SEED-Bench-2-Plus: Benchmarking Multimodal Large Language Models with Text-Rich Visual Comprehension

Bohao Li[3,1] Yuying Ge[1†] Yi Chen[1] Yixiao Ge[1,2] Ruimao Zhang[3†] Ying Shan[1,2]

[1]Tencent AI Lab
[2]ARC Lab, Tencent PCG
[3]School of Data Science, The Chinese University of HongKong, Shenzhen

Abstract

Comprehending text-rich visual content is paramount for the practical application of Multimodal Large Language Models (MLLMs), since text-rich scenarios are ubiquitous in the real world, which are characterized by the presence of extensive texts embedded within images. Recently, the advent of MLLMs with impressive versatility has raised the bar for what we can expect from MLLMs. However, their proficiency in text-rich scenarios has yet to be comprehensively and objectively assessed, since current MLLM benchmarks primarily focus on evaluating general visual comprehension. In this work, we introduce SEED-Bench-2-Plus, a benchmark specifically designed for evaluating ***text-rich visual comprehension*** *of MLLMs. Our benchmark comprises 2.3K multiple-choice questions with precise human annotations, spanning three broad categories: Charts, Maps, and Webs, each of which covers a wide spectrum of text-rich scenarios in the real world. These categories, due to their inherent complexity and diversity, effectively simulate real-world text-rich environments. We further conduct a thorough evaluation involving 34 prominent MLLMs (including GPT-4V, Gemini-Pro-Vision and Claude-3-Opus) and emphasize the current limitations of MLLMs in text-rich visual comprehension. We hope that our work can serve as a valuable addition to existing MLLM benchmarks, providing insightful observations and inspiring further research in the area of text-rich visual comprehension with MLLMs. The dataset and evaluation code can be accessed at https://github.com/AILab-CVC/SEED-Bench.*

1. Introduction

In recent years, through leveraging the strong generality of Large Language Models (LLMs) [8, 12, 37, 38, 43], Multi-modal Large Language Models (MLLMs) [1, 4, 9, 17, 19, 21, 24–26, 29, 30, 35, 36, 39, 39, 40, 42, 50, 56, 58] have showcased remarkable capabilities in comprehending multimodal data, which aim to mimic human-like understanding through multimodal perception and reasoning. To realize the practical applications of MLLMs in the real world, a significant challenge lies in comprehending text-rich visual data, which is pervasive in various contexts with images intertwined with texts. An MLLM capable of compre-

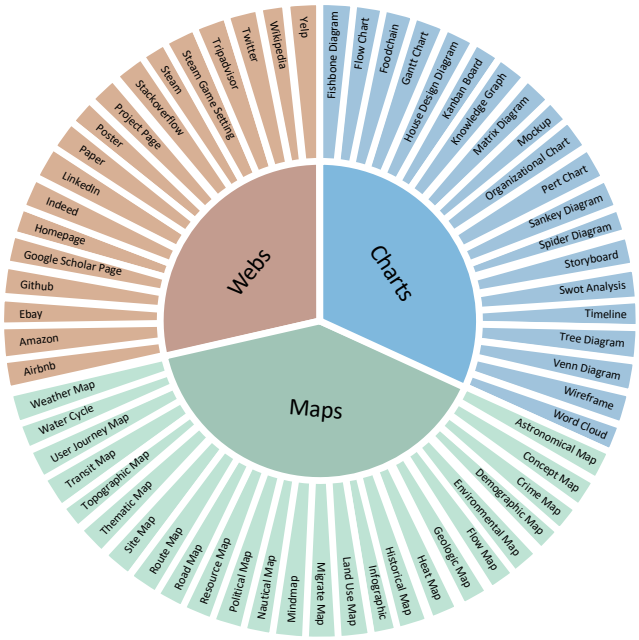


Figure 1. Overview of 63 data types within three major categories including Charts, Maps, and Webs in SEED-Bench-2-Plus, which encompasses an extensive range of **text-rich scenarios** for evaluating visual comprehension of MLLMs.

*This technical report is an extension to SEED-Bench-2 [22].
†Correspondence Author.

Table 1. Comparisons between existing MLLM benchmarks. "H/G Evaluation" denotes whether human or GPT is used for evaluation. "#Models" denotes the number of evaluated model.

