

# Web2Code: A Large-scale Webpage-to-Code Dataset and Evaluation Framework for Multimodal LLMs

**Sukmin Yun<sup>\*,1,4</sup>, Haokun Lin<sup>\*,1</sup>, Rusuru Thushara<sup>\*,1</sup>, Mohammad Qazim Bhat<sup>\*,1</sup>,  
Yongxin Wang<sup>\*,1</sup>, Zutao Jiang<sup>1</sup>, Mingkai Deng<sup>2</sup>, Jinhong Wang<sup>1</sup>, Tianhua Tao<sup>1,3</sup>,  
Junbo Li<sup>1</sup>, Haonan Li<sup>1</sup>, Preslav Nakov<sup>1</sup>, Timothy Baldwin<sup>1</sup>, Zhengzhong Liu<sup>1,5</sup>,  
Eric P. Xing<sup>1,2,5</sup>, Xiaodan Liang<sup>1,6</sup>, Zhiqiang Shen<sup>1</sup>**  
<sup>1</sup>MBZUAI, <sup>2</sup>CMU, <sup>3</sup>UIUC, <sup>4</sup>HYU ERICA, <sup>5</sup>Petuum, <sup>6</sup>SYSU  
<https://mbzuai-llm.github.io/webpage2code/>

## Abstract

Multimodal large language models (MLLMs) have shown impressive success across modalities such as image, video, and audio in a variety of understanding and generation tasks. However, current MLLMs are surprisingly poor at understanding webpage screenshots and generating their corresponding HTML code. To address this problem, we propose Web2Code, a benchmark consisting of a new large-scale webpage-to-code dataset for instruction tuning and an evaluation framework for the webpage understanding and HTML code translation abilities of MLLMs. For dataset construction, we leverage pretrained LLMs to enhance existing webpage-to-code datasets as well as generate a diverse pool of new webpages rendered into images. Specifically, the inputs are webpage images and instructions, while the responses are the webpage’s HTML code. We further include diverse natural language QA pairs about the webpage content in the responses to enable a more comprehensive understanding of the web content. To evaluate model performance in these tasks, we develop an evaluation framework for testing MLLMs’ abilities in webpage understanding and web-to-code generation. Extensive experiments show that our proposed dataset is beneficial not only to our proposed tasks but also in the general visual domain, while previous datasets result in worse performance. We hope our work will contribute to the development of general MLLMs suitable for web-based content generation and task automation. Our data and code are available at <https://github.com/MBZUAI-LLM/web2code>.

## 1 Introduction

Multimodal large language models (MLLMs) have achieved explosive growth in the past few years. Leveraging the rich commonsense knowledge in large language models (LLMs), MLLMs are remarkably successful at processing and reasoning about various modalities such as image [2, 35], video [60, 53], and audio [40] in a broad range of tasks such as recognition [52], reasoning [59], and question-answering [39], all using language as the intermediate representation. However, existing MLLMs are surprisingly poor at understanding webpage screenshots and generating the HTML code to express their latent states. For instance, given the instruction “Parse the HTML code for this webpage”, the well-known LLaVA-1.5 [33] generates generic, pale code that fails to preserve most of the original webpage’s features (see Figure 1), which hampers its utility in applications such as UI prototyping, automation agents, and accessibility (e.g., noting available buttons and options given webpage screenshot).

---

\*Equal Contribution.

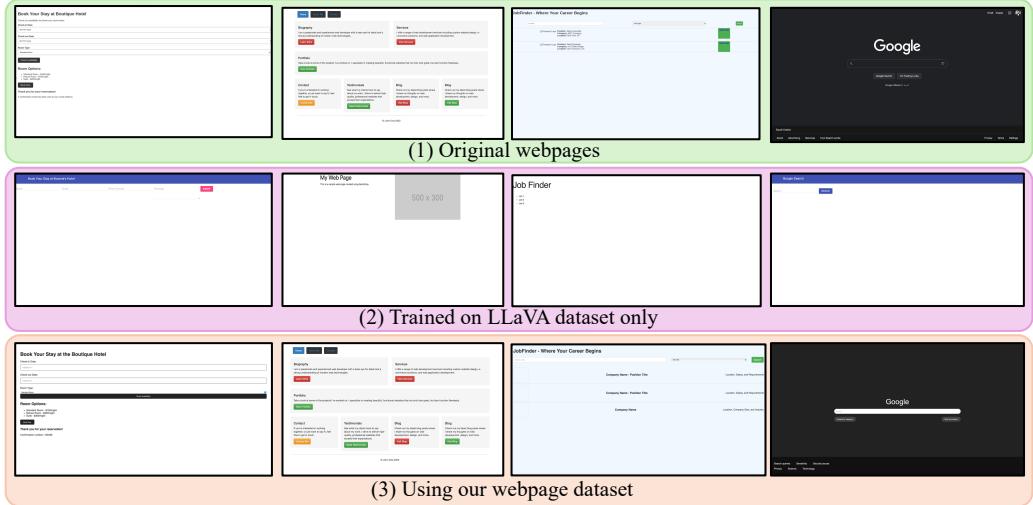


Figure 1: Our motivation for constructing the Web2Code dataset stems from the limitations of previous models, such as LLaVA [33], which are trained on general datasets and struggle to generate high-quality webpages, as in the second row. Our dataset aims to significantly enhance the quality of webpage generation as in third row while maintaining a strong level of general multimodal ability.

The essential ingredients behind the progress in MLLMs are arguably large-scale instruction datasets [9, 63] and evaluation benchmarks [16, 58] – the former for aligning multimodal inputs with the massive knowledge in LLMs [27, 35], and the latter for standardized comparison which facilitates model development. However, existing instruction datasets and benchmarks typically focus on general settings (e.g., visual QA and reasoning) and pay insufficient attention to webpage understanding and webpage-to-code generation, which requires a unique combination of capabilities such as optical character recognition (OCR), spatial reasoning, and long-text generation, among others. While previous work has developed datasets for these tasks [4, 22], they lack instruction information and are unsuitable for integration with general-purpose MLLMs. On the other hand, popular benchmarks [36, 25] evaluate some of the required capabilities in isolation, but not fully in combination for visual parsing and reasoning over webpages.

To fill this gap, we propose a new instruction tuning dataset and evaluation suite named Web2Code. Web2Code contains a total of 1179.7k webpage based instruction-response pairs. The responses consist of not only the HTML code, but also the structured questions and answers about the webpage, which assist a model in better understanding its information. For dataset collection, we use GPT-3.5 and GPT-4 to clean existing data (e.g. WebSRC [10]) as well as to generate completely new webpages in HTML code. To evaluate the MLLM’s success at webpage understanding and HTML parsing, we propose the Webpage Understanding Benchmark (WUB) and Webpage Code Generation Benchmark (WCGB), two tasks that test the model’s abilities to answer questions about a webpage and generate its HTML code, respectively. For the latter task, we find that traditional text similarity metrics are insufficient for evaluating the fidelity of the generated code, and instead propose to render the output HTML back to a webpage screenshot, and use GPT-4V [42] to evaluate the quality of the resulting webpage [61].

To demonstrate the utility of our dataset, we train LLaVA-style MLLMs with our dataset included in the instruction finetuning stage. Quantitative results show that finetuning on our dataset not only clearly improves the image-to-HTML-code translation ability of the MLLM, but also leads to slight improvements in the model’s perception and reasoning abilities in general benchmarks. In contrast, including previous datasets without our data conversion leads to degraded general performance, suggesting our dataset is suitable for inclusion in MLLM instruction-tuning for additional capabilities without impacting existing ones.

## 2 Related Work

**MLLM Dataset.** At present, there is a substantial amount of large-scale visual instruction data, primarily generated using GPT. SVIT [63] and LRV-Instruction [32] are both generated by GPT4

based on manual prompts to adjust the instruction data, including rich high-quality dialogue question and answer, complex reasoning question and answer, reference question and answer, and image detail description task datasets; similarly, ShareGPT4V [9], LLaVAR [62], LVIS-Instruct4V [55] use GPT-4V [42] to generate millions of high-quality image-text pairs, aiming to enhance the perception, reasoning and planning capabilities of MLLM. Commonly used image data sources include LAION [48], CC [49], SBU [46], COCO [31], VG [21], VQAv2 [18].

