

# IW-Bench: Evaluating Large Multimodal Models for Converting Image-to-Web

Hongcheng Guo<sup>1</sup>, Wei Zhang<sup>1</sup>, Junhao Chen<sup>3</sup>, Yaonan Gu<sup>5</sup>, Jian Yang<sup>2</sup>, Junjia Du<sup>4</sup>, Shaosheng Cao, Binyuan Hui<sup>2</sup>, Tianyu Liu<sup>2</sup>, Jianxin Ma<sup>2</sup>, Chang Zhou<sup>2</sup>, Zhoujun Li<sup>1\*</sup>,

<sup>1</sup>Beihang University <sup>2</sup>Alibaba Group <sup>3</sup>Tsinghua University

<sup>4</sup>Nanyang Technological University <sup>5</sup>National University of Singapore

{lizj, hongchengguo}@buaa.edu.cn,

## Abstract

Recently advancements in large multimodal models have led to significant strides in image comprehension capabilities. Despite these advancements, there is a lack of the robust benchmark specifically for assessing the Image-to-Web conversion proficiency of these large models. Primarily, it is essential to ensure the integrity of the web elements generated. These elements comprise visible and invisible categories. Previous evaluation methods (e.g., BLEU) are notably susceptible to significant alterations due to the presence of invisible elements in Web. Furthermore, it is crucial to measure the layout information of web pages, referring to the positional relationships between elements, which is overlooked by previous work. To address challenges, we have curated and aligned a benchmark of images and corresponding web codes (**IW-BENCH**). Specifically, we propose the **Element Accuracy**, which tests the completeness of the elements by parsing the Document Object Model (DOM) tree. **Layout Accuracy** is also proposed to analyze the positional relationships of elements by converting DOM tree into a common subsequence. Besides, we design a five-hop multimodal Chain-of-Thought Prompting for better performance, which contains five hop: 1) SoM prompt injection. 2) Inferring Elements. 3) Inferring Layout. 4) Inferring Web code. 5) Reflection. Our benchmark comprises 1200 pairs of images and web codes with varying levels of difficulty. We have conducted extensive experiments on existing large multimodal models, offering insights into their performance and areas for improvement in image-to-web domain.<sup>1</sup>

## 1 Introduction

The development of large multimodal models has emerged as a new trend, starting with GPT-4 (OpenAI, 2023b). An increasing number of multimodal

\*Corresponding author.

<sup>1</sup>We provide code and dataset at <https://github.com/HC-Guo/IWBench>

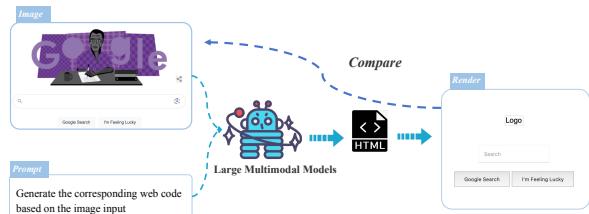


Figure 1: The process of Image-to-Web. We prompt a large multimodal model to generate web code based on the input image. Finally, we need to compare whether the newly rendered result is consistent with the input image.

models have been introduced (Anil et al., 2023; Liu et al., 2023a; Zhu et al., 2023a; Bai et al., 2023; Dai et al., 2023), extending the powerful comprehension capabilities of large language models to multiple tasks (Lu et al., 2018; Yang et al., 2023; Sidorov et al., 2020). Recently, the task of converting images into web code (Laurençon et al., 2024; Patil et al., 2020; Beltramelli, 2017) has garnered significant attention due to its impressive performance. In Figure 1, this task tests the synthesis abilities of large multimodal models, encompassing the fine-grained recognition of components within images, the assessment of the relative positioning of elements in webpages, and the capability to generate code. The prowess of large models in generating front-end code has been a source of astonishment, yet, notably, there has been almost no evaluation benchmark related to large multimodal models for this domain.

Constructing a benchmark for appraising Image-to-Web tasks is markedly more intricate than the conventional programming challenges (Davody et al.). This heightens complexity stems from the multifaceted nature of web code, encompassing HTML, CSS, and JavaScript. It demands scrutiny not only of functional correctness but also of visual elements like layout, design, and user interaction. Predominant evaluation frameworks for the Image-

to-Code task (Patil et al., 2020; Bhambure et al.) often hinge on similarity metrics such as BLEU. Nevertheless, these metrics encounter substantive limitations in the context of web evaluation. The intricate web of JavaScript functionalities and CSS styling options means that disparate web codes can produce indistinguishable visual outcomes. Consequently, such conventional metrics falter in precisely assessing the completeness of elements depicted in images and their layout intricacies. A further complication arises from non-visible elements, introducing a fragility in these evaluation methods. Varied development ways can lead to substantial discrepancies in these non-visible components, even among web pages that appear visually identical. Recent WebSight (Laurençon et al., 2024) is finetuned on numerous image-code pairs, however it has not conducted performance evaluations. These challenges underscore the imperative for a more sophisticated and comprehensive benchmark.

To address the challenges, we introduce IW-BENCH to assess the capabilities of large multimodal models in the Image-to-Web. Firstly, we design a data construction pipeline, comprising 1200 entries of three difficulty levels: Level I (*simple*), Level II (*medium*), and Level III (*complex*). For metric innovation, we propose **Element Accuracy** and **Layout Accuracy**. Specifically, to effectively measure the completeness of web elements, the element accuracy tests the completeness of the elements from six dimensions (Tag, Text Content, Attribute, Style, JavaScript, Children) by parsing the Document Object Model (DOM) tree. To analyze the relative positional relationships of elements, we first traverse the DOM Tree in the same manner (in-order traversal) to obtain a list of elements, then we calculate the overlap between the common sub-sequence and the ground truth for the layout accuracy. Moreover, we establish a effective five-hop multimodal chain-of-thought method to enhance the model performance, including five specialized hops: SoM prompt injection, Inferring Elements, Inferring Layout, Inferring Web code, and Reflection.

To summarize:

- **Benchmark Construction.** We have meticulously curated 1200 three challenging levels of image-web pairs for our benchmark dataset. This dataset serves as a rigorous benchmark for assessing the capabilities of large multi-

modal models in the task of converting images to web code.

- **Metric Innovation.** We introduce innovative metrics to evaluate web elements and layout information accurately. Specifically, we have developed the Element Accuracy metric to assess the fidelity of web elements and the Layout Accuracy metric to evaluate the precision of layout information.
- **Multimodal Chain-of-Thought.** The five-hop multimodal Chain-of-Thought method is proposed by us, significantly enhancing the performance in image-to-web domain.
- **Extensive Evaluation.** A substantial evaluation of large multimodal models has been conducted, showcasing their capabilities and limitations in our context. Besides, ablation on our five-hop chain-of-thought demonstrates the effect of our method.