Benchmark	Customized Question	Text-Rich Data	Text-Rich Scenes	#Answer Annotation	Answer Type	H/G Evaluation	#Models
LLaVA-Bench [30]	✓	✗	-	-	free-form	GPT	4
MME [13]	✓	✗	-	-	Y/N	N/A	10
M3Exam [57]	✓	✗	-	-	A/B/C/D	N/A	7
LAMM [52]	✗	✗	-	-	free-form	GPT	4
LVLM-eHub [49]	✗	✗	-	-	free-form	Human	8
MMBench [32]	✗	✗	-	-	free-form	GPT	14
VisIT-Bench [6]	✓	✗	-	-	free-form	Human/GPT	14
MM-VET [53]	✓	✗	-	-	free-form	GPT	9
Touchstone [5]	✓	✗	-	-	free-form	GPT	7
Q-bench [47]	✓	✗	-	-	Y/N & free-form	N/A	15
SEED-Bench-1 [23]	✓	✗	-	-	A/B/C/D	N/A	18
SEED-Bench-2 [22]	✓	✗	-	-	A/B/C/D	N/A	23
OCR-Bench [33]	✗	✓	28	-	free-form	N/A	6
CONTEXTUAL [44]	✓	✓	8	506	free-form	GPT	13
MathVista [34]	✓	✓	19	735	A/B/C/D & free-form	N/A	12
MMMU [54]	✓	✓	30	11.5K	A/B/C/D & free-form	N/A	23
SEED-Bench-2-Plus(Ours)	✓	✓	63	2.3K	A/B/C/D	N/A	34

hensively comprehending text-rich scenarios should be able to interpret texts, understand visual content, and discern the interactions between textual and visual contexts.

With the emergence of increasingly powerful and versatile MLLMs, such as GPT-4V [1], Gemini-Pro-Vision [42], and Claude-3-Opus [2], it naturally raises the question: *How do these models perform in text-rich scenarios?* Although recent benchmarks [13, 22, 23, 32] are specifically designed to evaluate MLLMs, their primary focus is on general visual comprehension (*e.g.*, images in different domains), leaving a significant gap in a comprehensive and objective evaluation of MLLM in text-rich contexts.

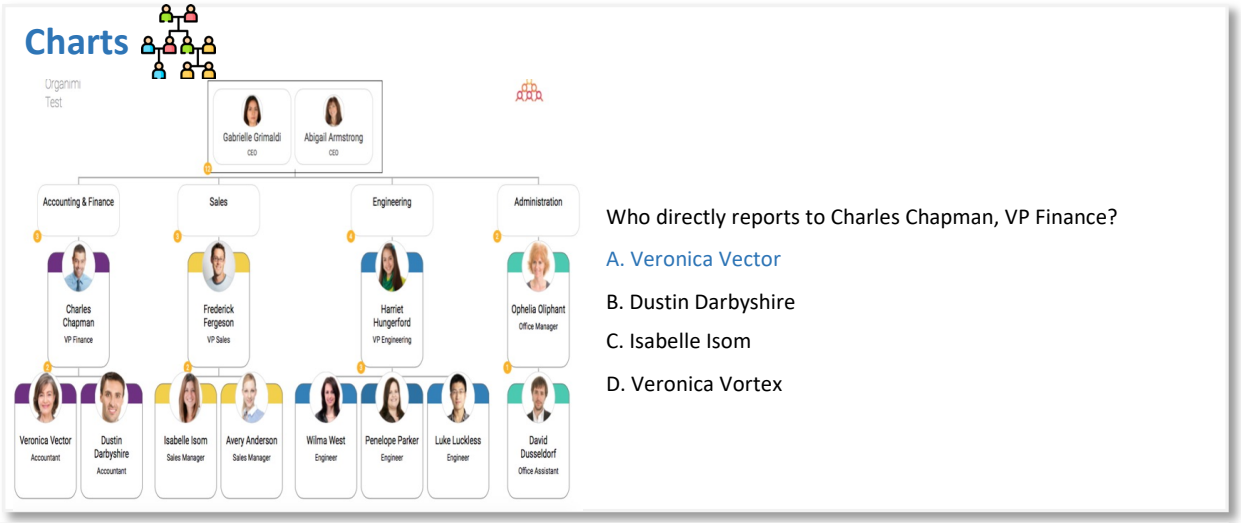
In this work, we introduce SEED-Bench-2-Plus, a comprehensive benchmark designed specifically to assess MLLMs' performance in comprehending text-rich visual data, which covers a wide spectrum range of text-rich scenarios in the real world. Specially, we meticulously craft 2.3K multiple-choice questions spanning three broad categories including Charts, Maps and Charts, as shown in Figure 2. The broad categories are further divided into 63 specific types as shown in Figure 1, to capture a more granular view of the challenges presented by text-rich visual comprehension. Each question in our benchmark is answered by human annotators, ensuring the accuracy and reliability of the ground-truth answer.