**MLLM.** Instruction Tuning MLLM has made great progress in recent years. The structure of MLLM usually contains a visual encoder, vision-language mapping module and LLM. LLaVA-v1.5 [33] only uses MLP as the vision-language mapping module and the successful application of instruction tuning on MLLM has inspired people. The community has explored various feasible structures, which can be divided into attention structures BLIP2 [27], InstructBLIP [12], Qwen-VL [3], ShareGPT4V [9] and non-attention structures LLaVA [33], Shikra [8] according to the vision-language mapping module. At the same time, various open source and more powerful LLMs, such as Vicuna1.5 [11], InternLM2 [5] also help MLLM achieve richer and more extensive instruction following capabilities. Qwen-VL [3], OtterHD [24], mPLUG-Owl [56], InternLM-XComposer2-4KHD [13] increase the resolution of images, while LLaVA-NeXT [34], Mini-Gemini [28], MM1 [41] split the input image into several image crops. In addition, BRAVE [19], MoVA [64], DeepSeek-VL [38], OmniFusion [17] apply supplementary vision encoders to obtain abundant visual features, e.g. DINOv2 [43], SAM [20]. Furthermore, more computer vision models are utilized for different tasks, which include image segmentation, detection and OCR, in MOAI [23], CuMo [26] and SpatialVLM [7]. Subsequently, MoE [30] was applied to MLLM to expand the scale of training data at the same computing scale.

**Code Study.** There are various code studies related to LLM. Sarker et al. [47] focused on generating code functions through formula hints, trying to improve the robustness of the syntax, and systematically tested the robustness of the syntax. From the perspective of security, Finkman et al. [15] said that these code assistance tools may inadvertently disclose the developer’s proprietary code to the code assistant service provider in the process of helping development, thus they proposed a complementary method to reduce the risk of code leakage while providing effective advice to developers. In addition to code leakage, the code generated by LLM has also caused concerns in industries and other fields. To address this issue, Ye et al. [57] proposed a new zero-shot synthetic code detector based on the similarity between code and its rewritten variants. In the evaluation work on code generation, Du et al. [14] proposed a new computational efficiency benchmark Mercury and a new metric Beyond for the efficiency evaluation of code. They experimentally showed that direct preference optimization can be used as a robust baseline for improving computational efficiency compared to supervised fine-tuning, which paves a promising path for future exploration of efficient code generation.

### 3 Dataset Construction

**Overview.** Our Web2Code instruction tuning dataset construction and instruction generation process involves four key components: **(1)** Creation of new webpage image-code pair data: We generated high-quality HTML webpage-code pairs following the CodeAlpaca prompt [6] using GPT-3.5 and convert them into instruction-following data. **(2)** Refinement of existing webpage code generation data: We transform existing datasets including WebSight [22] and Pix2Code [4] into an instruction-following data format similar to LLaVA data [33], so they can be used as instruction-following data to train MLLMs. **(3)** Creation of a new text question-answer pair data: We generated a new question-answer pair dataset utilizing our new GPT-3.5 generated data from (1) for webpage understanding. **(4)** Refinement of existing webpage understanding data: We refine the WebSRC [10] question-answer data to improve its quality using the GPT-4. Each component is elaborated in detail as follows:

**DWCG: Creation of new webpage image-code pair data for code generation.** To augment our dataset with high-quality data, we employed GPT-3.5 to generate 60K HTML pages following the guidelines and prompts in CodeAlpaca [6]. Using Selenium WebDriver, we then created web image screenshots from the generated HTML code. These web image-code pairs were subsequently converted into an instruction-following data format similar to the LLaVA data format [33], enabling their use in training Multimodal Large Language Models (MLLMs). The example of the instruction is shown in Figure 16. The generation of instruction was done in two stages using prompts fed to GPT-4: (a) During stage 1, the prompt shown in Figure 12 resulted in the creation of generic instructions. (b) This was followed by augmenting the instruction from (a) with the GPT generated instructions using

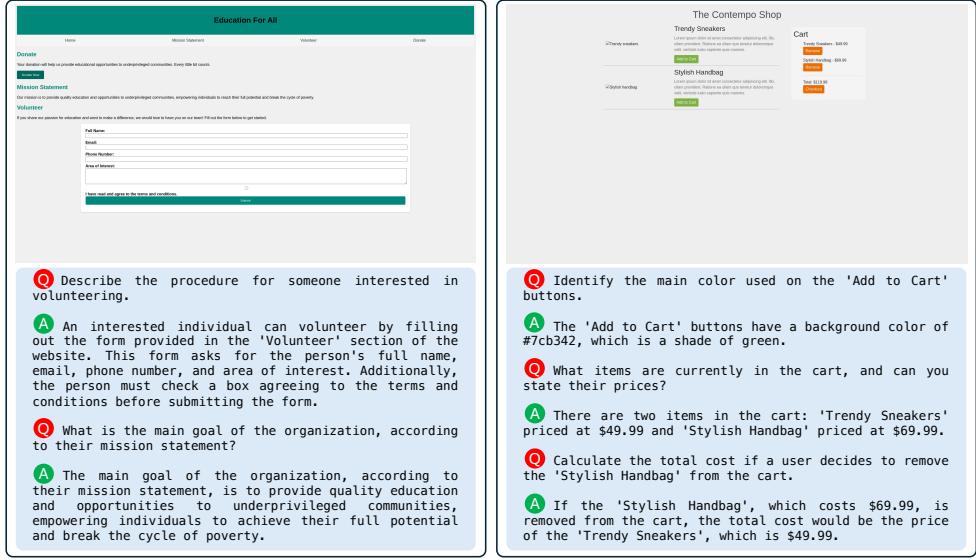


Figure 2: Qualitative example of generated question-answer pair dataset. Questions cover diverse aspects of the web page understanding.

<p>Question: Where will I be working if I choose Unlimited Commission – Work from Home? Answer: Western Cape</p> <p>Question: What is the location of Unlimited Commission – Work from Home? Answer: Western Cape</p> <p>Question: Where is the job Sales Representative located? Answer: Johannesburg</p> <p>Question: Where will I be working if I choose Sales Representative? Answer: Johannesburg</p> <p>Question: Where is the job Work from home – Call centre agent / lead warmer located? Answer: Gauteng</p> <p>Question: What is the location of Work from home – Call centre agent / lead warmer? Answer: Gauteng</p> <p>Question: When was the job Sales Representative posted? Answer: 21 Oct 2020</p> <p>Question: When was the job offer Sales Representative posted on the internet? Answer: 21 Oct 2020</p>	<p>Question: Where will I be working if I choose Unlimited Commission – Work from Home? Answer: Western Cape</p> <p>Question: What is the location of Unlimited Commission – Work from Home? Answer: The Unlimited Commission – Work from Home job is located in Western Cape.</p> <p>Question: Where is the job Sales Representative located? Answer: The Sales Representative job is based in Johannesburg.</p> <p>Question: Where will I be working if I choose Sales Representative? Answer: Johannesburg</p> <p>Question: Where is the job Work from home – Call centre agent / lead warmer located? Answer: The job is located in Gauteng.</p> <p>Question: What is the location of Work from home – Call centre agent / lead warmer? Answer: Gauteng</p> <p>Question: When was the job Sales Representative posted? Answer: The Sales Representative job was posted on 21 Oct 2020.</p> <p>Question: When was the job offer Sales Representative posted on the internet? Answer: 21 Oct 2020</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 3: WebSRC data refinement for improved Quality. **Left:** Before refinement; **Right:** After refinement, the quality has been improved and duplications have been excluded.

the prompt shown in Figure 13 to include stylistic information. This allowed the model to learn two styles: Modern and Bootstrap style as shown in Figure 21 and Figure 22, respectively.

**DWCG<sub>R</sub>: Refinement of existing webpage code generation data.** To enhance the capability of our model in the task of HTML code generation, we leverage the Pix2code [4] and WebSight [22] datasets. To mitigate the detrimental impact on model performance from random letters in Pix2Code data, we replace these random letters with meaningful text using GPT-4, thereby refining the webpages into diverse webpages encompassing product landing pages, personal portfolios, blogs, and other categories. We then visually rendered each sample by taking screenshots of the browser view of each webpage. Further, we convert all these data into LLaVA instruction following data format using the same strategy as used for DWCG. We note that DWCG and WebSight webpages follow Modern style while Pix2Code follow Bootstrap style.

**DWU: Creation of a new question-answer pair data for webpage understanding.** For the purpose of fine-tuning our models through an instruction-following manner, we utilized the capabilities of GPT-4 to generate webpage code-based question-answer pairs. We generated 10 question-answer pairs using GPT-4 for a subset of 24.35K webpage data, resulting in a total of 243.5K question-answer

data points. This includes, a set of 230K question-answer pairs for GPT-3.5 based webpages, a set of 13.5K newly generated question answer pairs for refined Pix2Code images. These pairs were meticulously crafted to align with our image-based evaluation criteria, ensuring that each question probes specific aspects of the visual and content quality reflected in the generated web images. This strategy enhances the model’s performance by integrating a nuanced understanding of the evaluation parameters into its learning process. Figure 2 shows a qualitative example from our generated question-answer pairs. The prompt used for question-answer generation is shown in Figure 11.