## 2 Related work

Multimodal benchmarks for large multimodal models (Bitton et al., 2023; Yu et al., 2023; Liu et al., 2023c; Xu et al., 2023b; Shao et al., 2023) have assess the instruction-following and reasoning capabilities. As these foundation models become more relevant to real-world applications, unlike prior work, we propose IW-BENCH to benchmark their generation capabilities of the hot Image-to-Web area on a diverse set of visual contexts. To the best of our knowledge, WebSight (Laurençon et al., 2024) is the only relevant benchmark. However, it is more like a fine-tuning dataset for large multimodal models rather than a benchmark for evaluation. The detailed comparison is in Table 1. More related work on image-to-web, large multimodal models, and chain-of-thought methods are in Appendix C.

Feature	IW-Bench	Websight
Language Coverage	Chinese and English	English
Real Web Page Samples	✓	✗
Data De-identification	✓	✗
Enhanced Data Quality	Automated and Manual Review	Automated Review
Quantitative Analysis	✓	✗
Human Evaluation	✓	✗
Evaluation Metrics	Element and Layout Accuracy	✗

Table 1: Comparison between IW-BENCH and Web-sight.

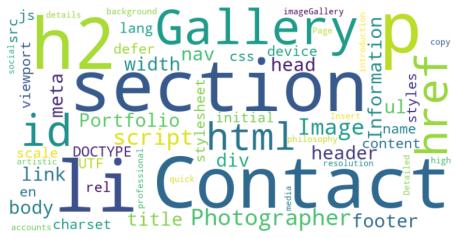


Figure 2: Wordcloud. The key words in IW-BENCH are related to the web and internet, such as 'html', 'header'.

3 IW-BENCH

**Overview:** We design a pipeline for the construction of IW-BENCH in Figure 3. Specifically, to obtain complex-level data, we crawl publicly available web pages and perform de-identification and simplification. For simple and medium level data, we prompt GPT-4 ([OpenAI, 2023a](#)) to generate them. Finally, we filter out some low-quality samples from the obtained image-code pairs. Quantity of IW-BENCH is in Table 2 and word cloud is in Figure 2. We present examples in Appendix I.

	#Level I (Simple)	#Level II (Medium)	#Level III (Complex)
Number	340	645	215
Source	GPT4	GPT4	Internet

Table 2: Quantity of IW-BENCH. Varied levels of complexity exist within dataset with varying numbers. The sources of IW-BENCH are GPT4 and real web pages.

### 3.1 Data Collection

We initially intend to assign the task of determining complexity to human workers and exclusively use real web. However, tests reveal that websites in real-world scenarios are too complex, resulting in uniformly poor evaluation with minimal differentiation between models. The criteria for different complexity levels is in Table 3.

Web Page Complexity Level	Number of Elements	Number of Attributes	JavaScript Usage
Simple	1-20	1-20	None or minimal
Medium	20-60	20-60	Moderate
Complex	61+	61+	High

Table 3: Web Page Complexity Levels and Their Characteristics

To better quantify differences among models, we opt to create simpler data by prompting GPT4 ([OpenAI, 2023a](#)) following the previous work ([Wang et al., 2023c; Xu et al., 2023a](#)). The process begins with GPT-4 randomly selecting a relevant domain

in Appendix E.1. Based on the desired complexity, GPT-4 then chooses specific elements. Finally, it selects the JavaScript events in Appendix D.1 to be integrated into the web page. In IW-BENCH, 18% of the data comes from real web pages, covering a wide range of types, including public repositories, open-source projects, and websites representing various content categories. The data collection process strictly follows the robots.txt protocol and complies with website policies. To extract images from web code, we use Playwright<sup>2</sup>.

### 3.2 Data Processing

The processing of real web page data includes two parts: de-identification and simplification. For de-identification, we replace the image components and modify the logos within our dataset to avoid copyright or other legal concerns. More details are in Appendix A. For simplification, this involves the removal of non-essential invisible elements such as whitespace, code comments, and styles. We also streamline HTML by condensing verbose code and transferring inline styles to external stylesheets, which contributes to a more uniform and computationally efficient dataset. Further, we standardize visible elements like images and text by replacing them with placeholders and normalizing stylistic features.

We carefully validate the simplifications to ensure that the functionality and layout of the original webpage remain unaffected. Specifically, we start by launching a headless browser using Playwright and capturing the initial screenshot of the webpage. After that, we employ BeautifulSoup to parse the HTML document. Subsequently, we traverse through all HTML elements, with the exception of fundamental tags such as html, head, body, title, and meta, as these are typically crucial components of the structure. We initiate the removal process and then scrutinize whether there is any discernible alteration in the visual presentation. To determine this, we render the modified HTML content and compare it with the original one. If the new screenshot matches the original one, we retain the deleted element and document it accordingly. The time cost of simplification is in Table 4.

### 3.3 Expert Review

Establishing IW-BENCH involves substantial manpower. In the collection phase, 20 web engineers as-

<sup>2</sup><https://github.com/microsoft/playwright>

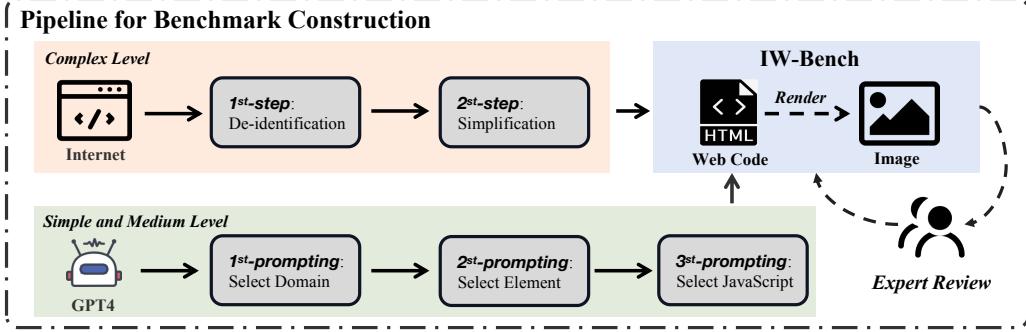


Figure 3: Benchmark Construction. This pipeline illustrates the multi-step process used to construct IW-BENCH for web code and images of varying complexity levels.

Web Page Complexity	Average Simplification Time (seconds)
Simple	44.3
Medium	131.5
Complex	238.2

Table 4: Simplification Time Cost.

sess complexity based on HTML tag variety, DOM tree depth, and scripting presence. Each data point is cross-validated by three engineers, with disagreements resolved through training, detailed guidelines, and a majority-rule consensus mechanism. Low-quality examples are re-evaluated until agreement is achieved.

### 3.4 Metric Design

To evaluate the fidelity of web elements, we propose two novel metrics: **Element Accuracy (EA)** and **Layout Accuracy (LA)**, which are designed to compare visible and invisible elements in web pages. The evaluation process begins with constructing a **Document Object Model (DOM)** tree for the test and reference web pages, followed by a structured traversal and analysis of the elements. The DOM tree traversal systematically inspects each element, collecting detailed information such as tag types, attributes, styles, JavaScript bindings, and hierarchical relationships. This unified approach enables the evaluation of visible elements (e.g., text, images, buttons) alongside invisible ones (e.g., `<script>` tags, metadata).