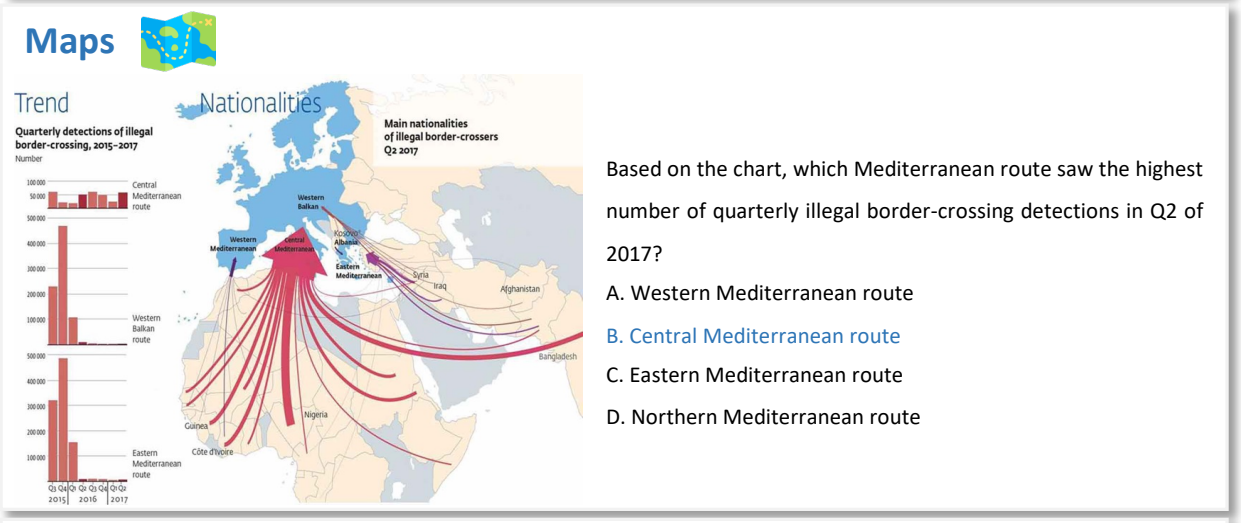
We further conduct extensive evaluation, encompassing 34 prominent MLLMs, including GPT-4V, Gemini-Pro-Vision, and Claude-3-Opus, which reveals critical insights into the current limitations and strengths of these models in comprehending text-rich information. Our findings underline several key observations: the inherent complexity of text-rich visual data, the varying difficulty levels across different data types, and the performance disparities among leading MLLMs.

Through this evaluation, we aim not only to benchmark current MLLM performance but also to catalyze further research in enhancing MLLMs' proficiency of multimodal comprehension in text-rich scenarios. Serving as a valuable supplement to Seed-Bench-2 [22], both the dataset and the evaluation code of Seed-Bench-2-Plus are made publicly available. We also consistently maintain a leaderboard to facilitate ongoing contributions to this area.

2. Related Work

Multimodal Large Language Models. Following the remarkable success of Large Language Models (LLM) [8, 12, 43], recent studies work on generative Multimodal Large Language Models (MLLMs) [4, 9, 19, 21, 24, 25, 29, 30, 39, 40, 50, 56, 58] aim to enhance multimodal comprehension by aligning the visual features of pre-trained image encoders with LLMs on image-text datasets. Some research [26, 35, 36] has further incorporated video inputs, leveraging the vast capabilities of LLMs for video understanding tasks. Recent work [10, 16, 17, 27, 41, 48] take significant strides in equipping MLLMs with the capacity for generating images beyond texts. However, their capability in text-rich scenarios has yet to be comprehensively and objectively assessed, which is the primary motivation of SEED-Bench-2-Plus.