**DWU<sub>R</sub>: Refinement of existing webpage understanding data.** To increase our instruction-following dataset with high-quality instruction-following examples for webpages, we integrated the WebSRC dataset into our training regime. Prior to inclusion, we meticulously filtered the existing question-and-answer pairs from the WebSRC dataset to ensure relevance and quality. This involves duplication removal and quality optimization as shown in Figure 3. Specifically, we found that WebSRC data contains several questions related to the same answer. To this end, we first removed those duplicates and then employed GPT-4 to assess and enhance the quality of answers. This process not only refined the dataset into 51.5K high-quality instruction data but also ensured that the model’s training was influenced by high-fidelity, instructionally sound data, thereby improving its ability to follow complex web-based instructions.

### 3.1 Statistics and Analysis

Figure 4 shows the word cloud of the answer set of our question-answer dataset. The word cloud highlights the most frequently occurring terms, with "section," "color", "button", and "website" being the most prominent, indicating a strong emphasis on structural and design elements in the data. This reflects the detailed focus on the layout and visual aspects of the dataset.

Figure 5 illustrates the distribution of the most common HTML tags in our GPT-3.5 generated HTML data. The distribution shows a high frequency of essential structural tags such as `<div>`, `<p>`, `<meta>`, `<img>`, and `<a>`, indicating that the generated pages include a diverse range of elements necessary for rich and varied web content. The significant presence of `<h2>`, `<input>`, `<html>`, `<head>`, and `<body>` tags further reinforces the completeness and structural integrity of the generated HTML documents.

To estimate the difficulty levels of our HTML-based webpage dataset, we provide several quantitative measures and compare them with recent and similar existing datasets, namely WebSight [22], Design2Code [50], and Pix2Code [4] (See Table 1).

Design2Code is primarily used for testing and has a small size of 484 examples, limiting its versatility and robustness. In contrast, our dataset, intended for both training and testing, is significantly larger (884.7K examples) and more complex, making it more suitable for developing robust models. Overall, our benchmark examples are more challenging and cover a broader spectrum of complexities compared to prior efforts such as WebSight.

Dataset	WebSight [22]	Design2Code [50]	Pix2Code [4]	DWCG (ours)	DWCG <sub>R</sub> (ours)
<b>Instruction Source</b>	-	-	-	✓	✓
<b>Size</b>	Synthetic 823K	Real-World 484	Synthetic 1.7K	Synthetic 60K	Synthetic 824.7K
<b>Avg Length (tokens)</b>	647±216	31216±23902	658.7±98.0	471.8±162.3	652.85±157.0
<b>Avg Tag Count</b>	19±8	158±100	51.6±8.0	28.1±10.6	35.3±9.0
<b>Avg DOM Depth</b>	5±1	13±5	8.0±0.0	5.3±1.0	6.5±1.0
<b>Avg Unique Tags</b>	10±3	22±6	17.0±0.0	13.6±2.7	13.5±2.5

Table 1: Comparison of dataset statistics among webpage code generation datasets: WebSight, Design2Code, Pix2Code, our DWCG, and our DWCG<sub>R</sub>. DWCG is a newly generated GPT-3.5-based dataset, while DWCG<sub>R</sub> is the refined dataset that utilizes WebSight and Pix2Code datasets.

### 3.2 Distribution

Our instruction-following dataset contains 1,179.7K instruction data points. This includes 884.7K website image-code pairs and 295K question-answer pairs.

The 295K question-answer pairs consist of 243.5K GPT-4 based question-answer pairs (DWU Data) and 51.5K question-answer pairs from WebSRC image-based data, as shown in Table 2. Our evaluation dataset comprises 1,198 webpage screenshot

Dataset	DWU	DWU <sub>R</sub>
<b>Instruction Size</b>	✓ 243.5K	✓ 51.5K

Table 2: Distribution of DWU and DWU<sub>R</sub> datasets. Both datasets include high-quality question-answer pairs for webpage understanding.



Figure 4: Word Cloud for the answer set of the GPT4 based DWU dataset.

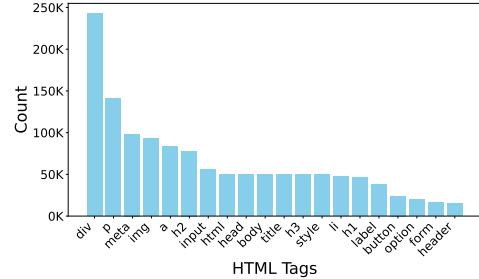


Figure 5: The distribution of most common 20 tags in GPT-3.5 based HTML data.

images, sourced from diverse origins, including WebSight, Pix2Code, GPT-3.5 based data, and manual processes, ensuring a broad representation of web content. Additionally, we utilize 5,990 "yes" / "no" question-answer pairs generated from the GPT-4 Vision API for our Webpage Understanding Benchmark, as shown in Section 4.1.

#### 4 A New Evaluation Framework for Webpage

Our proposed evaluation framework includes two schemes: (1) **Webpage Understanding Benchmark (WUB)**: An offline evaluation using "yes" / "no" questions. (2) **Webpage Code Generation Benchmark (WCGB)**: An online evaluation (using GPT-4 Vision) based on image similarity.

#### 4.1 Evaluation Metric for HTML Code Generation

In the realm of assessing code quality, particularly in terms of final visual appeal and overall functionality, existing methods that rely on code similarity metrics fall short. These traditional approaches often lack the precision and reliability needed for nuanced evaluations of code effectiveness. To address these shortcomings, we have developed a novel approach: regenerating the webpage using the model's predicted HTML code and capturing screenshots of these generated webpages. This process, automated using the Selenium WebDriver extension in Python, shifts the focus from the less reliable code similarity assessments to a more accurate and visually oriented method. By comparing images of the generated webpages, we can more effectively evaluate the aesthetic and functional aspects of the code, offering a more comprehensive understanding of its quality.

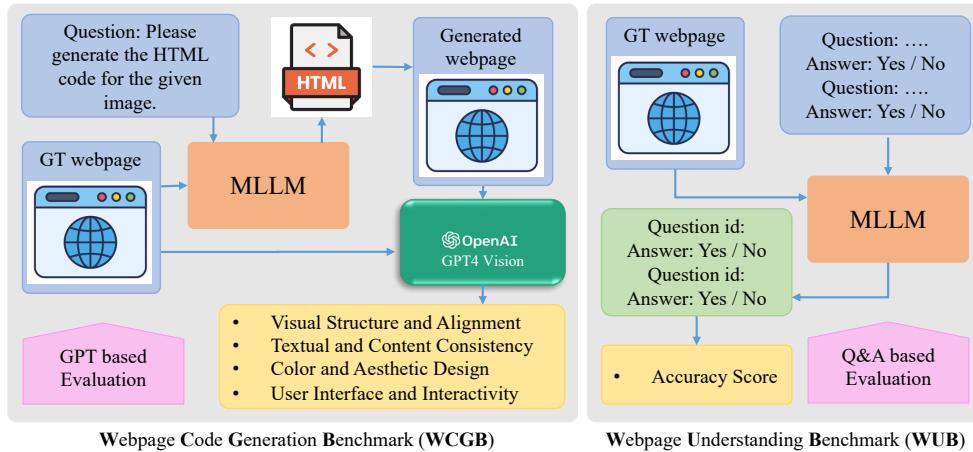


Figure 6: Evaluation benchmark for webpage generation and webpage understanding. **Left:** WCGB utilizes GPT4 Vision based online evaluation for image level comparison; **Right:** WUB employs an offline evaluation based on question-answer pairs.

We propose two benchmarks for assessing webpage understanding and code generation capabilities.

**WUB:** This benchmark comprises 5,990 high-quality question-answer pairs generated from GPT-4 Vision API (See prompt 15), based on 1,198 webpage screenshot images, where each answer is

either "yes" or "no". These images are sourced from diverse data origins, including WebSight, Pix2Code, GPT-3.5, and manual processes, ensuring a broad representation of web content. Figure 10 shows a qualitative sample data we used for WUB. We test these pairs on various multimodal image understanding models by comparing the predicted answers to the ground truth, with the final accuracy score serving as the evaluation metric as depicted on the right side of Figure 6. Qualitative data examples in our WUB benchmark are shown in Figure 10.

**WCGB:** Utilizing the same images as the WUB, this benchmark evaluates a multimodal model tasked with generating HTML code from webpage images based on specific instructions. Unlike traditional code-level evaluations, this benchmark assesses the generated webpage's fidelity at the image level. We convert the predicted HTML codes back into images using Selenium WebDriver to allow a direct visual comparison with the ground truth images. The evaluation, depicted on the left side of Figure 6, considers 10 different aspects, which are further categorized into four evaluation matrices using the GPT-4 Vision API. This image-level evaluation provides a more accurate measure of the model's code generation capabilities, acknowledging that identical webpages can be constructed from varying codes. The prompt used for evaluation is shown in Figure 14. This framework consists of 10 distinct criteria, which we group into four categories, each encompassing specific criteria that are scored on a 0-10 scale, as shown in Section D of Appendix.