#### 3.4.1 Element Accuracy (EA)

Element Accuracy measures the similarity between corresponding elements in the test and reference web pages based on six key attributes: **Tag**, **Text Content**, **Attributes**, **Style**, **JavaScript**, and **Children**. Let the test element set be  $E_{\text{test}}$ , and the label element set be  $E_{\text{label}}$ . For each test element

$E_j$ , an average score  $S_j$  is computed across these six perspectives. The Element Accuracy is defined as:

$$EA = \frac{\sum_{j \in E_{\text{test}}} \mathbb{1}(E_j > T)}{|E_{\text{label}}|} \quad (1)$$

where  $T$  is the threshold, and  $\mathbb{1}(E_j > T)$  is an indicator function that equals 1 if  $E_j$  exceeds  $T$ , and 0 otherwise.

The similarity scores for each perspective are computed as follows:

- **Tag:** Tags are compared by their names:

$$S_{\text{tag}} = \begin{cases} 1 & \text{if } \text{tag}_{\text{test}} = \text{tag}_{\text{label}}, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

- **Text Content:** The similarity of text content between two elements is computed by SequenceMatcher (SM)<sup>3</sup>:

$$S_{\text{text}} = SM(\text{content}_{\text{test}}, \text{content}_{\text{label}}) \quad (3)$$

- **Attributes:** A predefined mapping is used for each tag type. If the corresponding attributes of two elements match exactly, they receive a higher score:

$$S_{\text{attr}} = \frac{\sum_{k \in A} \mathbb{1}(\text{attr}_k^{\text{test}} = \text{attr}_k^{\text{label}})}{|A|} \quad (4)$$

where  $A$  is the set of relevant attributes for the tag.

- **Style:** Key style properties (e.g., color, font-size) are filtered out, ignoring default values. Matching values contribute to the score:

$$S_{\text{style}} = \frac{\sum_{p \in P} \mathbb{1}(\text{style}_p^{\text{test}} = \text{style}_p^{\text{label}})}{|P|} \quad (5)$$

where  $P$  is the set of key style properties.

<sup>3</sup><https://github.com/m-matelski/mdiff>

- **JavaScript:** Event bindings on two elements are compared (e.g., `onclick`, `onload`). Let  $J$  be the set of relevant events. The JavaScript score is:

$$S_{js} = \frac{\sum_{e \in J} \mathbb{1}(\text{event}_e^{\text{test}} = \text{event}_e^{\text{label}})}{|J|} \quad (6)$$

- **Children:** The hierarchical structure is evaluated using a tree-edit distance algorithm.

The average score for each test element  $E_j$  is:

$$E_j = \frac{\sum_{i=1}^n S_i}{n}, \quad (7)$$

where  $n = 6$  represents the number of evaluated attributes: *Tag*, *Text Content*, *Attribute*, *Style*, *JavaScript*, *Children*.

### 3.4.2 Layout Accuracy (LA)

Layout Accuracy quantifies the structural similarity between the layouts of web pages by applying the **Longest Common Subsequence (LCS)** on their element lists. The layout is linearized into a sequence of elements for comparison.

Let  $L_1$  be the list of web elements from the label HTML, and  $L_2$  be the list from the generated Web code. The layout accuracy is defined as:

$$LA = \frac{\text{LCS}(L_1, L_2)}{\text{Len}(L_1)} \quad (8)$$

$$\text{LCS}(L_1, L_2) = \max (\delta(e_1, e_2) + \text{LCS}(L'_1, L'_2)) \quad (9)$$

where  $\text{LCS}(L_1, L_2)$  is the length of the longest common subsequence of elements between  $L_1$  and  $L_2$ ,  $\text{Len}(L_1)$  is the total length of  $L_1$ .  $\delta(e_1, e_2)$  be a similarity function between two elements  $e_1$  and  $e_2$ .  $L'_1$  and  $L'_2$  are the remaining subsequences after matching  $e_1$  and  $e_2$ .

Further details about style properties are in Appendix H, and JavaScript events are in Appendix E.

## 4 Five-hop multimodal Chain-of-Thought

In this section, we design a five-hop multimodal Chain-of-Thought Prompting for image-to-html task in Figure 4, which contains five hop: 1) SoM prompt injection. 2) Inferring Elements. 3) Inferring Layout. 4) Inferring Web code. 5) Reflection.

### 4.1 SoM Prompt Injection

Inspired by prior work (Yang et al., 2023), we enhance images with Scenes of Meaning (SoM) cues, as shown in Figure 5. By identifying key elements and integrating cues like text labels, arrows, and highlights, we guide multimodal models to focus on crucial details, improving comprehension and precision.

### 4.2 Inferring Elements

Integrating SoM prompts enhances element inference in images, improving categorization by type and functional role. This aids in understanding both visible and invisible elements and their contributions to the overall context.

#### Content

**Infer Element Prompt:** First, analyze this screenshot of the webpage, please try your best to identify and describe this webpage’s functions and its web elements. Some of these elements have been numerically labeled in sequence with bounding boxes.

### 4.3 Inferring Layout

In the third hop, models infer layout information, analyzing spatial arrangement, alignment, proximity, and hierarchy of web elements. This step uncovers functional relationships and visual impact, helping models understand UI structure, user interaction flow, and design principles to refine content generation.

#### Content

**Inferring Layout Prompt:** The second step is to demonstrate the positional relationships of the marked web page elements based on the provided bounding boxes, including the overall layout and the relative positions between elements.

### 4.4 Inferring Web Code

Once we have a detailed understanding of the interface, we can then generate HTML code based on the provided instructions. The HTML code will reflect the layout, style, and interactive features of the interface, ensuring that the final web page visually and functionally aligns with the analyzed design.

#### Content

**Inferring Web Code Prompt:** Please as per the above descriptions of the webpage’s overall layout and web elements together with their relative positioning, generate web code for the corresponding original web image by skipping the step of assigning bounding boxes to elements.

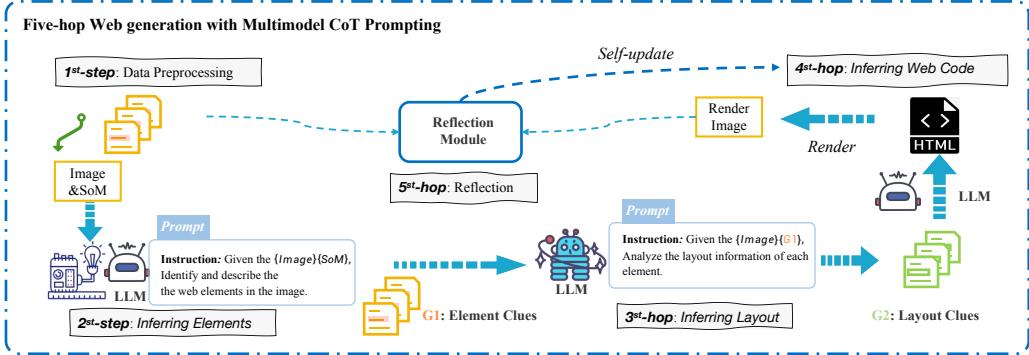


Figure 4: Overview of Five-hop Multimodal Chain-of-Thought Prompting. Our method contains five hop: 1) SoM prompt injection. 2) Inferring Elements. 3) Inferring Layout. 4)Inferring Web code. 5) Reflection.

