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

Sukmin Yun[*,1,4], Haokun Lin[*,1], Rusiru Thushara[*,1], Mohammad Qazim Bhat[*,1],Yongxin Wang[* ,1], Zutao Jiang[1], Mingkai Deng[2], Jinhong Wang[1], Tianhua Tao[1,3],Junbo Li[1], Haonan Li[1], Preslav Nakov[1], Timothy Baldwin[1], Zhengzhong Liu[1,5], Eric P. Xing[1,2,5], Xiaodan Liang[1], Zhiqiang Shen[1]

[1]MBZUAI, [2]CMU, [3]UIUC, [4]HYU ERICA, [5]Petuum

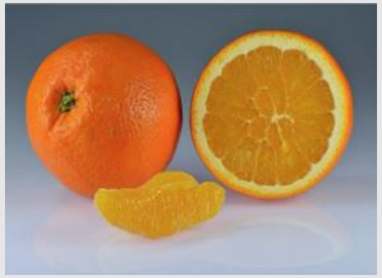
Introduction: Datasets for Multimodal LLMs

- Previous dataset (e.g., WebSight)
  - Lack instruction information
  - Do not suitable for general MLLM

- Popular benchmarks (e.g., MMBench)
  - Evaluate in isolation
  - Do not fully integrate visual information

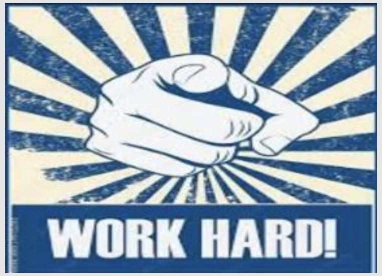
Examples of MMBench

Attribute recognition



Q: what is the color of this object?
A.Purple
B.Pink
C.Gray
D.Orange
GT: D

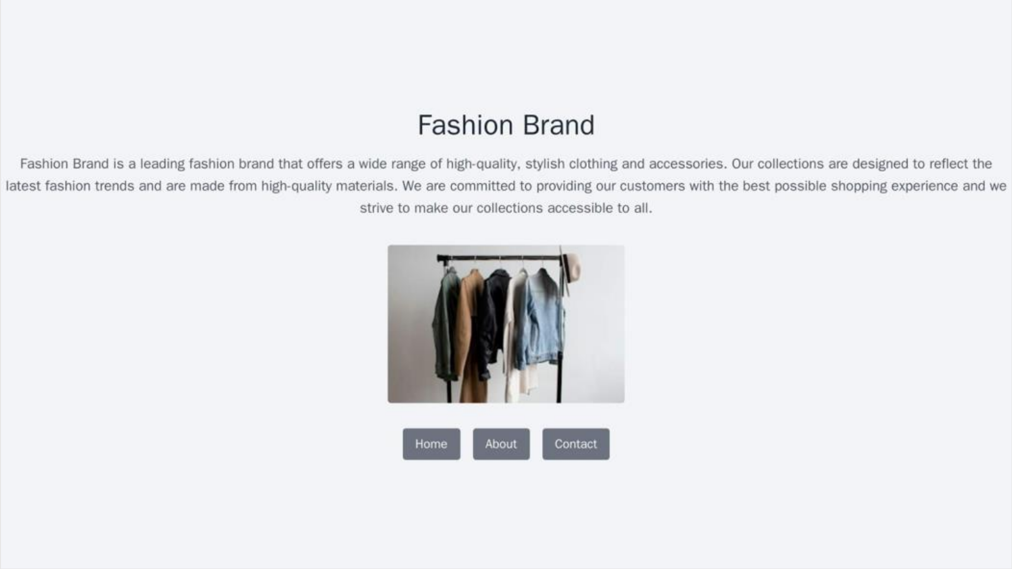
OCR



Q: What does this picture want to express?
A.We are expected to care for green plants.
B.We are expected to care for the earth.
C.We are expected to stay positive.
D.We are expected to work hard.
GT: D