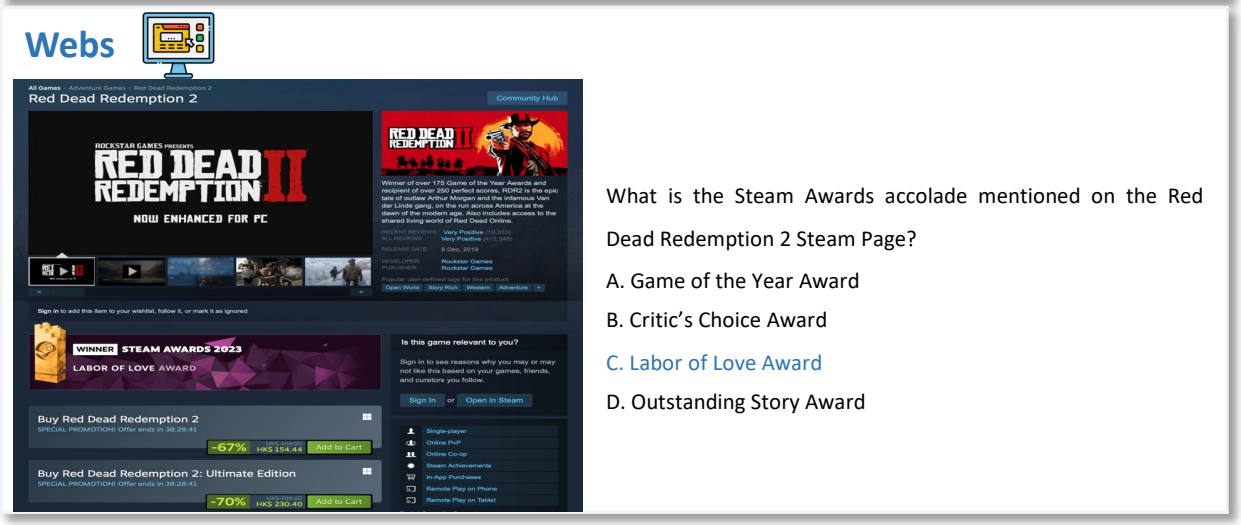


Figure 2. Data samples in SEED-Bench-2-Plus spanning three broad categories: Charts, Maps, and Webs. These categories cover a wide spectrum of multiple-choice questions in text-rich scenarios, with ground truth options derived from human annotations.

Rank	Model	Accuracy(%)
1	GPT-4V	53.8
2	Gemini-Pro-Vision	52.8
3	InternLM-Xcomposer-VL-2	51.5
4	SPHINX-v2-1k	48.0
5	SEED-X	47.1

(1) All Text-Rich Data

Rank	Model	Accuracy(%)
1	GPT-4V	54.8
2	Gemini-Pro-Vision	52.1
3	InternLM-Xcomposer-VL-2	49.4
4	SEED-X	46.9
5	Claude-3-Opus	43.7

(2) Charts

Rank	Model	Accuracy(%)
1	GPT-4V	49.4
1	Gemini-Pro-Vision	49.4
3	InternLM-Xcomposer-VL-2	47.1
4	Claude-3-Opus	43.9
5	SEED-X	43.3

(3) Maps

Rank	Model	Accuracy(%)
1	SPHINX-v2-1k	60.5
2	InternLM-Xcomposer-VL-2	58.0
3	GPT-4V	57.2
4	Qwen-VL-Chat	57.0
5	Gemini-Pro-Vision	56.8

(4) Webs

Figure 3. Leaderboard of SEED-Bench-2-Plus.

Benchmarks for Multimodal Large Language Models. In tandem with the rapid development of Multimodal Large Language Models (MLLMs), several concurrent studies [5, 13, 32, 49, 52] have proposed various benchmarks for evaluating MLLMs. For instance, GVT [45] constructs a benchmark by aggregating two semantic-level understanding tasks (VQA and Image Captioning) and two fine-grained tasks (Object Counting and Multi-class Identification). However, its evaluation is limited to specific aspects of visual understanding. LVLM-eHub [49] combines multiple existing computer vision benchmarks and develops an online platform where two models are prompted to answer a question related to an image, and human annotators are employed to compare the models' predictions. The involvement of human annotators during evaluation not only introduces bias but also incurs significant costs. LLaVA-Bench [30], LAMM [52], and Touchstone [5] utilize GPT to evaluate the relevance and accuracy of answers in relation to the ground truth. The reliance on entity extraction and GPT metric can affect the accuracy and reliability of the evaluation. MME [13] and MMBench [32] aim to enhance the objective evaluation of MLLMs by constructing 2194 True/False Questions and 2974 Multiple Choice Questions across various ability dimensions, respectively. MMMU [54] generates numerous college-level multi-discipline QAs to evaluate MLLMs' knowledge-ability. In SEED-Bench-2-Plus, we focus on evaluating MLLMs' performance in comprehending text-rich visual data, which covers a wide spectrum range of text-rich scenarios in the real world.