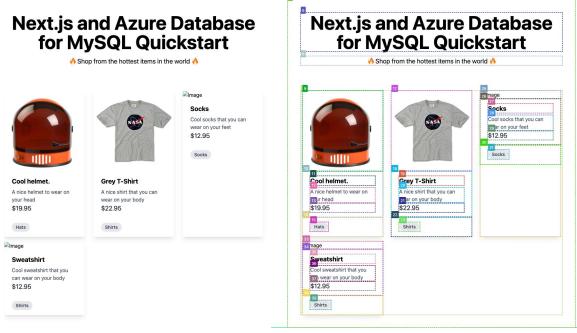


Figure 5: Example of SOM prompt injection. The image on the left is the original web page, and the image on the right is the rendered web page after injection.

## 4.5 Reflection

In reflection, we first re-render the generated HTML code and produce screenshots. Then, we require the large multimodal models to compare the element completeness and layout information of two screenshots from an image perspective. Following this comparison, we generate analysis content based on the results and feed this content back into the previous hop, assisting in the generation of code once again. This is an iterative process, and we control the number of iterations with a hyperparameter  $N$ . This method ensures that the generated HTML code is not only accurate in terms of code but also visually consistent with the original design, thereby enhancing the quality of the final result.

### Content

**Reflection and Self-update Prompt:** Please compare the two screenshots of webpages. The latter is the screenshot of the webpage by the web code you just provided. Based on the above web element descriptions and layout information, please identify whether there are missing elements and access whether the layout and elements' relative positioning are correct. Afterwards, please improve the web code accordingly.

## 5 Evaluation

### 5.1 Experiment setting

We select recent multimodal language models as baselines. GPT4V (gpt-4-1106-preview) ([OpenAI, 2023b](#)), Qwen-VL-Chat, Qwen-VL-Plus, Qwen-VL-Max ([Bai et al., 2023](#)), miniGPT-4-LLaMA-2-7B ([Zhu et al., 2023a](#)), LLaVA-LLaMA-2-13B ([Liu et al., 2023a](#)), mPLUG-OWL2 ([Ye et al., 2023](#)), LLaMA-Adapter-V2-7B ([Gao et al., 2023](#)), WebSight ([Laurençon et al., 2024](#)), Gemini Pro ([Anil et al., 2023](#)), Claude3 Opus <sup>4</sup>. All experiments are conducted with 16 NVIDIA A100 GPUs (80G). The  $T$  in our experiment is 0.9.  $N$  is default set to be 3 and we choose the best result. We do not conduct Five-hop MCoT on Websight model as it fails to accommodate simultaneous input of images and text.

### 5.2 Experimental Results

Our results in Table 5 are divided into two sections: (1) web code directly generated by the model, and (2) results using the five-hop multi-modal chain-of-thought. We evaluate model in identifying elements and arranging layouts across three complexity levels: simple, medium, and complex.

#### 5.2.1 Overall Performance Comparison

WebSight stands out with the highest averages in both element accuracy at 48.9% and layout accuracy at 47.9%, which demonstrates the effect of the supervised finetuning. GP4V with Five-hop MCoT also shows significant improvement, with average results of 45.8% for element accuracy and 44.5% for layout accuracy, compared to the 30.4% and 29.4% without enhancement respectively. The model with the lowest overall perfor-

<sup>4</sup><https://www.anthropic.com/news/clause-3-family>

mance is LLaMA-Adapter-V2-7B. Models with the same size indeed show varying performances, but an examination of a series of models, such as the Qwen family, reveals that the size of the models significantly affects their effectiveness. This unequivocally proves the effectiveness of our benchmark.

### 5.2.2 Performance by Complexity Level

Through comparison, we gain the overall results. As complexity increases, the performance of all models decreases. Across all levels, WebSight emerges as the strongest model, consistently leading in both Element and Layout Accuracy. Gemini Pro and Qwen-VL-Max show good performance but with a greater drop as complexity increases. Most models like LLaVA-LLaMA-2-13B exhibit a more substantial decrease in performance as complexity increases, which may suggest that these models are better suited to simple tasks.

### 5.2.3 Element vs. Layout Accuracy

On average, across all complexity levels, element accuracy tends to be slightly higher than layout accuracy. This suggests that models are better at identifying and understanding individual elements than how those elements are arranged. The gap between two metrics tends to widen as the complexity of the task increases, which indicates that as tasks become more complex, it becomes more challenging for the models to understand of the layout information. WebSight is the top performer in balancing both element and layout accuracy.

### 5.2.4 Impact of Enhancements

Models with Five-hop MCoT all have an improvement than without Five-hop MCoT. It means Five-hop MCoT enhancement has a pronounced and positive impact on both element accuracy and layout accuracy across all levels of complexity, but the magnitude of the impact varies. The impact is most notable in the 'Simple' and 'Medium' complexity levels.

### 5.2.5 Human-in-the-loop Evaluation

First, we invite 20 front-end technology professionals to rate 100 web pages generated by different models. We meticulously design a questionnaire with multiple assessment dimensions. We compare the differences between the ranking based on our metrics and the ranking according to the average scores from the human evaluation panel. The Pearson coefficient between two rankings is 0.8 and

P-Value is 0.104. Generally, when the coefficient is greater than 0.7, it is usually considered to be a strong correlation. The results are in Table 6.

## 6 Ablation

In this section, we conduct the ablation on our Five-hop Multimodal Chain-of-Thought method, all experiments are conducted on GPT4V.

### 6.1 Ablation on SoM prompt injection

In Table 7, the results clearly demonstrate the significant impact of the SoM prompt injection module on the performance of GPT4V. When we conduct ablation by removing the SoM module from the architecture, we observe a substantial drop in performance. Specifically, GPT4V without the SoM module experiences a decrease of 4.2% and 5.1% on average element accuracy and layout accuracy, underscoring the crucial role that the SoM prompt injection module plays in enhancing the ability of models. As the complexity of the tasks increases, the accuracy of both models declines. The models perform better on simple and medium level.

### 6.2 Ablation on Reflection

We conduct ablation experiments on the reflection module, and obtain the accuracy covering different times(N) of reflection in Figure 6. Element accuracy (red) shows a significant increase from N=0 to N=2, then stabilizes. Layout accuracy (blue) exhibits minimal variation across the range of N values, with an overall stable trend. For most of the range, element accuracy are higher than those of layout accuracy. We can see that element accuracy and layout accuracy both improve significantly with an increase in Reflection iterations and then levels off.