An example of WebSight dataset

image



Fashion Brand

Fashion Brand is a leading fashion brand that offers a wide range of high-quality, stylish clothing and accessories. Our collections are designed to reflect the latest fashion trends and are made from high-quality materials. We are committed to providing our customers with the best possible shopping experience and we strive to make our collections accessible to all.

text

```
<html> <link
href="https://cdn.jsdelivr.net/npm/tailwindcss@2.2.19/dist/tailwind.min.css"
rel="stylesheet"> <body class="bg-gray-100"> <div class="flex flex-col items-
center justify-center h-screen">
```

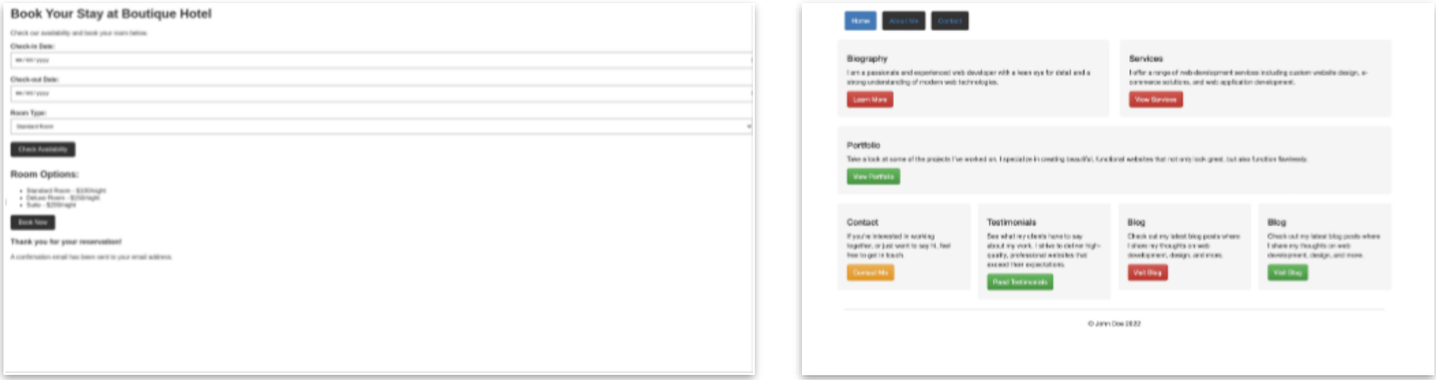
Llm_generated_idea

Fashion Brand: A visually stunning layout with a full-width, rotating image carousel showcasing their latest collections, a bold, center-aligned logo, and a bottom navigation menu. The color palette is inspired by the latest fashion trends.

Introduction: Webpage-to-Code Generation

Webpage(image)-to-code Generation Task

- Existing MLLMs still have difficulty understanding web page screenshots
- The web pages restored by the generated html code is very different from the original web page



Book Your Stay at Boutique Hotel

Check our availability and book your room below.

Check-in Date: mm/dd/yyyy

Check-out Date: mm/dd/yyyy

Room Type: Standard Room

Room Options:

- Standard Room - $100/night
- Deluxe Room - $150/night
- Suite - $200/night

Thank you for your reservation!

A confirmation email has been sent to your email address.

Biography

I am a passionate and experienced web developer with a keen eye for detail and a strong understanding of modern web technologies.

Services

I offer a range of web development services including custom website design, e-commerce solutions, and web application development.

Portfolio

Take a look at some of the projects I've worked on. I specialize in creating beautiful, functional websites that not only look great, but also function flawlessly.

Contact

If you're interested in working together, or just want to say hi, feel free to get in touch.

Testimonials

See what my clients have to say about my work. I strive to deliver high-quality, professional websites that exceed their expectations.

