11-777 Report 1: Dataset Proposal and Analysis

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

1 Problem Definition and Dataset Choice

We have decided to work with the VisualWebBench (Liu et al., 2024) dataset. Our goal is to develop a model/framework to maximize the accuracies of various tasks in this dataset.

1.1 What phenomena or task does this dataset help address?

In the past couple of years, MLLMs have shown huge potential in web-related tasks, but prior to VisualWebBench, evaluating their performance in the web domain was a big challenge. There are a few famous web-related datasets such as WebArena (Zhou et al., 2023), VisualWebArena (Koh et al., 2024a), Mind2Web (Deng et al., 2024). However, these benchmarks are either designed for general multimodal tasks or they focus on end-to-end web agent tasks. This greatly limits their capability to measure the fine-grained abilities such as OCR, grounding, and semantic understanding. Measuring them is extremely important as they are the foundation for complex web-related tasks. VisualWebBench introduces a lot of granularity in measuring these abilities and hence helps in the development of a much more capable MMML model specifically for the web domain.

1.2 What about this task is fundamentally multimodal?

The various tasks in the dataset such as WebQA, action grounding, action prediction is fundamentally multimodal because it involves understanding and processing both visual and textual information of the website. There is a lot of information scattered around the webpage and it consists of images, text, and interactive elements like buttons.

To succeed in all the tasks, the model needs to extract text from images for OCR tasks, and for grounding tasks, it has to link visual elements - such as buttons, search bar, links - to their corresponding textual description or functionalities.

1.3 Hypothesis

We are proposing three ideas (more like avenues that we are interested to explore).

1.3.1 Training

We plan to train a model with the VisualWebBench dataset, particularly using Fusion or Fusion+Gating. We feel that the biggest challenge in training is that the model will overfit to the training data and will not generalize well. This is because the dataset has only 1.5k samples and 7 tasks, so the number of samples per task is even less. However, we still want to try to develop a good shared representation and implement gating so that the model can focus on one modality over the other whenever required. We were also thinking about fine-tuning using a reward model setting. Per this GitHub issue, the authors are releasing the training data in one week. Once it becomes available, we are assuming that the problem of lack of data points will be mitigated.

1.3.2 Survey to extract max performance for web-related tasks

The other avenue we are interested to explore is to conduct a comprehensive survey on the exact combination of multimodal inputs that improves the accuracy of an open/closed source MLLM to perform web-related tasks. This includes how much data, what modalities to include, performance shift when including/not including a modality. In this case, we may extend the evaluation to other web agent datasets such as WebArena or Mind2Web to test the hypothesis.

*Everyone Contributed Equally – Alphabetical order

1.3.3 Integration of set of marks and reasoning tags

We also plan to add reasoning tags to set of marks (Yang et al., 2023) segmentations. These reasoning tags provide an explanation of what the action is expected to achieve. By incorporating SoM, reasoning tags, and gating mechanisms, we aim to develop a more robust and interpretable multimodal agent for web related tasks. We believe that this will enable the model to reason more effectively about the elements it interacts with, improving its overall decision making capabilities.

1.4 Expertise

We have the following expertise in the underlying modalities required by this task:

1. **Akshay Badagabettu:** Took ANLP and IDL in Spring 2024, worked with agents before. I've mostly worked in the language domain, but have some experience working with images too.

2. **Sai Sravan Yarlagadda:** Took NLP in Spring 2024, Worked with segmentation models and have knowledge of CV.

3. **Nikolaj Hindsbo** Taken a few AI courses at CMU (most work in LSTMs, CNNs, NNs), but also some experience with transformer architecture model implementation. Worked on a chatbot with "agent-like" ability, general coding background, general mechanical engineering background.

4. **Aayush Shah:** Has work experience in multimodal large language models and took CV course in Spring 2024.

2 Dataset Analysis

2.1 Dataset properties

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.