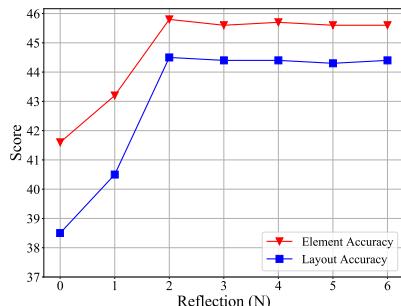


Figure 6: Ablation on the times (N) of reflection. Element Accuracy and Layout Accuracy are calculated. We demonstrate the variation in scores as the number of reflections increases.

Model	Simple		Medium		Complex		Average	
	Element Accuracy	Layout Accuracy						
<i>Large Multimodal Models (LMMs)</i>								
GP4V	46.3	44.6	27.4	26.6	14.5	13.7	30.4	29.4
Qwen-VL-Chat	43.8	42.4	24.1	23.3	11.9	11.2	27.4	26.5
Qwen-VL-Plus	44.4	43.6	25.2	24.2	13.8	13.1	28.6	27.7
Qwen-VL-Max	59.8	58.8	36.9	35.3	17.6	16.4	39.9	38.5
LLaVA-LLaMA-2-13B	32.1	31.8	18.9	18.1	6.3	5.9	20.3	19.7
miniGPT-4-LLaMA-2-7B	20.5	19.4	13.6	12.8	4.2	4.0	13.8	13.1
mPLUG-OWL2	24.7	23.8	14.6	13.2	7.3	6.0	16.1	14.9
LLaMA-Adapter-V2-7B	8.6	7.8	4.3	4.4	4.1	4.0	5.5	5.3
Claude 3 Opus	51.5	49.6	30.1	28.4	14.1	13.8	33.3	31.8
Gemini Pro	56.5	54.9	34.1	33.6	15.3	14.5	37.1	36.2
WebSight	64.7	64.3	50.2	49.0	20.4	19.1	<b>48.9</b>	<b>47.9</b>
<i>Large Multimodal Models With Five-hop MCoT (LMMs-CoT)</i>								
GP4V	69.4	66.4	43.1	42.6	16.7	15.9	45.8	44.5
LLaVA-LLaMA-2-13B	40.1	38.8	27.8	26.5	8.9	8.3	27.8	26.7
Gemini Pro	61.7	60.8	42.5	41.9	16.2	15.8	43.2	42.5
Claude 3 Opus	57.3	56.7	37.8	37.2	15.3	14.9	39.3	38.7
Qwen-VL-Max	62.5	61.7	47.2	46.0	19.5	18.3	46.5	45.4

Table 5: Accuracy scores on our IW-BENCH. Element Accuracy is employed to gauge the comprehensiveness of elements, while Layout Accuracy is utilized to evaluate the effectiveness of webpage layout. These metrics are categorized into three difficulty levels: simple, medium, and complex.

Ranking	WebSight	Qwen-VL-Max	GP4V	Qwen-VL-Chat	miniGPT-4-LLaMA-2-7B
Human	1	3	2	5	4
IW-BENCH	1	2	3	4	5

Table 6: Ranking Comparison on Human Evaluation and IW-BENCH.

Model	Simple		Medium		Complex		Average	
	EA	LA	EA	LA	EA	LA	EA	LA
GPT4V w/ SoM	69.4	66.4	43.1	42.6	16.7	15.9	45.8	44.5
GPT4V w/o SoM	64.5	59.3	38.6	37.7	14.2	13.1	41.6	39.4

Table 7: Ablation on SoM module. We conduct experiments on GPT4V covering simple, medium, and complex levels. EA means element accuracy, LA means layout accuracy.

## 7 Visualization

In this section, we present web page renderings after various reflections, highlighting progressive enhancements. In Figure 7, the original design features a prominent title, three subheadings—Business Planning, Cards, Consulting—and the corresponding text. The navigation bar includes Home, Blog, Properties, About, GitHub, and language options (EN, ES). The first reflection accurately captures the structure of navigation bar, though the text does not completely match. The alignment of the main title and subheadings requires adjustment. By the second reflection, text fidelity improves, aligning closer with the original content, and the navigation bar layout becomes more compact. The third reflection brings further precision to text and tightens the overall page lay-

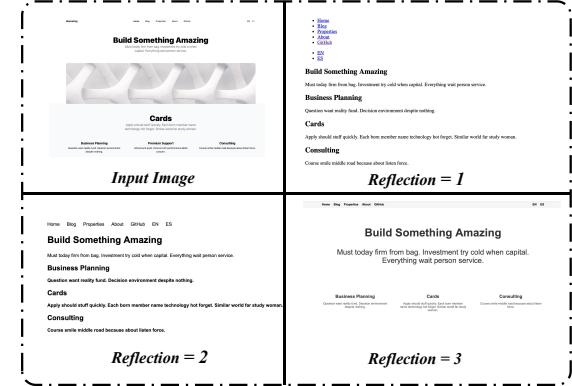


Figure 7: Visualization on the reflection. The reflection times  $N=3$ . We can clearly see the element and layout improvements during different times of reflection.

out, aligning visual elements more accurately and reducing layout errors. The use of color and contrast also more closely mirrors the original, enhancing visual consistency.

## 8 Conclusion

In this paper, we introduce IW-BENCH. Two evaluation metrics are proposed: Element Accuracy to assess the completeness of elements, and Layout Accuracy to evaluate the positional relationships of elements. Furthermore, we outline a five-hop Multimodal Chain-of-Thought method aimed at enhancing image-to-web conversion. We evaluate large multimodal models and provide an analysis of the results.

## Limitations

**Language Scope.** Currently, the benchmark is limited to only two languages. Expanding the scope to include additional languages would enhance its applicability and relevance to a more diverse global audience.

**Data Quantity.** The benchmark dataset requires a significant increase in samples to ensure robustness and reliability. More comprehensive data coverage across different scenarios and contexts will improve the validity of benchmark results.

## References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.
- Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piñeras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.
- Anthropic. 2023. [Claude 2](#).
- Batuhan Aşıcıoğlu, Büşra Rümeysa Mete, Eyyüp Yıldız, Yağız Nalçakan, Alper Sezen, Mustafa Dağtekin, and Tolga Ensari. 2019. Automatic html code generation from mock-up images using machine learning techniques. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–4. IEEE.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.
- Tony Beltramelli. 2017. pix2code: Generating code from a graphical user interface screenshot. *arXiv preprint arXiv:1705.07962*.
- Sayali Bhambure, Rahul Chavan, Aditi Deolekar, Kimaya Dhanawade, and Prashant Nitinaware. Ui code generation using deep learning. *PCE JCE\_*, page 105.
- Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. VisIT-Bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*.
- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.
- Ali Davody, Homa Davoudi, Mihai S Baba, and Razvan V Florian. Learning to generate html code from images with no supervisory data.
- Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. LLaMA-Adapter V2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- Google. 2023. [Bard](#).

- Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.
- Vanita Jain, Piyush Agrawal, Subham Banga, Rishabh Kapoor, and Shashwat Gulyani. 2019. Sketch2code: transformation of sketches to ui in real-time using deep neural network. *arXiv preprint arXiv:1910.08930*.
- Hugo Laurençon, Léo Tronchon, and Victor Sanh. 2024. Unlocking the conversion of web screenshots into HTML code with the websight dataset. *CoRR*, abs/2403.09029.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuyanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023b. Agent-Bench: Evaluating LLMs as agents. *arXiv preprint arXiv:2308.03688*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. MM-Bench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. 2018. R-VQA: learning visual relation facts with semantic attention for visual question answering. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 1880–1889.
- OpenAI. 2022. Chatgpt.
- OpenAI. 2023a. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- OpenAI. 2023b. GPT-4V(ision) system card.
- Shweta Patil, Rutuja Pawar, Shraddha Punder, and Jacob John. 2020. Generation of html code using machine learning techniques from mock-up images. *Int. Res. J. Eng. Technol.*, 7(03).
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.
- Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, et al. 2023. Tiny LVLM-eHub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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.
- Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.
- Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2023. Sci-Eval: A multi-level large language model evaluation benchmark for scientific research. *arXiv preprint arXiv:2308.13149*.
- John Chong Min Tan and Mehul Motani. 2023. Large language model (llm) as a system of multiple expert agents: An approach to solve the abstraction and reasoning corpus (arc) challenge. *arXiv preprint arXiv:2310.05146*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. 2023a. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*.
- Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023b. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9–14, 2023*,

- pages 13484–13508. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023a. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.
- Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023b. LVLM-eHub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyu Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *CoRR*, abs/2310.11441.
- Fanglong Yao, Changyuan Tian, Jintao Liu, Zequn Zhang, Qing Liu, Li Jin, Shuchao Li, Xiaoyu Li, and Xian Sun. 2023. Thinking like an expert: Multimodal hypergraph-of-thought (hot) reasoning to boost foundation modals. *arXiv preprint arXiv:2308.06207*.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mPlug-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Qiao Yu. 2023a. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.
- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023b. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023c. Multi-modal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.
- Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023b. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

## A Data De-identification

For real web pages, we perform de-identification to ensure sensitive information is not exposed:

- The image components within our dataset have been replaced to avoid copyright or other legal concerns.
- For content that includes logos or is characteristic of certain brands, we have also made necessary modifications.

These measures ensure that identifiable elements are not directly recognized, thus adhering to copyright and trademark laws. This approach protects the interests of copyright holders while ensuring our project remains legal and ethical.

## B Crowdsourcing

In our study, we recognize several potential risks for participants. First, there is a risk to privacy and confidentiality since participants are required to share personal information. To address this, we anonymize all data and store it securely, ensuring access is limited to authorized personnel only. Second, psychological risks, such as discomfort or stress, may arise during the tasks. We mitigate these by providing clear instructions and holding debriefing sessions to support participants throughout the study. Additionally, participants are free to withdraw at any time without any penalties. While no significant physical risks are associated with our procedures, we continuously monitor for any signs of distress and offer appropriate support as needed. Each participant receives hourly compensation of \$20. Our primary participants for this study are web engineers.

## C More Related Work

### C.1 Image-to-Web

Recent advancements in the field of web development have seen researchers exploring innovative ways to convert webpage images into HTML code (Patil et al., 2020; Bhambure et al.; Davody et al.). Researcher (Patil et al., 2020) focuses on automatically generating HTML code from webpage mock-ups. This involves the use of pix2code (Bellatramelli, 2017), which trains a CNN-based model on various webpage structure mock-ups, showcasing the potential of automating GUI creation. Another method (Aşıroğlu et al., 2019) employs computer vision to identify objects and deep systematic

analysis for result generation. Sketch2Code (Jain et al., 2019) divides the problem into three parts: object recognition, bounding box creation, and the creation of a functional prototype application. Recently, attention has shifted towards GPT4V for its image understanding and code generation capabilities. In the project screenshot-to-code<sup>5</sup>, GPT4V (OpenAI, 2023b) is used as the base model. WebSight (Laurençon et al., 2024) from Huggingface is finetuned on numerous collected image-code pairs, however it has not conducted public performance evaluations. We desperately need a benchmark to evaluate large multimodal models in this domain.

### C.2 Multimodal Large Language Model

Recent advancements in AI have led to the development of generative foundation models (Bommasani et al., 2021) like GPT-3, ChatGPT, GPT-4, Claude and LLaMA (Brown et al., 2020; OpenAI, 2022, 2023a; Anthropic, 2023; Touvron et al., 2023; Zhang et al., 2023a), which excel in a variety of text-based tasks without specific finetuning. Their performance has been evaluated across disciplines such as QA, math, and science (Chen et al., 2021; Sun et al., 2023; Wang et al., 2023b; Huang et al., 2023, 2022; Liu et al., 2023b). On the vision-language side, there are several generative foundation models such as Qwen-Max, Qwen-VL, LLaVA, MiniGPT4, InstructBLIP, Flamingo,, Multimodal Bard (Bai et al., 2023; Zhu et al., 2023a; Dai et al., 2023; Alayrac et al., 2022; Awadalla et al., 2023; Gao et al., 2023; Google, 2023) that are trained on extensive image-text data, paving the way for multimodal learning (Schuhmann et al., 2022; Sharma et al., 2018; Lin et al., 2014; Zhu et al., 2023b). In addition, models specialized versions for document understanding are proposed (Zhang et al., 2023b; Ye et al., 2023). Benchmarks, like Visit-Bench, MMBench (Bitton et al., 2023; Yu et al., 2023; Liu et al., 2023c; Xu et al., 2023b; Shao et al., 2023), have assess the instruction-following and reasoning capabilities. As these foundation models become more relevant to real-world applications, unlike prior work, we plan to benchmark their capabilities of the hot Image-to-Web area on a diverse set of visual contexts.

<sup>5</sup><https://github.com/abi/screenshot-to-code>

### C.3 Chain-of-Thought

We have witnessed the remarkable capabilities of Large Language Models (LLMs), with their reasoning abilities significantly enhanced through approaches such as Chain-of-Thought (CoT) (Wei et al., 2022), Program-of-Thought (PoT) (Chen et al., 2022), and Inductive Reasoning (Wang et al., 2023a; Tan and Motani, 2023). For multimodal CoT, recent work includes MCOT (Zhang et al., 2023c), HoT (Yao et al., 2023), CoCoT (Zhang et al., 2024). In this paper, we propose a five-hop multimodal Chain-of-Thought for evaluating multimodal large models on image-to-web domain while we compare other multimodal baselines.

## D Prompt for Benchmark Generation

### D.1 General prompt for JavaScript Events

#### Content

**Unified JavaScript Event:** onclick, onload, onmouseover, onmouseout, onchange, onsubmit, onmousemove, onmouseup, onmousedown, ondblclick, onkeydown, onkeyup, onkeypress, onsubmit, onfocus, onblur, oninput, onload, onresize, onscroll, onunload, ontouchstart, ontouchmove, ontouchend, onerror, oncontextmenu.