Blog

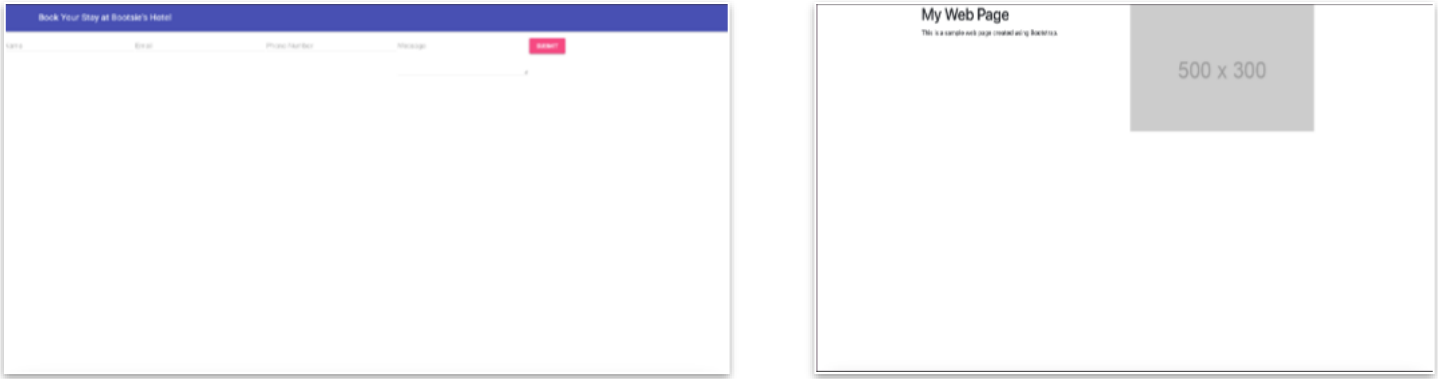
Check out my latest blog posts where I share my thoughts on web development, design, and more.

Blog

Check out my latest blog posts where I share my thoughts on web development, design, and more.

© John Doe 2022

Original webpages



Book Your Stay at Bootsie's Hotel

My Web Page

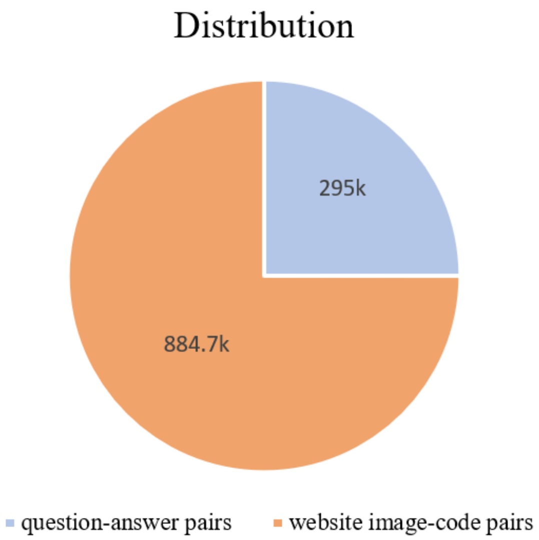
This is a sample web page created using Bootstrap.