2.2 Compute Requirements

The paper does not provide explicit details on compute requirements. However, we can infer the following:

- **Files:** With 1.5K high-resolution screenshots, the dataset 1.18 GB, so this part should not require much memory allocation.

- **Models:** The benchmark evaluates large multimodal models like GPT-4V, Claude, and various open-source models up to 34B parameters. In general, the larger models are the ones that had non-trival accuracy reports. We aim to focus on the open-source models, likely the

7B parameter ones identified such as LLaVA. These larger models would require high-end GPUs (like A100s), which we would only be able to get through AWS (or one of our research labs less likely).

2.3 Modality analysis

Tables 1,2,3 and 4 provides a detailed summary of the modalities used in VisualWebBench along with some initial data analysis.

Task	Number of datapoints
WebQA	314
Action Grounding	103
Element Grounding	413
Action Prediction	281
Element OCR	245
Heading OCR	46
Webpage captioning	134

Table 1: Number of datapoints in each task

Task	Average Bounding Boxes
WebQA	-
Action Grounding	8.0
Element Grounding	7.91
Action Predictin	1
Element OCR	1
Heading OCR	1
Webpage Captioning	-

Table 2: Task details with average number of objects detected per image

2.4 Baselines

VisualWebBench is a new multimodal evaluation benchmark. Because of the niche and recent introduction (April 2024) there are only five citations related to its paper. We have analyzed relevancy to four papers which cited VisualWebBench.

TroL: Traversal of Layers for Large Language and Vision Models

TroL is composed of a vision encoder, a vision projector, and a backbone multimodal large language model (MLLM) based on a pre-trained LLM (Lee et al., 2024). The novelty in this paper is that they have enabled the reusing of layers in a token-wise manner. This approach simulates the effect of retracing the answering stream, while increasing the number of forward propagation layers without increasing the number of layers.

TroL has been trained on 2.3M samples from a diverse dataset consisting of image/text samples, documents, charts, diagrams, symbols, and math samples. The authors have released 3 variants of the model - 1.8B, 3.8B, and 7B parameters. This architecture has been evaluated on a number of datasets including VisualWebBench where it outperformed very large open-source MMML models such as LLaVA-NeXT-34B in few of the tasks.

Tree Search for Language Model Agents

The paper (Koh et al., 2024b) talked about improving decision making capabilities of language models agents through integration of a tree search algorithm. During inference, the language model agent operates within a partially observable Markov decision framework, where it uses the tree search algorithm to evaluate and select the best action paths based on a value function. The search function explores different states, receives feedback and it also backtracks whenever necessary. The value function uses GPT-4 that helps the agent estimate the reward of different states and hence it improves the decision making over time. This approach helps in improving the agent's ability to handle environments like websites, where agents needs to understand about multi-step interactions. This proposed algorithm achieved a relative increase in the success rate of 39.7% compared to the GPT-4o agent without search when evaluated in the VisualWebArena dataset . They set a success rate of 26.4% for this task.

MMR: Evaluating Reading Ability of Large Multimodal Models

This paper introduces a novel Multi-Modal Reading (MMR) benchmark which assesses LLMs' capabilities for the task of text-rich image comprehension (Chen et al., 2024). It consists of pairs of 11 visual question answering tasks on text-rich images, which can be categorized into text recognition, spatial relationships, localization, and grounding. The authors evaluated the performance of seven open-source and five proprietary models on their benchmark, observing that although proprietary models, specifically GPT-4o and Claude 3.5 Sonnet, exhibit superior performance, the open-source models still outperform them on some benchmarks. This paper is related to VisualWebBench since web-page screenshots can be considered as text-rich images as well. Some of their insights, like which models performed better on which benchmarks, could

Task	Average Answer length	Lexical Diversity
WebQA	2.51	84%
Action Grounding	-	-
Element Grounding	-	-
Action Prediction	6.71	0.81%
Element OCR	47	40.68%
Heading OCR	7	74%
Webpage captioning	32	43%