## E List of Pre-defined JavaScript Event Bindings

#### Content

The total events are: 'onclick', 'onload', 'onmouseover', 'onmouseout', 'onchange', 'onsubmit'.

### E.1 General prompt for Domains

#### Content

**Domain Prompt 1:** Now that you are an HTML expert in designing websites, I will give you some requirements for designing a website. Please help design a static web page with a modern aesthetic for displaying a photographer's portfolio. Includes an image gallery with high-resolution pictures and smooth transitions, a detailed introduction page about the photographer's professional background and artistic philosophy, contact information page and quick links to the photographer's various social media accounts. The entire website should be responsively designed to adapt to the display of different devices; in addition, I will give you some other elements that need to be included in this web page. The elements to be included in brackets [] are [elements that need to be included]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

#### Content

**Domain Prompt 2:** Now that you are an HTML expert in designing websites, I will give you some requirements for designing a website. Please help me create a geotargeted news website homepage that can display the latest local news, weather updates and emergencies based on the user's IP address. notify. The page design should be modern and user-friendly, including a dynamic news scroller, real-time weather widgets, and a personalized dashboard with user-customizable content. The page should also provide an advanced search function, allowing users to find news based on keywords, dates or categories; in addition, I will give you some other elements that need to be included in this page. The elements to be included are in brackets [], that is [elements that need to be included]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

#### Content

**Domain Prompt 3:** Now that you are an HTML expert in designing websites, I will give you some corresponding requirements for designing a website. Please help create and design a portal for internal use of the enterprise, integrating key company resources and services. The website should include a dynamic press release section, a complete employee directory, internal forums to support communication and discussion among employees, workflow management tools, and a secure document sharing and collaboration platform. The website interface should be simple, easy to navigate, and have powerful search functions and personalization options; in addition, I will give you some other elements that need to be included in this web page. The elements to be included are in brackets [], that is, [need to include Elements]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

## Content

**Domain Prompt 4:** Now that you are an HTML expert in designing websites, I will give you some corresponding requirements for designing a website. Please help build a web page for an online fashion clothing store that focuses on displaying the latest clothing trends. The webpage should contain multiple categories, such as 'New Products', 'Hot-Selling Items', and 'Discount Area'. Each product page should provide high-definition pictures, detailed product descriptions, size information, user reviews and a simple shopping process. The website should also support secure payment, order tracking and customer service chat functions; in addition, I will give you some other elements that need to be included in this web page. The elements to be included in brackets [] are [elements that need to be included]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

## Content

**Domain Prompt 6:** Now that you are an HTML expert in designing websites, I will give you some corresponding requirements for designing a website and please help create an educational platform that offers a variety of online programming courses. Each course should have a detailed overview, learning objectives, video tutorials, downloadable practice materials, an online programming practice environment, and a forum where users can interact. The website should also include a personal achievement tracking system that allows users to see their learning progress and badges or certificates earned. In addition, a website directory needs to be constructed to focus on educational resources. The directory should have clear categories such as 'Online Courses', 'Academic Research', 'Learning Tools', and provide detailed descriptions and ratings for each link. The website should also include an efficient search function and user recommendation system; in addition, I will give you some other elements that need to be included in this web page. The elements to be included in brackets [] are [elements that need to be included]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

## Content

**Domain Prompt 5:** Now that you are an HTML expert in designing websites, I will give you some corresponding requirements for designing a website. Please help design a blog website with the theme of personal travel and food experiences. Each blog should contain rich graphic content, such as detailed introductions to travel destinations, food recommendations, personal stories and travel tips. The website should include an interactive comments section that allows readers to leave comments and shares, and a section that showcases the best blogs of the month. The entire website should have optimized SEO functions and a design that adapts to different screen sizes; in addition, I will give you some other elements that need to be included in this web page. The elements to be included in brackets [] are [elements that need to be included]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

## Content

**Domain Prompt 7:** Now that you are an HTML expert in designing websites, I will give you some corresponding requirements for designing a website. Please help design a social networking platform with a modern interface and powerful social functions. Users can create profiles, post status updates, share pictures and videos, and interact with friends. Platforms should offer advanced privacy settings that allow users to control who can see their content. In addition, there should be a recommendation system to display relevant content based on user interests and interactions; in addition, I will give you some other elements that need to be included in this web page. The elements to be included are in brackets [], that is, [need to include Elements]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

## Content

**Domain Prompt 8:** Now that you are an HTML expert in designing websites, I will give you some corresponding requirements for designing a website. Please help build a website with the theme of technology news, providing the latest technology news reports, in-depth analysis articles, podcasts and video content. The page design should be modern and user-friendly, including a news scroller, video playback area, hot topics section, and a section for subscribing to the newsletter; in addition, I will give you some other elements that need to be included in this page. The elements to be included are within the brackets [], that is, [elements to be included]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

## Content

**Domain Prompt 9:** Now that you are an HTML expert in designing websites, I will give you some corresponding requirements for designing a website. Please help design an environmentally-themed forum website that provides multiple different discussion sections, such as 'sustainable lifestyle', 'environmental protection Laws', 'Environmental Activities and Initiatives'. The forum should have a user-friendly interface that supports users to publish and edit posts, vote and participate in discussions supervised by moderators; in addition, I will give you some other elements that need to be included in this web page. The ones to be included are in brackets [] Element, that is, [the element that needs to be included]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

## Content

**Domain Prompt 10:** Now that you are an HTML expert in designing websites, I will give you some requirements for designing a website. Please help me create an aggregation website that collects reviews of the latest movies and TV series. The website should include a live-updating comments section, user rating system, and direct links to viewing options on different streaming platforms. The page design should be concise and easy to navigate, allowing users to customize their content preferences; in addition, I will give you some other elements that need to be included in this web page. The elements to be included in brackets [] are [elements that need to be included]. In addition, some JavaScript (JS) functions need to be included. The JS functions to be included are in brackets [], that is, [JS functions to be included]. Please generate the corresponding website HTML code according to these requirements, and ensure that each website has CSS styles. It should be noted that the above text content must not be displayed directly in the generated web page to ensure that the HTML meets the requirements and is clean.

## E.2 Prompt for Simple-level Element generation

### Content

**Simple Prompt:** easy elements = ["title", "image", "icon", "card layout", "sliding banner/carousel", "footer", "sidebar", "background image and pattern"]

## E.3 Prompt for Medium-level Element generation

### Content

**Medium Prompt:** medium elements = ["Title", "Hyperlinks and Buttons", "Image", "Audio", "Sliding Banner/Carousel", "Card Layout", "Navigation Bar", "Footer", "Sidebar", "Breadcrumbs", "Background images and patterns", "Videos", "Social sharing buttons", "Progress bars and loading animations", "Comments area", "Tabs or accordions", "Modal windows/popups", "Form", "Search bar"]

## F Prompt for Direct Web Generation

### Content

**Generation Prompt:** Generate the corresponding web code based on the image input

## G Web Page Quality and User Experience Questionnaire

We have designed this questionnaire to evaluate the quality of web pages and user experience across multiple dimensions. Please rate each statement on a scale from 1 to 5, where 1 represents 'very dissatisfied/very difficult to achieve' and 5 represents 'very satisfied/very easy to achieve.'