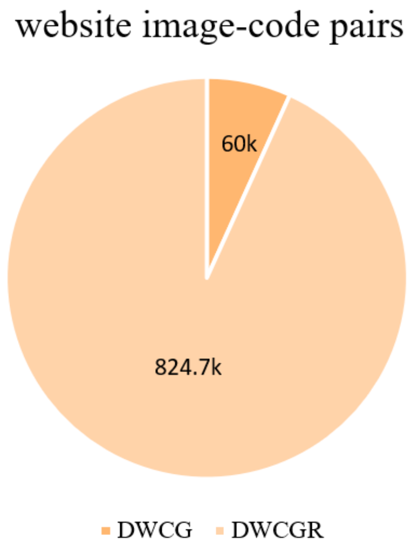
Trained on LLaVA dataset only

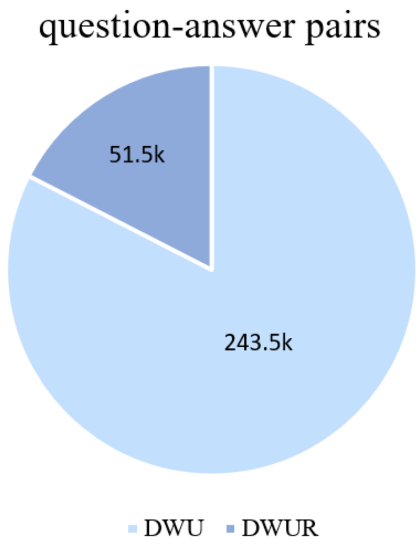
Web2Code: Overall distribution

We propose a **Large-scale Webpage-to-Code Dataset** and **Evaluation Framework** for MLLMs

- Web2Code contains a total of 1179.7k web-based instruction-response pairs
- These pairs include Question-Code pairs and Questions-Answer pairs

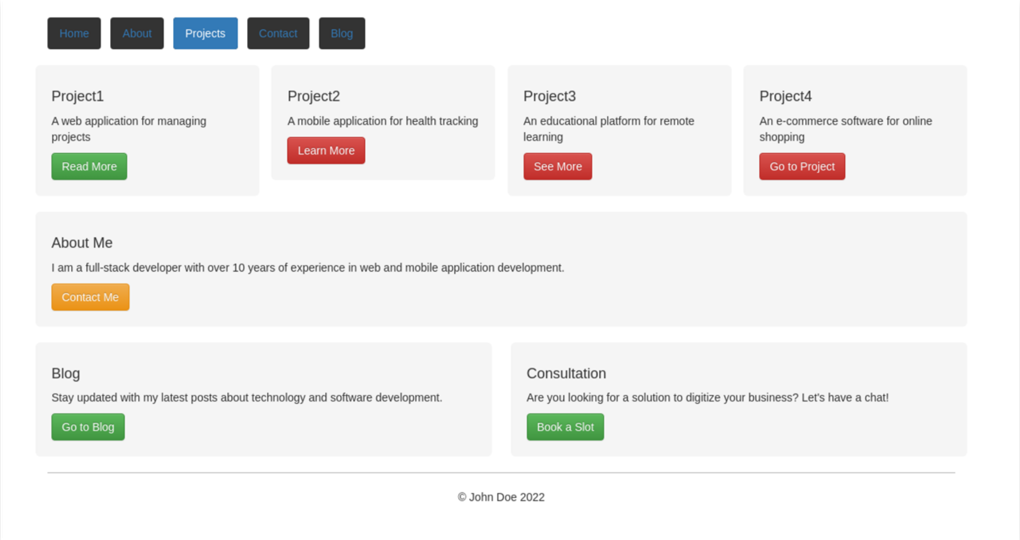






Web2Code: Examples

Website image–code pairs



Project1

A web application for managing projects

Project2

A mobile application for health tracking

Project3

An educational platform for remote learning

Project4

An e-commerce software for online shopping

About Me

I am a full-stack developer with over 10 years of experience in web and mobile application development.

Blog

Stay updated with my latest posts about technology and software development.

Consultation

Are you looking for a solution to digitize your business? Let's have a chat!

