Attention map and corresponding image for Example 1.

B.2 Example 2

Prompt: You are given a screenshot of a webpage. Please generate the main text within the screenshot, which can be regarded as the heading of the webpage. You should directly tell me the main content, and do not output any explanation or any other contents.
Ground Truth: The Gold Standard in Online Invitations & Digital Cards
Predicted Answer: Gold Standard in Cards is a boutique card studio specializing in custom handmade cards for all occasions.

Figure 16: Attention map and corresponding image for Example 2

B.3 Example 3

Prompt: In this website screenshot, I have labeled IDs for some HTML elements as candicates. Tell me which one I should click to complete the following task: explore fishing charters in a specific location
Correct Answer: E
Predicted Answer: E

Figure 17: Attention map and corresponding image for Example 3