## 4.2 Quantitative Evaluation for HTML Code Generation of MLLMs

We have evaluated the trained models using various data configurations and backbones on our WUB and WCGB benchmarks. The performance of the models on the code generation benchmark is presented in Table 3, while the results for webpage understanding are shown in Table 4.

LLM Backbone	DWCG	DWU	DWCG <sub>R</sub>	DWU <sub>R</sub>	VSA ↑	CAD ↑	TCC ↑	UII ↑	Overall ↑
CystalCoder-7B [37]	-	-	-	-	3.832	3.678	3.411	3.992	3.728
	✓	-	-	-	7.812	7.899	8.138	8.112	7.990
	✓	✓	-	-	8.010	<b>8.102</b>	8.266	8.124	8.126
	✓	✓	✓	✓	<b>8.023</b>	8.071	<b>8.321</b>	<b>8.334</b>	<b>8.187</b>
CrystalChat-7B [37]	-	-	-	-	4.714	4.572	4.865	5.147	4.825
	✓	✓	-	-	7.900	8.001	8.204	8.215	8.080
	✓	✓	✓	✓	<b>8.384</b>	<b>8.287</b>	<b>8.417</b>	<b>8.488</b>	<b>8.394</b>
Vicuna1.5-7B [11]	-	-	-	-	3.042	3.250	3.333	3.167	3.198
	✓	✓	✓	✓	<b>7.876</b>	<b>7.687</b>	<b>7.267</b>	<b>7.563</b>	<b>7.598</b>
LLaMA3-8B [1]	✓	✓	✓	✓	<b>8.522</b>	<b>8.564</b>	<b>8.421</b>	<b>8.611</b>	<b>8.530</b>

Table 3: The performance of different LLM backbones under various data configurations on our Webpage Code Generation Benchmark (WCGB). "VSA" denotes Visual Structure and Alignment, "CAD" represents Color and Aesthetic Design, "TCC" represents Textual and Content Consistency, and "UII" denotes User Interface and Interactivity.

LLM Backbone	DWCG	DWU	DWCG <sub>R</sub>	DWU <sub>R</sub>	Accuracy (%)
CrystalCoder-7B [37]	-	-	-	-	73.54
	✓	-	-	-	71.81
	✓	✓	-	-	<b>73.74</b>
	✓	✓	✓	✓	73.61
CrystalChat-7B [37]	-	-	-	-	73.94
	✓	✓	-	-	73.48
	✓	✓	✓	✓	<b>74.14</b>
Vicuna1.5-7B [11]	-	-	-	-	71.12
	✓	✓	✓	✓	<b>71.23</b>
LLaMA3-8B [1]	✓	✓	✓	✓	<b>74.84</b>

Table 4: The accuracy of webpage understanding under various data configurations and LLM backbones. All models are instruction-tuned and evaluated on our WUB benchmark. We note that the general domain data (i.e., LLava) is included in all data configuration as default.

To be specific, our dataset components have an orthogonal contribution to the overall improvements on both the WUB and the WCGB benchmarks. Table 3 demonstrates improvements in webpage

code generation quality when adding DWU, DWCG + DWU, and DWCG<sub>R</sub> + DWU<sub>R</sub>. For example, the results based on instruction-tuned CrystalChat show step-wise improvements on the WCGB benchmark when adding DWCG + DWU, and DWCG<sub>R</sub> + DWU<sub>R</sub> ( $4.825 \rightarrow 8.080 \rightarrow 8.530$  on the overall metric). Interestingly, the instruction-tuned CrystalChat-7B with the general domain data only shows poor WCGB performances while it shows on par WUB performance compared to the use of further webpage datasets. Similar trends are also found in the Vicuna1.5-7B backbone, while adding the proposed dataset shows significant improvements, achieving an overall score of 7.598.

Similarly, Table 4 further demonstrates the effectiveness of the proposed dataset on the webpage comprehension capabilities. For example, the four rows of the instruction-tuned CrystalCoder-7B show that not only using DWCG with DWU can prevent degradation in language understanding, but also the full components even enhance the webpage comprehension capabilities. Overall, we found the proposed dataset can enhance both webpage understanding capability and webpage code generation abilities under various LLM backbones, and LLaMA3-8B achieves the best performance among all on both webpage code generation and webpage understanding.

### 4.3 Visualizations for Qualitative Evaluation

As shown in Figure 7, we compare the results between the original image which is the real-world webpage sample, the rendered image generated by using LLM backbones of Vicuna1.5-7B and CrystalChat-7B, respectively. CrystalChat-7B is a code-enhanced LLM and our visualization demonstrates that it achieves the better quality of generation than Vicuna1.5-7B even the performance is slightly worse on general multimodal domain, as presented in Table 6. Moreover, as in Figure 8, our rendered webpage from the model trained on our web dataset closely resembles the original image, indicating the positive impact of the web2code dataset. We further visualize our generation in Figure 9 when the input is a hand-drawn webpage to examine the adaptation ability of our model.

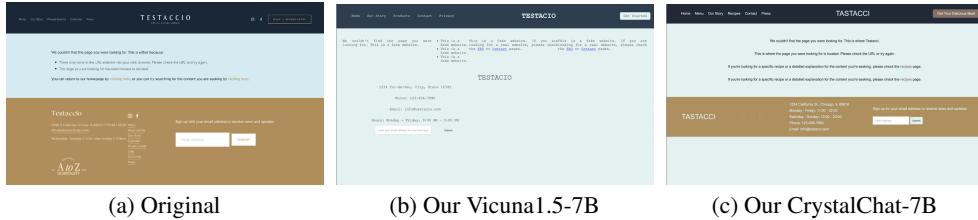


Figure 7: Visualization comparison using different backbones. Using the code-enhanced LLM backbone CrystalChat-7B achieves better quality of generation than Vicuna1.5-7B.



Figure 8: Visualization comparison between ground-truth code generated image and our result. The style and layout of the generated webpage image are similar to the ground-truth image.

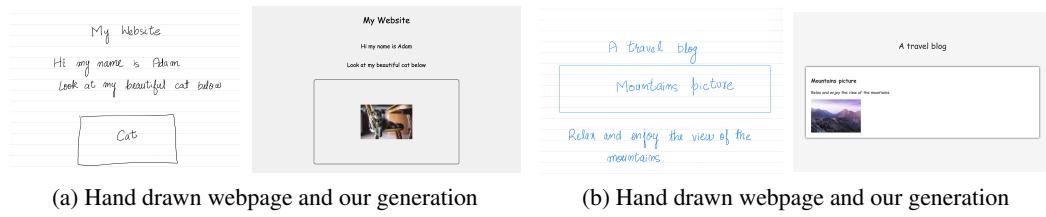


Figure 9: Visualization of our CrystalChat-7B generation when the input is a hand-drawn webpage.

## 5 General Evaluation of MLLMs Using Web2Code

**Setup and Overview.** Our model training framework mainly follows the design of LLaVA-1.5 [33] where we leverage the capabilities of both a pre-trained visual encoder, an LLM and a projector to connect visual features into the word embedding space. The model consists of (1) a pre-trained CLIP ViT-L/14 [44] visual encoder with a resolution of  $336 \times 336$  and a patch size of 14, which has good feature representation already aligned with the text embedding space. (2) As for the LLM backbones, we leverage CrystalChat [37] as the base model and compare it with other latest LLM backbones like Vicuna1.5 [11], LLaMA2 [54], LLaMA3 [1] and CrystalCoder [37]<sup>2</sup>. Training details and hyperparameters are presented in the Appendix A.

**General Evaluation Metrics for MLLMs.** MME [16] serves as an extensive evaluative benchmark, aiming to assess the perceptual and cognitive capability of MLLMs within 14 sub-tasks. Additionally, we also evaluate the performance of our models on text-oriented visual question-answering tasks employing a diverse set of benchmark datasets including ScienceQA [39] and TextVQA [51]. Furthermore, We assess our models’ ability toward anti-hallucination through POPE [29].

LLM Backbone	DWU	DWCG	DWU <sub>R</sub>	DWCG <sub>R</sub>	MME-P	MME-C	POPE	SciQA	TextVQA
CrystalChat-7B [37]	-	-	-	-	1456.53	<b>308.21</b>	86.86	67.77	57.84
	✓	-	-	-	1438.51	292.14	<b>87.10</b>	68.27	<b>58.15</b>
	✓	✓	-	-	<b>1478.82</b>	297.14	86.13	67.92	57.41
	✓	✓	✓	✓	1449.54	279.64	86.53	<b>68.32</b>	57.86

Table 5: Component analysis on CrystalChat-7B backbone under various data configurations. We note that the general domain data (i.e., LLaVA) is included in all data configuration as default.