© John Doe 2022

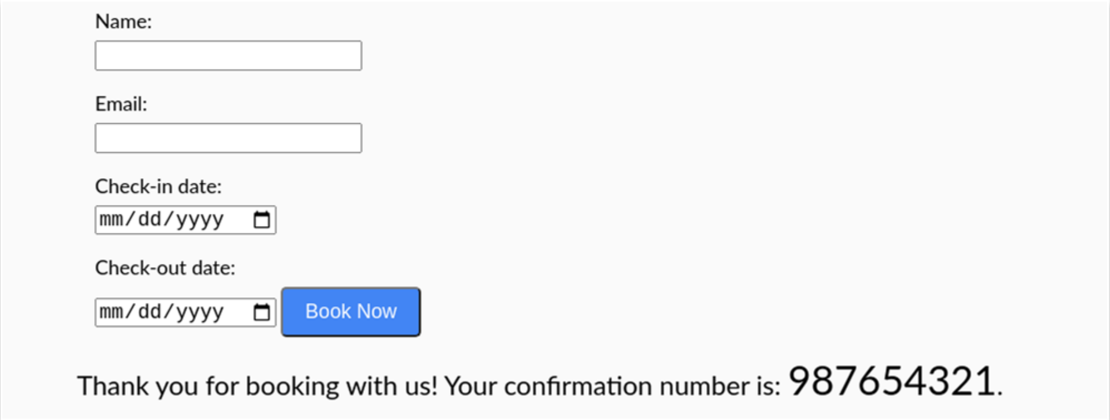
Instruction:

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

Code:

```
<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/boo
tstrap.min.css\" integrity=\"sha384……
```

Question answer pairs


Name:

Email:

Check-in date:

mm/dd/yyyy

Check-out date:

mm/dd/yyyy

Thank you for booking with us! Your confirmation number is: 987654321.

Question 1:

<image>\nWhat is the main purpose of this website as indicated on the page?

Answer:

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.

Question 2:

Describe the colors utilized for the submit button, both in its default state and upon hovering.

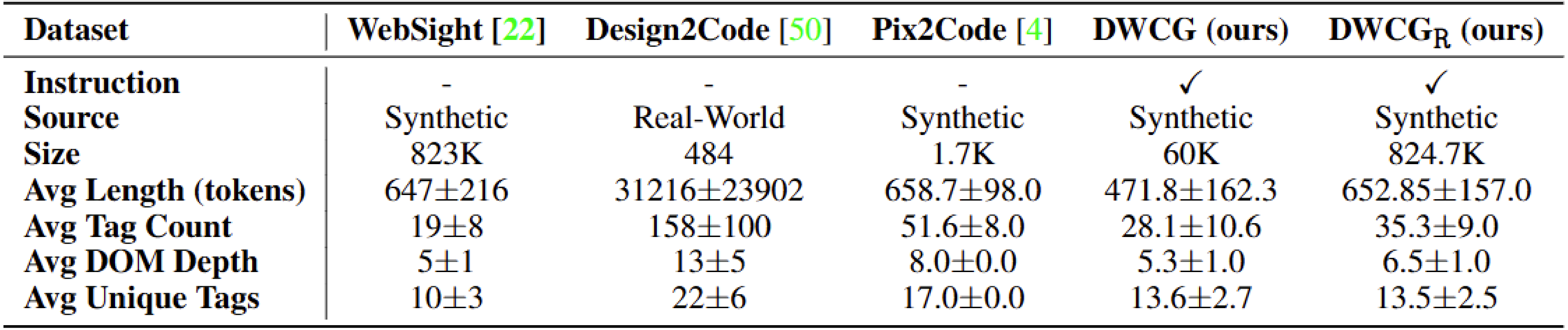
Answer:

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 #366BC5, which is a darker shade of blue.

Web2Code: Statistics

Our dataset

- has larger samples for webpage code generation
- includes more complex interactions
- is more suitable for developing robust models across diverse web-based tasks



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

Comparison of dataset statistics among webpage code generation datasets: WebSight, Design2Code, Pix2Code, our DWCG, and our DWCGR. DWCG is a newly generated GPT-3.5-baseddataset, while DWCGR is the refined dataset that utilizes WebSight and Pix2Code datasets.