Table 3: Average sentence length and lexical diversity

Features	WebQA	Action Grounding	Element Grounding	Action Prediction	Element OCR	Heading OCR	Webpage Captioning
id	✓	✓	✓	✓	✓	✓	✓
task_type	✓	✓	✓	✓	✓	✓	✓
website	✓	✓	✓	✓	✓	✓	✓
image	✓	✓	✓	✓	✓	✓	✓
image_size	✓	✓	✓	✓	✓	✓	✓
raw_image	✗	✓	✓	✗	✗	✗	✗
options	✗	✓	✓	✓	✗	✗	✗
instruction	✗	✓	✗	✗	✗	✗	✗
question	✓	✗	✗	✗	✗	✗	✗
bbox	✗	✗	✗	✓	✓	✓	✗
elem_desc	✗	✗	✓	✓	✓	✗	✗
answer	✓	✓	✓	✓	✓	✓	✓

Table 4: Feature availability across different tasks

be transferred to VisualWebBench and serve as an inspiration for our project.

LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models

The paper titled "LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models" is one of the few papers that expands on multimodal model evaluations by introducing their new multimodal model evaluation, LiveBench (Zhang et al., 2024). LiveBench provides a dynamic evaluation framework that assesses models' real-time generalization abilities using constantly updated data from news and online forums. This approach emphasizes low-cost, zero-contamination evaluations, highlighting the challenges in balancing comprehensive coverage with practical constraints in multimodal model assessments.

When compared to VisualWebBench, both LiveBench and VisualWebBench share the objective of testing multimodal models, but each has a different focus. LiveBench evaluates how well models handle rapidly changing, real-world information, while VisualWebBench centers around evaluating models' abilities to comprehend and interact with complex, web-based environments. Both frameworks contribute to advancing the evaluation of multimodal models by testing their adaptability and contextual reasoning in diverse, real-world scenarios.

In the context of multimodality, both LiveBench and VisualWebBench underscore the importance of evaluating models in multimodal and dynamic environments. While LiveBench provides a broader range of continuously updated data, VisualWebBench focuses specifically on granularity in web interactions, offering detailed assessments of models' abilities to navigate, interpret, and reason with web content. Together, these papers offer a (hopefully better) way to benchmark multimodal model abilities.

Set of Marks Prompting Strategy

In the paper (Yang et al., 2023) the authors worked on a new way of visual prompting that improved the visual grounding abilities of large language models. The main idea of SoM is to overlay images with boxes or masks, thus allowing the language model to reference specific regions in an image. This approach helped GPT-4V to answer fine-grained visual questions by using the marked reguibs to improve upon its reasoning and grounding capabilities. The methodology involved partitioning image into different regions using segmentation models like SAM and MaskDINO. In the paper VisualWebArena (Koh et al., 2024a) the authors mentioned that SoM improved navigability, boosting overall success rate from 15.05% to 16.37%. The authors

stated that most websites have smaller sized images that are arranged very closely and using SoM representations with strong vision language model proved critical for accurately clicking on the right button.

2.5 Metrics used

Table 5 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 5: The benchmark uses different metrics for different tasks.

3 Team member contributions

Nikolaj Hindsbo worked on and wrote Section 2 except modality analysis. Paper analysis - LMMs-Eval

Sai Sravan Yarlagadda worked on and wrote 1.3.3 and modality analysis. Paper analysis - Tree Search, Set of Marks

Akshay Badagabettu worked on and wrote Section 1 except 1.3.3. Paper analysis - TroL

Aayush Shah worked on and wrote Section 2 except modality analysis. Paper analysis - MMR

References

Jian Chen, Ruiyi Zhang, Yufan Zhou, Ryan Rossi, Jiuxiang Gu, and Changyou Chen. 2024. Mmr: Evaluating reading ability of large multimodal models.

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

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

Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. 2024b. Tree search for language model agents.

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

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. 2024. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding?

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024. Lmms-eval: Reality check on the evaluation of large multimodal models.

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*.