**Effects of Web2Code on General Domain.** Here, we first perform instruction tuning using Web2Code on various LLM backbones and then we evaluate those MLLMs on the general domain of visual language understanding. Throughout extensive experiments under various data configurations, we observed that the proposed dataset Web2Code can be incorporated with the conventional visual language instruction tuning dataset of LLaVA [33] without harming performances on the general domain. Table 5 summarizes the results<sup>3</sup>. Specifically, both proposed Web Understanding data (DWU or DWU<sub>R</sub>) and Web Code Generation data (DWCG or DWCG<sub>R</sub>) do not hurt or even can be beneficial to the visual language understanding. For example, we observed that adding DWU to CrystalChat achieves comparable or even better performances on POPE (86.86→87.10), SciQA (67.77→68.27), and TextVQA (57.84→58.15). Somewhat surprisingly, we further found that adding DWCG can even improve visual language understanding. For example, the second and third rows of CrystalChat show +40.31 and +5.00 points higher improvements in MME-P and MME-C benchmarks, respectively. Moreover, adding refined data DWU<sub>R</sub> and DWCG<sub>R</sub> are still effective in the visual language domain, by achieving comparable (or even better) performances on overall benchmarks. For example, the last row indicates that adding DWU<sub>R</sub> and DWCG<sub>R</sub> preserves comparable performances on overall benchmarks and even achieves +0.4 higher points on the SciQA benchmark.

## 6 Conclusion

We have presented Web2Code, a benchmark that consists of a high-quality, large-scale webpage-to-code instruction tuning dataset containing 1179.7k entries and an evaluation suite for the webpage understanding and webpage-to-HTML translation abilities of MLLMs. Through extensive experiments, we have demonstrated that our proposed dataset is clearly effective at enhancing these abilities of MLLMs as well as general visual proficiency, while existing datasets lead to inferior performance. We hope our work will attract the community’s attention and facilitate progress toward foundation models serving as virtual assistants for content generation and task automation.

**Limitations and Ethics Statement.** The Web2Code project provides a comprehensive dataset and evaluation framework for fine-grained multimodal large language models. This can significantly enhance the capabilities of LLMs in understanding and generating web code from instructions, leading to advancements in web development automation, improved coding assistance tools and

<sup>2</sup>CrystalCoder [37] and CrystalChat [37] are an open-source code LLM pre-trained and instruction-tuned models, respectively. They are trained on publicly available language and code datasets.

<sup>3</sup>We observe that the conventional visual language domain data (i.e., LLaVA) is a crucial component for visual language understanding, i.e., instruction-tuned MLLMs without the general domain data are weak.

platforms. By enabling more accurate and context-aware code generation, it can boost productivity for developers and make coding more accessible to beginners. However, the primary limitations of the Web2Code include the potential for biased dataset that may not cover all possible HTML coding scenarios, potentially leading to gaps in model performance, some webpages that include humans may be privacy sensitive. Ensuring high-quality annotations and comprehensive coverage of all possible HTML and code structures is challenging. Also, handling complex, real-world HTML and code scenarios might still be beyond the current capabilities of the models trained on this dataset. Moreover, the proposed evaluation framework may not capture all aspects of the code generation quality, such as code efficiency, readability, or adherence to best practices.

## Acknowledgements

We are thankful to Veselin Stoyanov for discussions on multimodal large language models. We also thank the support provided by the MBZUAI IT/corporate service teams (Ian Mathews, John Murphy, Padma Pavani, Tapas Sen, Walid Omari) and CIAI engineering team (Guowei He, Yun Xu, Yue Peng) for organizing High Performance Computing resources and services. Z.S. and E.X. would like to thank the MBZUAI-WIS Joint Program for AI Research. Z.S. also would like to thank the Google Research award grant for the support.

## References

- [1] AI@Meta. Llama 3 model card. 2024. URL [https://github.com/meta-llama/llama3/blob/main/MODEL\\_CARD.md](https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md).
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.
- [4] Tony Beltramelli. pix2code: Generating code from a graphical user interface screenshot. *arXiv preprint arXiv:1705.07962*, 2017.
- [5] Zheng Cai, Maosong Cao, Haojong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.
- [6] Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation. <https://github.com/sahil280114/codealpaca>, 2023.
- [7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024. URL <https://arxiv.org/abs/2401.12168>.
- [8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic, 2023.
- [9] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- [10] Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*, 2021.
- [11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.

- [12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 36, 2024.
- [13] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, and et al Bin Wang. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd, 2024.
- [14] Mingzhe Du, Anh Tuan Luu, Bin Ji, and See-Kiong Ng. Mercury: An efficiency benchmark for llm code synthesis. *arXiv preprint arXiv:2402.07844*, 2024.
- [15] Amit Finkman, Eden Bar-Kochva, Avishag Shapira, Dudu Mimran, Yuval Elovici, and Asaf Shabtai. Codecloak: A method for evaluating and mitigating code leakage by llm code assistants. *arXiv preprint arXiv:2404.09066*, 2024.
- [16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2023.
- [17] Elizaveta Goncharova, Anton Razzhigaev, Matvey Mikhalkuk, Maxim Kurkin, Irina Abdullaeva, Matvey Skripkin, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. Omnifusion technical report, 2024.
- [18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017.
- [19] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models, 2024.
- [20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.
- [21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.
- [22] Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024.
- [23] Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Moai: Mixture of all intelligence for large language and vision models, 2024.
- [24] Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multi-modality model, 2023.
- [25] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [26] Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts, 2024.
- [27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [28] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models, 2024.
- [29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023.

- [30] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.
- [31] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.
- [32] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.
- [33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [34] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-v1.github.io/blog/2024-01-30-llava-next/>.
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [36] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
- [37] Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. Llm360: Towards fully transparent open-source llms, 2023.
- [38] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhushu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.
- [39] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- [40] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.
- [41] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, and et al. Mm1: Methods, analysis & insights from multimodal llm pre-training, 2024.
- [42] OpenAI. Gpt-4 technical report, 2024.
- [43] Maxime Oquab, Timothée Dariset, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

- [45] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémie Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- [46] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.
- [47] Laboni Sarker, Mara Downing, Achintya Desai, and Tevfik Bultan. Syntactic robustness for llm-based code generation. *arXiv preprint arXiv:2404.01535*, 2024.
- [48] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarezyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [49] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soriciu. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [50] Chenglei Si, Yanzhe Zhang, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: How far are we from automating front-end engineering? *arXiv preprint arXiv:2403.03163*, 2024.
- [51] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020.
- [52] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
- [53] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023.
- [54] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [55] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.
- [56] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023.
- [57] Tong Ye, Yangkai Du, Tengfei Ma, Lingfei Wu, Xuhong Zhang, Shouling Ji, and Wenhui Wang. Uncovering llm-generated code: A zero-shot synthetic code detector via code rewriting. *arXiv preprint arXiv:2405.16133*, 2024.
- [58] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.
- [59] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.
- [60] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023.
- [61] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

- [62] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.
- [63] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023.
- [64] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.

## Appendix

### A Training Details and Hyperparameters

We follow the instruction-tuning protocol of LLaVA-1.5 [33]. In the pretraining step, we employ the caption data to optimize the projector, while keeping the vision encoder and LLM frozen. Meanwhile, we optimize the projector and LLM in the instruction tuning step. During the pretraining phase, we utilize a batch size of 256, while for the instruction tuning phase, we employ a batch size of 128. The learning rate is set at  $1e^{-3}$  during pretraining and adjusted to  $2e^{-5}$  for instruction tuning, with both phases incorporating a cosine decay schedule. We also apply a learning rate warmup with a decay factor of 0.03, and no weight decay is used. Both pretraining and instruction tuning are conducted for one epoch each, consistently using the AdamW optimizer.

### B More Effects of Web2Code on General Domain

Here, we first perform instruction tuning using Web2Code on various LLM backbones and then we evaluate those MLLMs on the general domain of visual language understanding.

**Comparison on different LLM backbones.** We compare the general domain abilities of various LLM backbones under the same data configuration of LLaVA + DWU + DWCG; Table 6 summarizes the results of instruction-tuned MLLMs. Specifically, we found that instruction-tuned CrystalChat-7B, Vicuna1.5-7B, and LLaMA2-7B show superior performances in the general domain compared to CrystalCoder and CodeLlama. For example, CrystalChat shows +132.89 points higher than CodeLlama in MME-P (i.e. perception domain). Somewhat surprisingly, instruction-tuned CrystalChat showed the strongest performance on TextVQA, which requires visual reasoning based on text in images.