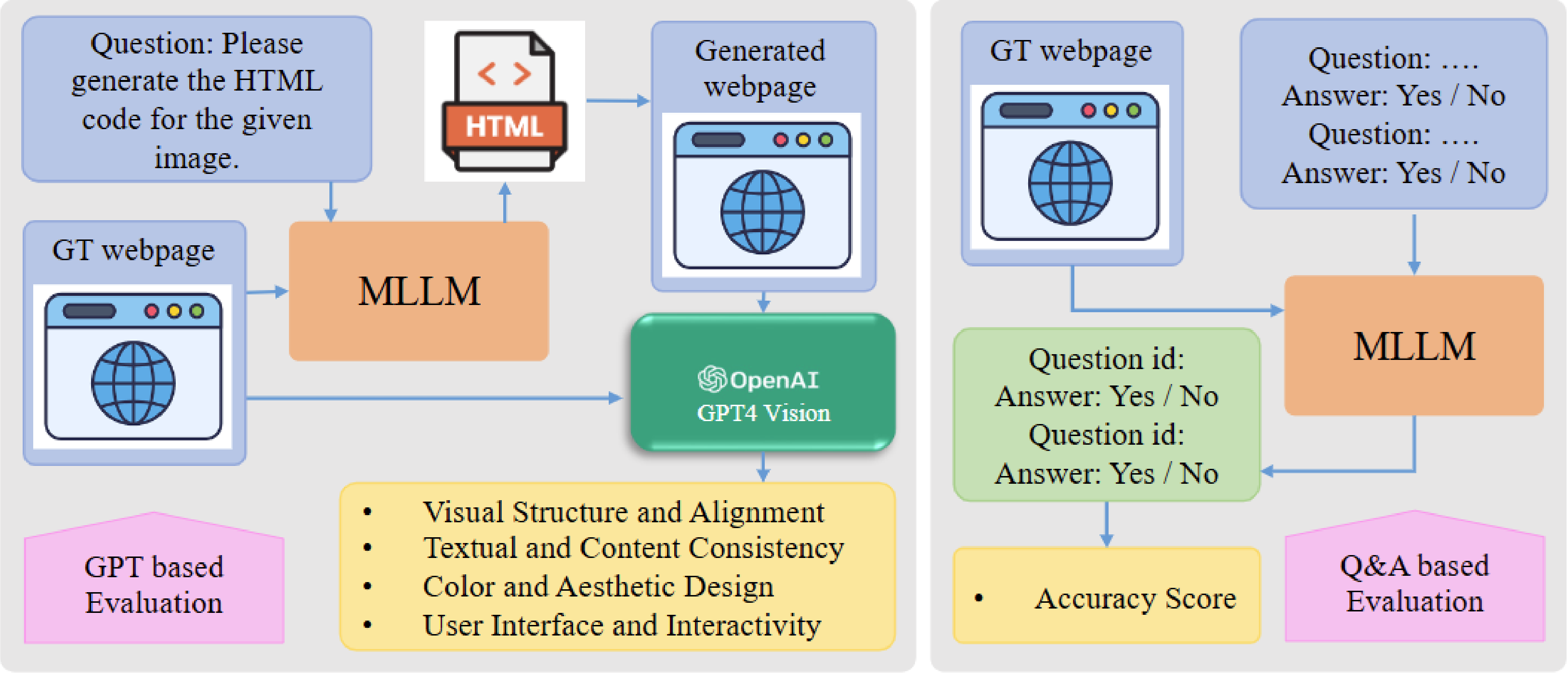
Web2Code: Evaluation Framework

Webpage Code Generation Benchmark (WCGB):

- This benchmark sets a series of tasks to generate HTML code from web page images, using GPT to evaluate the consistency of images recovered from HTML with real images

Web Understanding Benchmark (WUB):