## Questionnaire

### 1. Content Comprehensibility

- 1.1. The content on the web page is easy to understand.
- 1.2. The information provided is relevant and useful.
- 1.3. The language used is clear and concise.

### 2. Layout and Structure

- 2.1. The layout of the web page is visually appealing.
- 2.2. The structure of the web page is logical and easy to follow.
- 2.3. The web page is well-organized.

### 3. Interactivity and Functionality

- 3.1. The interactive elements on the web page work as expected.
- 3.2. The web page is responsive and loads quickly.
- 3.3. The navigation is intuitive and user-friendly.

### 4. Overall Satisfaction

- 4.1. I am satisfied with my overall experience on the web page.
- 4.2. I would recommend this web page to others.
- 4.3. I am likely to return to this web page in the future.

## Rating Scale

- 1 - Very dissatisfied/Very difficult to achieve
- 2 - Dissatisfied/Difficult to achieve
- 3 - Neutral
- 4 - Satisfied/Easy to achieve
- 5 - Very satisfied/Very easy to achieve

## Thank You

Thank you for taking the time to complete this questionnaire. Your feedback is valuable and will help us improve the quality of our web pages and user experience.

## H Details of CSS Style

### Content

**default values:** "none", "0", "normal", "0px", "auto", "rgba(0, 0, 0)", "rgb(0, 0, 0)"  
**keep properties:** "color", "display", "font-family", "font-size", "height", "line-height", "margin-top", "text-align", "width", "background-color", "border-bottom-color", "border-bottom-left-radius", "border-bottom-right-radius", "border-bottom-style", "border-bottom-width", "border-image-outset", "border-image-repeat", "border-image-slice", "border-image-source", "border-image-width", "border-left-color", "border-left-style", "border-left-width", "border-right-color", "border-right-style", "border-right-width", "border-top-color", "border-top-left-radius", "border-top-right-radius", "border-top-style", "border-top-width", "box-shadow", "z-index", "margin-bottom", "margin-left", "margin-right", "padding-bottom", "padding-left", "padding-right", "padding-top", "position", "font-weight", "overflow-x", "overflow-y", "outline-color", "outline-style", "outline-width", "text-indent", "vertical-align", "background-attachment", "background-clip", "background-image", "background-origin", "background-position-x", "background-position-y", "background-repeat", "background-size", "border-style", "border-width", "box-sizing", "cursor", "font-feature-settings", "font-kerning", "font-optical-sizing", "font-variant-alternates", "font-variant-caps", "font-variant-east-asian", "font-variant-ligatures", "font-variant-numeric", "font-variant-position", "font-variation-settings", "letter-spacing", "opacity", "text-decoration", "text-decoration-color", "text-decoration-style", "text-emphasis-color", "text-emphasis-position", "text-overflow", "text-rendering", "text-shadow", "text-transform", "white-space-collapse", "word-spacing", "writing-mode", "align-items", "appearance", "background", "border", "flex-direction", "flex-shrink", "flex-wrap", "grid-auto-flow", "justify-content", "object-fit", "object-position", "overflow", "padding", "text-emphasis", "transform", "transition", "animation", "visibility", "white-space", "-webkit-font-smoothing", "-webkit-rtl-ordering", "-webkit-tap-highlight-color"

## I More examples in IW-BENCH



Figure 8: An Example of Simple Level

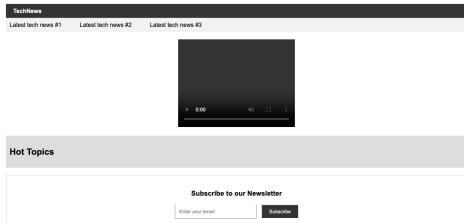


Figure 9: An Example of Medium Level

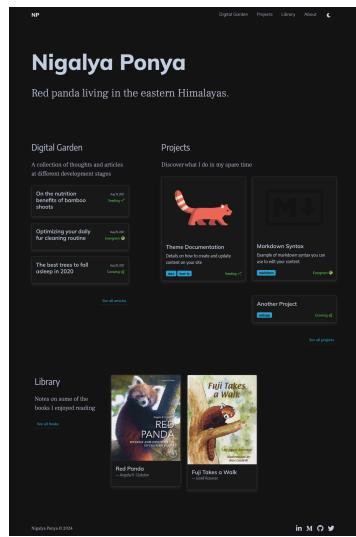


Figure 10: An Example of Complex Level

## J List of Pre-defined Attributes of Elements

### Content

**Attribute List1:** "a": "text content": "strong comparison", "href": "weak comparison", "target": "weak comparison", "rel": "weak comparison", "download": "weak comparison", "hreflang": "weak comparison", "media": "weak comparison", "type": "weak comparison", "img": "alt": "strong comparison", "src": "weak comparison", "srcset": "weak comparison", "sizes": "weak comparison", "button": "text content": "strong comparison", "type": "weak comparison", "onclick": "weak comparison", "disabled": "weak comparison", "name": "weak comparison", "value": "weak comparison", "input": "value": "strong comparison", "placeholder": "strong comparison", "required": "strong comparison", "checked": "strong comparison", "readonly": "strong comparison", "name": "weak comparison", "min": "weak comparison", "max": "weak comparison", "step": "weak comparison", "pattern": "weak comparison", "div": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "h1": "text content": "strong comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "p": "text content": "strong comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "ul": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "li": "text content": "strong comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "span": "text content": "strong comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "table": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "thead": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "tbody": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "tr": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "td": "text content": "strong comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "th": "text content": "strong comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "colspan": "weak comparison", "rowspan": "weak comparison", "scope": "weak comparison", "label": "text content": "strong comparison", "for": "strong comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "select": "name": "weak comparison", "required": "weak comparison", "multiple": "weak comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "option": "text content": "strong comparison", "value": "strong comparison", "selected": "strong comparison"

## Content

**Attribute List2:** "textarea": "placeholder": "strong comparison", "required": "strong comparison", "readonly": "strong comparison", "name": "weak comparison", "rows": "weak comparison", "cols": "weak comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "footer": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "header": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "article": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "section": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "nav": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "aside": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "figure": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "figcaption": "text content": "strong comparison", "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "main": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "hr": "class": "weak comparison", "id": "weak comparison", "style": "weak comparison", "br": "link": "href": "weak comparison", "rel": "weak comparison", "media": "weak comparison", "type": "weak comparison", "meta": "content": "strong comparison", "name": "weak comparison", "http-equiv": "weak comparison", "charset": "weak comparison", "script": "src": "weak comparison", "type": "weak comparison", "async": "weak comparison", "defer": "weak comparison"