LLM Backbone	MME-P	MME-C	POPE	SciQA	TextVQA
CodeLlama-7B [45]	1345.93	258.92	85.28	61.87	55.23
CrystalCoder-7B [37]	1351.22	274.64	86.05	61.63	50.11
CrystalChat-7B [37]	1478.82	297.14	86.14	67.92	<b>57.41</b>
Vicuna1.5-7B [11]	<b>1488.26</b>	268.21	<b>87.05</b>	<b>69.31</b>	56.40
LLaMA2-7B [54]	1448.58	<b>304.29</b>	86.87	67.87	55.62

Table 6: Comparison of different LLM backbones on visual language understanding benchmarks. All models are instruction-tuned on the general domain data (i.e. LLaVA) with DWU and DWCG.

### C Qualitative Data Examples in WUB Benchmark

The qualitative data examples in our WUB benchmark are shown in Figure 10. It covers different aspects of webpage understanding based on "yes" / "no" question-answer pairs.

### D WCGB Criteria

Our proposed WCGB framework consists of 10 distinct criteria, which we group into four categories, each encompassing specific criteria that are scored on a 0-10 scale:

#### 1. Visual Structure and Alignment

- *Layout Consistency*: Measures the arrangement of structural webpage elements like headers, footers, and sidebars.
- *Element Alignment*: Assesses the alignment of images, buttons, and text boxes.
- *Proportional Accuracy*: Checks for consistency in sizes and aspect ratios of visual elements.
- *Visual Harmony*: Examines the overall balance and harmony in design.

#### 2. Color and Aesthetic Design

- *Color Scheme and Aesthetic Match*: Focuses on the similarity in color schemes, including hues and saturation.

**Welcome to the 2024 Tennis Tournament**

**Tournament Overview**  
Our annual tennis tournament will take place in the sunny courts of Springfield from July 10th to 17th, 2024. It's open to players of all skill levels across various age groups.

**Register Now**

**Schedule**

- Day 1**
  - Opening Ceremony
  - Junior Matches
- Day 2**
  - Adult Round 1
  - Senior Match

**Frequently Asked Questions**

Q: Can users find information about specific match timings for the tournament directly on this page? Output YES or NO. A: NO

Q: Does the webpage employ a monochromatic color scheme for its design? Output YES or NO. A: NO

Q: Is there a consistent font style used throughout the different sections of the webpage? Output YES or NO. A: YES

**Global Gateway Airport**

**Live Flight Information**

**Departures**

Flight	Destination	Scheduled	Status
GG802	New York (JFK)	10:15 AM	Delayed
GG543	Los Angeles (LAX)	11:45 AM	Delayed
GG131	London (LHR)	12:30 PM	Delayed
GG995	Dubai (DXB)	1:40 PM	Delayed

Q: Does the visual structure of the webpage indicate that the flight information is meant for simultaneous display of both departures and arrivals? Output YES or NO. A: NO

Q: Are there interactive elements visible that would allow a user to update the flight status information on the webpage? Output YES or NO. A: NO

Q: Do the flight numbers follow a consistent naming convention that assists users in recognizing the airline's code and flight sequence? Output YES or NO. A: YES

Figure 10: Qualitative data examples in our WUB benchmark. It covers different aspects of webpage understanding based on "yes" / "no" question-answer pairs.

- *Aesthetic Resemblance*: Looks at the overall aesthetic appeal and style (modern, minimalist, traditional, etc.).

### 3. Textual and Content Consistency

- *Font Characteristics and Consistency*: Assesses uniformity in font type, size, style, and weight.
- *Textual Content Match*: Evaluates the match in words and sentences.
- *Numeric and Special Character Accuracy*: Checks for consistency in numbers, dates, and special characters.

### 4. User Interface and Interactivity

- *User Interface Consistency*: Assesses the similarity in design language and appearance of UI elements like menus, buttons, and forms.

## E Prompt Templates

### E.1 Prompt Used to Question and Answer Generation for DWU Data

```
System: You are an advanced AI model who can identify html code and interpret the compiled webpage.

User: You are asked to come up with a set of 10 diverse website understanding task instructions with the corresponding source codes. These task instructions will be given to a GPT model and we will evaluate the GPT model for completing the instructions.

You should output 10 questions and the answers for them, for each html code.
Your output should be formatted as follows. Don't include any other text, connective phrases other than the formatted text.
output a comma separated list of 10 instruction pairs in following format
[{"Q": "<question>" , "A": "<answer>"}, {"Q": "<question>" , "A": "<answer>"}]

Here are the requirements:
1. Try not to repeat the verb for each instruction to maximize diversity.
2. The languages used for the instruction should be diverse. For example, you should combine questions with imperative instructions.
3. The instructions should be in English.
4. The instructions should be at least 1 to 2 sentences long. Either an imperative sentence or a question is permitted.
5. Instructions should be clear and precise.
6. Instructions can either be simple queries like 'what text is displayed on the button?', to more complex tasks requiring analysis, such as 'what would be the total cost of ordering items A and B from the menu?'.
7. Avoid direct html code related instructions. i.e. Don't include questions related to html tags, font sizes, and etc. Instead include real webpage image related instructions that a human can ask after looking into the webpage. For example, What is the main event that this website is allowing users to register for?, State the variety of dishes featured in the desserts section of the menu.
8. The answer should be an appropriate response to the instruction and the input.
9. If the answer isn't immediately clear from the content, the output should detail the thought process. For instance, if the task is to determine the age of someone born in 1983, the response should explain: 'Considering the current year is 2024, the age would be 2023 minus 1984, which equals 40.'
10. Try to make questions covering the following aspects i.e. Visual Structure and Alignment: Layout Consistency: Evaluates the match in the placement of headers, footers, and sidebars. Element Alignment: Assesses alignment of images, buttons, and text boxes. Proportional Accuracy: Checks for consistency in sizes and aspect ratios of visual elements. Visual Harmony: Examines the overall balance and harmony in design. Color and Aesthetic Design: Color Scheme and Aesthetic Match: Focuses on the similarity in color schemes, including hues and saturation. Aesthetic Resemblance: Looks at the overall aesthetic appeal and style (modern, minimalist, traditional, etc.). Textual and Content Consistency: Font Characteristics and Consistency: Assesses uniformity in font type, size, style, and weight. Textual Content Match: Evaluates the match in words and sentences. Numeric and Special Character Accuracy: Checks for consistency in numbers, dates, and special characters. User Interface and Interactivity: User Interface Consistency: Assesses the similarity in design language and appearance of UI elements like menus, buttons, and forms.
11. Answers should be more informative and descriptive. Don't output single word or numbers for the outputs, make them descriptive with one or two sentences.
12. Don't shrink the output. Need everything expanded. Avoid outputs like (...)

Here's the html code.
<!DOCTYPE html>
<html>
<head>
<title>Charity Walk Registration</title>
</head>
<body>
<h1>Welcome to the Charity Walk Registration Site</h1> ...
```

Figure 11: Prompt used to generate Question Answer pairs using GPT4 for DWU data.

## E.2 Prompts Used For Instruction generation of DWCG and DWCG<sub>R</sub>

**User:**  
Here are some instructions that are being fed to a generative AI model as instruction tuning during the training of a Vision language model.

"In the provided webpage screenshot, generate HTML to replicate the layout and styling of the webpage",  
"Given the web application interface shown, write HTML code to implement the interactive features visible in the image",  
"Analyze the webpage structure in the screenshot and provide HTML and CSS code to design a responsive webpage with similar layout and components",  
"Examine the provided webpage screenshot and generate HTML code to implement the responsive layout",  
"Given the webpage screenshot, provide the HTML code that represents its structure",  
"Create HTML code for the webpage depicted in the image provided",  
"Craft HTML code to replicate the visual design and structure of the webpage captured in the provided screenshot",  
"Create HTML and CSS code to imitate the appearance and layout of the web interface shown in the provided image",  
"Generate HTML and CSS code snippets to mirror the layout and styling of the website visible in the given screenshot"

**Note:**

- 1.Try not to repeat the verb for each instruction to maximize diversity.
- 2.The languages used for the instruction should be diverse. For example, you should combine questions with imperative instructions.
- 3.The instructions should be in English.
- 4.The instructions should be at least 1 to 2 sentences long. Either an imperative sentence or a question is permitted.
- 5.Instructions should be clear and precise.
- 6.Please provide 150 examples.

Figure 12: Prompt used to generate instructions for DWCG using GPT4, feeding input as Seed instructions and output as GPT generated instructions shown in Figure 16.

**User:**  
Here are some instructions that are being fed to a generative AI model as instruction tuning during the training of a Vision language model.