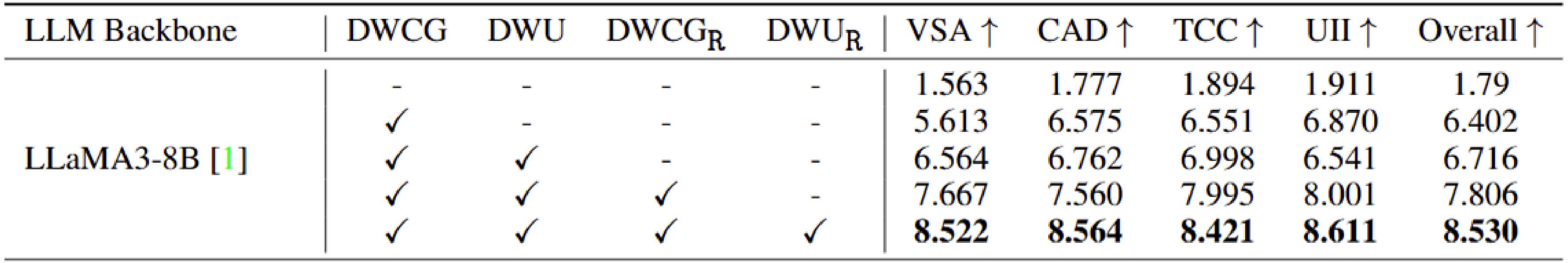
- The benchmark sets a series of web image QA tasks to detect the accuracy of the predicted answers



Webpage **C**ode **G**eneration **B**enchmark (**WCGB**) **W**ebpage **U**nderstanding **B**enchmark (**WUB**)

Web2Code: Experimental Results

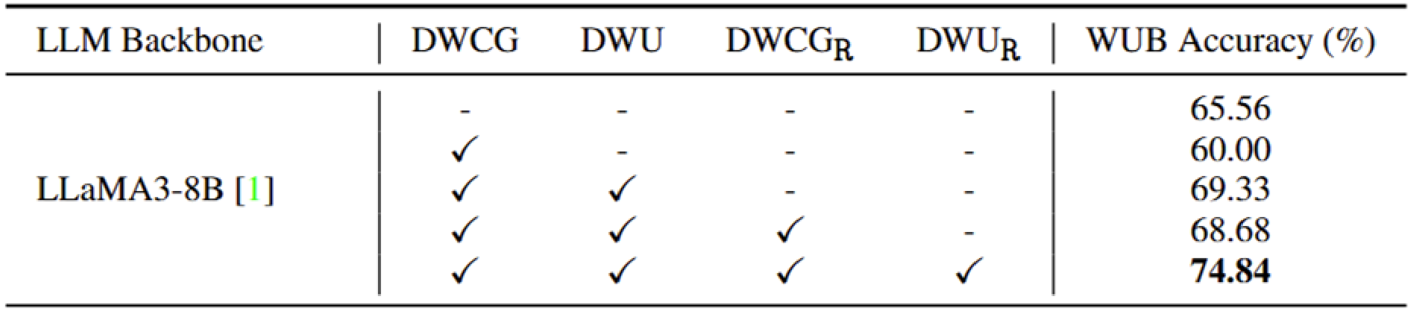
- **WCGB:** Our results show the improvement in the quality of webpage code generation when we incrementally add Web2Code sub-datasets: +DWCG, +DWU,+DWCG$_R$, and +DWU$_R$



LLM Backbone	DWCG	DWU	DWCG$_R$	DWU$_R$	VSA ↑	CAD ↑	TCC ↑	UII ↑	Overall ↑
LLaMA3-8B [1]	-	-	-	-	1.563	1.777	1.894	1.911	1.79
	✓	-	-	-	5.613	6.575	6.551	6.870	6.402
	✓	✓	-	-	6.564	6.762	6.998	6.541	6.716
	✓	✓	✓	-	7.667	7.560	7.995	8.001	7.806
	✓	✓	✓	✓	**8.522**	**8.564**	**8.421**	**8.611**	**8.530**

Performance comparison 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.