"Please provide the code in material design style.",  
"Share the code adhering to the principles of material design.",  
"Ensure the code follows the guidelines of material design aesthetics.",  
"Provide the code in a style consistent with material design principles.",  
"Code should reflect the design language of material design.",  
"Maintain a coding style aligned with material design aesthetics.",  
"Present the code in accordance with material design styling."

**Note:**

- 1.Try not to repeat the verb for each instruction to maximize diversity.
- 2.The languages used for the instruction should be diverse. For example, you should combine questions with imperative instructions.
- 3.The instructions should be in English.
- 4.The instruction should be clear.
- 5.Please provide 40 examples.

Figure 13: Prompt used to generate webpage style instruction for DWCG using GPT4, feeding input as Seed instructions and output as GPT generated webpage style instructions shown in Figure 16.

### E.3 Prompt Used For GPT4-Vision Evaluation in WCGB benchmark

**System:** You are an advanced AI model equipped with OCR and image understanding capabilities, capable of analyzing visual elements in detail.

**User:** Your task is to assess two webpage images and output a score between 0 and 10 for each of the following questions.

If the answer to a question is a definite YES, output a score of 10, signifying perfect similarity. Conversely, a definite NO should yield a score of 0, indicating no similarity.

For answers that fall in between, assign a score accordingly, where a higher number indicates a greater degree of similarity. Only provide the numerical score for each question, without any additional text. Example contexts are provided for clarity. Examples provides the idea, but you can output any number in 0-10 range accordingly.

Only output a comma separated list containing 10 numbers. DO NOT give score of 10 for any category unless otherwise the two images are identical.

Layout Consistency (Score: 0-10): Does the placement of headers, footers, and sidebars match in both webpages? (e.g., A score of 10 for identical layouts, 5 for similar but not exact placements, and 0 for completely different layouts.)

Element Alignment (Score: 0-10): Are elements like images, buttons, and text boxes aligned similarly on both pages? (e.g., A score of 10 for perfectly aligned elements, 6 for slight misalignments, and 0 for major misalignments.)

Proportional Accuracy (Score: 0-10): Do the sizes and aspect ratios of images, buttons, and text boxes appear consistent across both pages? (e.g., A score of 10 for exact proportions, 4 for noticeable size differences, and 0 for drastic inconsistencies.)

Visual Harmony (Score: 0-10): Do both webpages exhibit a similar level of visual harmony and balance in their design? (e.g., A score of 10 for harmonious designs, 5 for some dissonance, and 0 for clashing designs.)

Color Scheme and Aesthetic Match (Score: 0-10): How closely do the color schemes of the two webpages align in terms of background and text colors? Evaluate the similarity in hues, saturation, and overall color aesthetics. (e.g., A score of 10 for perfectly matching color schemes, including identical hues and saturation levels, 6 for similar color palettes with minor variations, and 0 for starkly different color schemes that create entirely different visual impacts.)

Aesthetic Resemblance (Score: 0-10): Is the overall aesthetic appeal (modern, minimalist, traditional, etc.) similar on both pages? (e.g., A score of 10 for identical aesthetics, 4 for somewhat similar but distinguishable styles, and 0 for completely different aesthetics.)

Font Characteristics and Consistency (Score: 0-10): Assess the degree of consistency in font attributes across both webpages. This includes not only the font type and size but also the nuances of font style (italic, bold) and weight (light, regular, bold). (e.g., A score of 10 for complete uniformity in font type, size, style, and weight across both pages, 5 for consistency in font type and size but variations in style or weight, and 0 for wide disparities in font type, size, style, or weight, leading to a distinctly different textual appearance.)

Textual Content Match (Score: 0-10): Do the words and sentences match between the two webpages? (e.g., A score of 10 for identical text, 5 for some similar paragraphs or sections, and 0 for completely different textual content.)

Numeric and Special Character Accuracy (Score: 0-10): Are numbers, dates, and special characters (like email addresses) consistent between the two pages? (e.g., A score of 10 for exact matches, 6 for minor discrepancies, and 0 for major differences.)

User Interface Consistency (Score: 0-10): Do the user interface elements (like menus, buttons, and forms) on both pages share a similar design language and appearance? (e.g., A score of 10 for identical UI elements, 6 for slight design variations, and 0 for completely different UI designs.)

<GROUND TRUTH IMAGE>

<PREDICTED IMAGE>

Figure 14: The prompt utilized for evaluation in WCGB with the employment of GPT4-Vision.

## E.4 Prompt Used For QA Generation for WUB benchmark

**User:** Generate 5 QA pairs for assessing an AI model's comprehension and response accuracy in relation to a given webpage screenshot.

The answers should be evenly balanced between YES and NO. These QA pairs will be used as a benchmark to evaluate another model's understanding of the webpage's visual and textual elements.

Ensure the output is formatted as QA pairs in the following dictionary format: '{ 'Q': \"question\", 'A': \"answer\"}'.

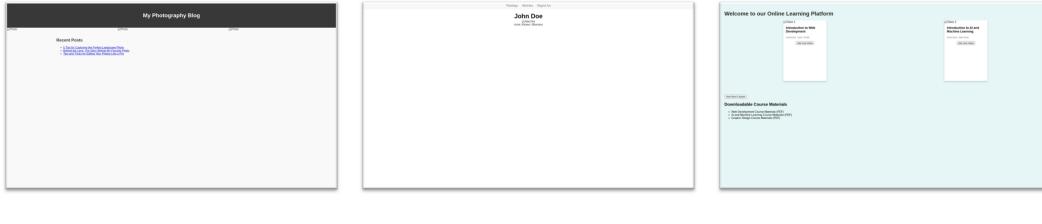
The QA pairs should cover the following aspects:

- Advanced Logical Reasoning and Textual Understanding: The questions should challenge the AI's ability to understand complex text and infer meanings or implications that are not explicitly stated.
- Visual Structure and Alignment Evaluation: Questions should test the AI's ability to interpret and analyze the webpage's layout, including the positioning and relationship of various visual elements.
- Color and Aesthetic Design Analysis: Include questions that assess the AI's understanding of the webpage's color scheme and overall design aesthetics, and how they contribute to the page's purpose or message.
- Consistency in Text and Content: Formulate questions that evaluate the AI's recognition of textual consistency and coherence across different sections of the webpage.
- User Interface and Interactivity Insights: Develop questions to assess the AI's comprehension of the webpage's interactive elements and user interface design, including navigational features and response mechanisms.

Remember, the goal is to create questions that will test another AI model's ability to analyze and interpret a webpage screenshot in a comprehensive manner. Do not include additional text other than the output dictionary entry.

Figure 15: The prompt employed to generate "yes" / "no" Question Answer pairs for the WUB benchmark through the utilization of GPT4-Vision.

## F Data samples



### Instruction:

<image>[using the provided webpage screenshot as inspiration, produce HTML code to mimic its design. Display code which integrates seamlessly with material design styling.]

### Response:

```
<!DOCTYPE html><html lang="en"><head><meta charset="UTF-8"><meta name="viewport" content="width=device-width, initial-scale=1.0"><title>My Photography Blog</title><style></style></head><body>
```

### Instruction:

<image>[Devise HTML/CSS code that mimics the layout and styling of the website from the provided screenshot. Elaborate on code that aligns with the material design standards.]

### Response:

```
<!DOCTYPE html><html lang="en"><head><meta charset="UTF-8"><meta name="viewport" content="width=device-width, initial-scale=1.0"><title>Online Learning Platform</title><style></style></head><body>
```

### Instruction:

<image>[Offer the HTML and CSS code that will replicate the given webpage screenshot. Grant code characterized by facets of material design.]

### Response:

```
<!DOCTYPE html><html lang="en"><head><meta charset="UTF-8"><meta name="viewport" content="width=device-width, initial-scale=1.0"><title>Online Learning Platform</title><style></style></head><body>
```

Figure 16: Examples of webpage to code generation instruction tuning data. These web image-code pairs were converted into an instruction-following data format close to the LLaVA data format.



**Instruction:**  
<image>\nAnalyze the image provided and replicate its web design using HTML and CSS code. Will you provide the code modeled by Bootstrap guidelines?

```
Response:
<html><n-in-header><n-meta charset="utf-8"/><n-meta
content="width=device-width, initial-scale=1"
name="viewport"><n-link crossorigin="anonymous"
href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.7/css/bootstrap.min.css"
integrity="sha384-BVYiiKIFiBeQzKk@@ocsQvZt+giRchq7zC63" />
```

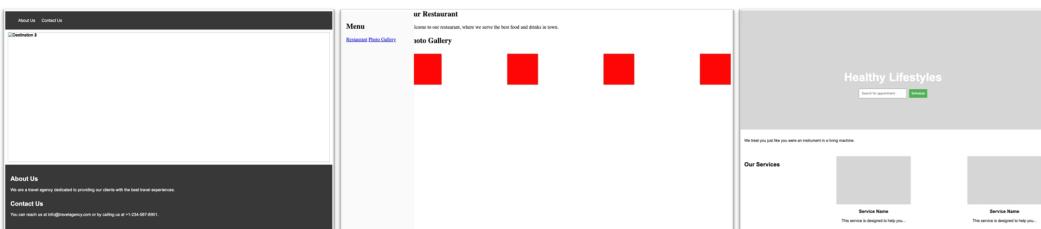
**Instruction:**  
<image>\nSeeing the webpage screenshot, can you generate HTML to replicate its layout? Could you deliver the code with Bootstrap conformities?

```
Response:  
<html>\n<header>\n<meta charset=\"utf-8\"/\>\n<meta  
content=\"width=device-width, initial-scale=1\"  
name=\"viewport\"/\>\n<link  
crossorigin=\"anonymous\"  
href=\"https://maxcdn.bootstrapcdn.com/bootstrap/3.3.  
7/css/bootstrap.min.css\"\ integrity=\"sha384-  
7z8JwDCEGv3SzfHdDk6t8t8EeGmKuZB1jwDlqYBzqyZLc\"/>  
</head>\n<body>\n<div class=\"container\" style=\"text-align: center; margin-top: 10px;\">\n<h1>Hello, World!</h1>\n<hr style=\"border: 1px solid black; margin: 10px 0;\"/\>\n<h2>This is a simple example of a responsive bootstrap page</h2>\n<hr style=\"border: 1px solid black; margin: 10px 0;\"/\>\n<h3>Check it out on your mobile device or tablet!</h3>\n</div>\n</body>\n</html>
```

**Instruction:**  
<image>\nSeeing the webpage screenshot, can you generate HTML to replicate its layout? Can you exhibit the code capturing the essence of Bootstrap?

```
Response:
<html><n><head><n><meta charset="utf-8"/><n><meta
content="width=device-width, initial-scale=1"
name="viewport"/><n><link crossorigin="anonymous"
href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.7/css/bootstrap
.min.css" integrity="sha384-
BXVJLoRAjLdJdWYtC8vXgI6PZTlJXaAqOgWQbB2M1Oy
AkozA1A UP
```

Figure 17: Examples of webpage to code generation instruction tuning data in DWCG. These web image-code pairs were converted into an instruction-following data format close to the LLaVA data format.

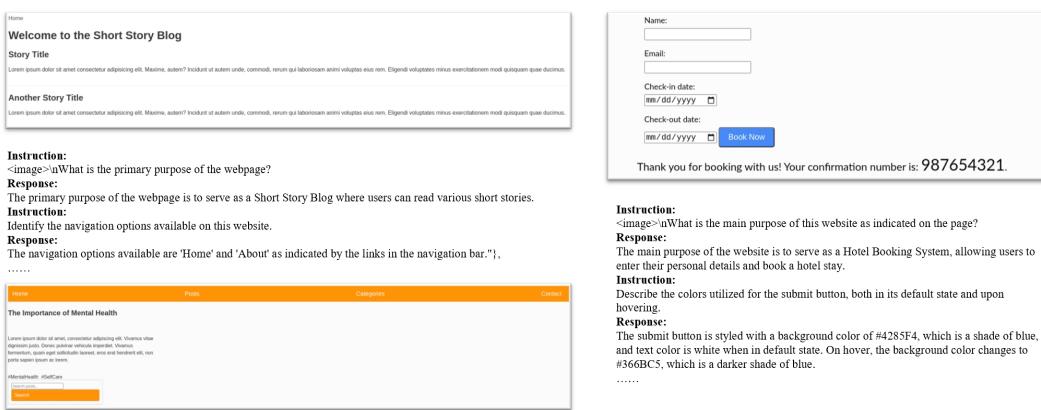


**Instruction:**  
<image></image>Using the webpage screenshot, create a simulating HTML/CSS code. Show the code developed according to material design principles.

**Instruction:**  
<image></image>Design HTML and CSS code to replicate the layout and visuals of the user interface visible in the image given. Provide code organized around the norms of material design.

**Instruction:** <image></image> inMold the HTML syntax to match with the layout of the webpage indicated in the screenshot. Explain code that synchronizes with the material design analogy.

Figure 18: Examples of webpage to code generation instruction tuning data in DWCG<sub>R</sub>. The refined instruction tuning dataset for webpage code generation, by utilizing GPT-4.



**Instruction:** `<image>`What is the central topic of the latest blog post?  
**Response:** The central topic of the latest blog post is 'The Importance of Mental Health' as indicated by the heading in the blog section.  
**Instruction:** Identify the primary colors used in the navigation bar.  
**Response:** The primary colors used in the navigation bar are (specifically) #FFD700 for the background color for the top bar.

Name:	<input type="text"/>
Email:	<input type="text"/>
Check-in date:	<input type="text"/> mm/dd/yyyy <input type="button" value="..."/>
Check-out date:	<input type="text"/> mm/dd/yyyy <input type="button" value="..."/>
<input type="button" value="Book Now"/>	

**Instruction:**  
<image>\nWhat is the main purpose of this website as indicated on the page?  
**Response:**  
The main purpose of the website is to serve as a Hotel Booking System, allowing users to

enter their personal details and book a hotel stay.

**Response:**  
The submit button is styled with a background color of #4285F4, which is a shade of blue and text color is white when in default state. On hover, the background color changes to

Figure 19: Examples of webpage understanding data in DWU. For a webpage, there is a set of diverse question-answer pairs about web understanding.

**Instruction:** <image>\nWhat's the "HT" of the player Willy Hernangomez?  
**Response:** Willy Hernangomez's height is 6 feet 11 inches.

**Instruction:** Who holds the "DOB" of 5/27/1994?  
**Response:** Willy Hernangomez was born on May 27, 1994

.....

**Instruction:** <image>\nTell me the value of STL in career  
**Response:** The STL value over the career is 0.4 per game.

**Instruction:** How about the STL value in regular season?  
**Response:** During the regular season, the STL value is 0.2 per game.

.....

**Instruction:** <image>\nWhat's the lowest level do the graphics have to be for this game?  
**Response:** The minimum graphics card requirement for this game is either a GTX 660Ti with 3GB of VRAM or an R9 270X with 4GB of VRAM.

**Instruction:** What are the recommended graphics card requirements for this game?  
**Response:** For an optimal gaming experience, it is recommended to have a GTX 1060 with 6GB of VRAM or an RX 480 with 8GB of VRAM.

.....

Figure 20: Examples of webpage understanding data in DWUR. The refined instruction tuning dataset for web understanding tasks.

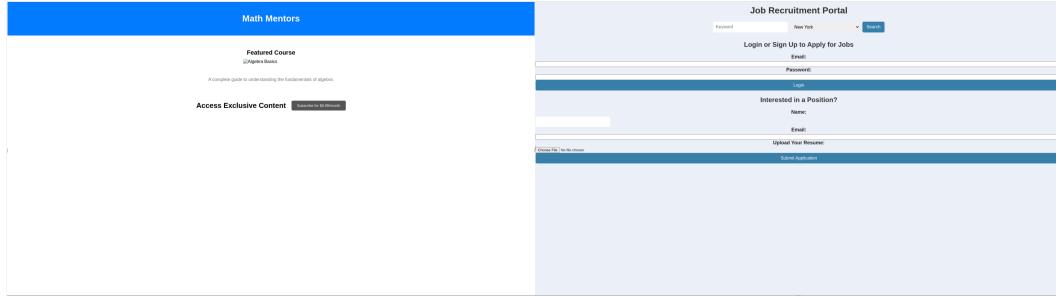


Figure 21: Examples of Modern styles webpages.

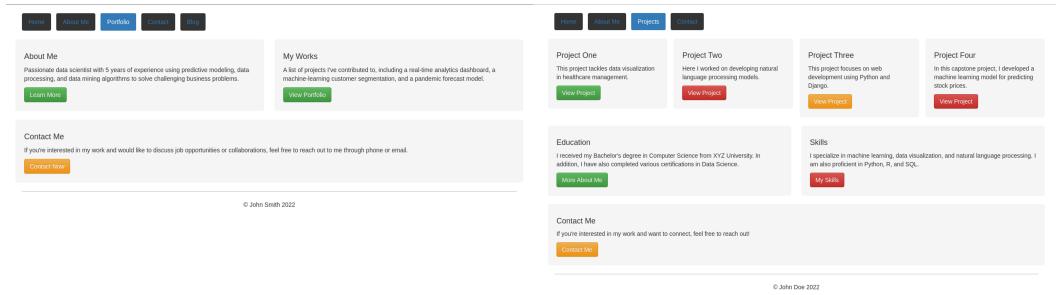


Figure 22: Examples of Bootstrap styles webpages.