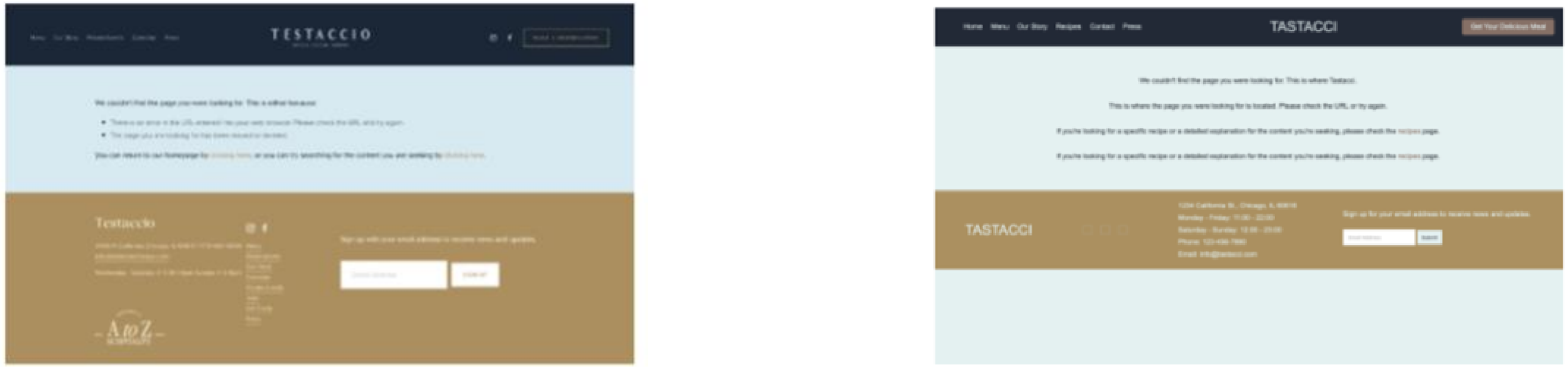
- **WUB:** Our results demonstrate the effectiveness of the proposed dataset in web page comprehension



LLM Backbone	DWCG	DWU	DWCG$_R$	DWU$_R$	WUB Accuracy (%)
LLaMA3-8B [1]	-	-	-	-	65.56
	✓	-	-	-	60.00
	✓	✓	-	-	69.33
	✓	✓	✓	-	68.68
	✓	✓	✓	✓	**74.84**

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.

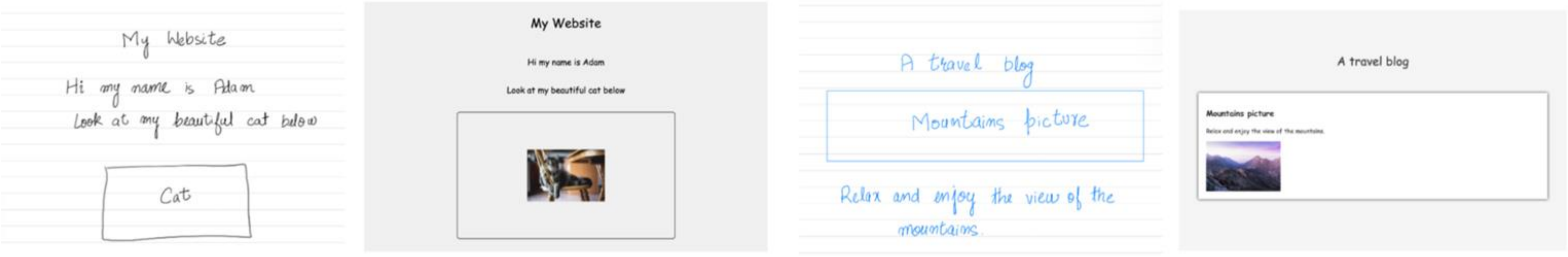
Web2Code: Qualitative Examples



Original

CrystalChat-7B

Visualization comparison using different backbones. The code-enhanced LLM backbone CrystalChat-7B achieves generation quality close to the original image.


My Website
Hi my name is Adam
Look at my beautiful cat below
Cat
My Website
Hi my name is Adam
Look at my beautiful cat below
A travel blog
Mountains picture
Relax and enjoy the view of the mountains.
A travel blog
Mountains picture
Relax and enjoy the view of the mountains.

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

Thanks for your attention!

Github repo

Github page