3. SEED-Bench-2-Plus

Our SEED-Bench-2-Plus benchmark incorporates 2K multiple-choice questions, all of which are accompanied by accurate human annotations and span three broad categories including Chats, Maps and Webs. In this section, we first the broad categories of SEED-Bench-2-Plus in Sec. 3.1. We then introduce the data source in Sec. 3.2, and finally, we describe the evaluation strategy for MLLMs to answer multiple-choice questions in Sec. 3.3.

3.1. Broad Categories

To thoroughly evaluate the capabilities of MLLMs in comprehending text-rich data, SEED-Bench-2-Plus encompasses 3 broad categories (see Figure. 2), which can be further divided into 63 data types (see Figure. 1).

Charts. This category pertains to the information contained within the chart image. In this task, the model is required to understand the specific semantics of each chart type, extract relevant information, and answer questions based on the context and spatial relationships.

Maps. This category corresponds to the information present in the map image. The model is expected to identify symbols, text, and spatial relationships, and use this information to answer questions that often require geographical or domain-specific knowledge.

Webs. In this category, the model needs to understand the layout and design of different websites, extract relevant information from various elements, and answer questions that may relate to the website's content, functionality, or design based on the given website screenshot.

3.2. Data Source

To construct a comprehensive benchmark that encapsulates a variety of evaluation scenarios, it is imperative to amass an extensive collection of data. This data should predominantly include images that are rich in textual information, thereby creating a text-rich dataset.

Table 2. Evaluation results of various MLLMs in SEED-Bench-2-Plus. The best (second best) is in bold (underline).

Model	Language Model	Text-Rich Data			
		Average	Charts	Maps	Webs
BLIP-2 [25]	Flan-T5-XL	29.8	33.6	30.2	25.6
InstructBLIP [9]	Flan-T5-XL	29.2	31.7	28.7	27.2
InstructBLIP Vicuna [9]	Vicuna-7B	30.9	30.0	32.7	29.9
LLaVA [30]	LLaMA-7B	30.1	29.9	29.0	31.3
MiniGPT-4 [58]	Vicuna-7B	30.2	30.5	30.4	29.7
VPGTrans [55]	LLaMA-7B	30.3	30.0	31.3	29.6
MultiModal-GPT [19]	Vicuna-7B	31.7	30.5	32.7	32.0
Otter [24]	LLaMA-7B	31.3	29.5	32.3	32.0
OpenFlamingo [3]	LLaMA-7B	31.7	30.5	32.7	32.0
LLaMA-Adapter V2 [14]	LLaMA-7B	30.6	29.9	30.8	31.1
GVT [45]	Vicuna-7B	29.7	29.3	30.2	29.7
mPLUG-Owl [50]	LLaMA-7B	31.8	30.6	33.5	31.1
Qwen-VL [4]	Qwen-7B	37.0	38.2	37.0	55.9
Qwen-VL-Chat [4]	Qwen-7B	43.4	37.3	35.9	57.0
LLaVA-1.5 [29]	Vicuna-7B	36.8	36.5	35.1	38.8
IDEFICS-9B-Instruct [21]	LLaMA-7B	32.1	31.0	31.8	33.5
InternLM-Xcomposer-VL [56]	InternLM-7B	40.6	39.9	39.0	43.0
VideoChat [26]	Vicuna-7B	28.6	27.8	29.7	28.3
Video-ChatGPT [36]	LLaMA-7B	29.8	29.9	29.0	30.5
Valley [35]	LLaMA-13B	29.2	29.1	27.4	31.1
Emu [41]	LLaMA-13B	33.5	32.4	34.2	34.0
NExt-GPT [48]	Vicuna-7B	26.2	26.3	26.6	25.7
SEED-LLaMA [17]	LLaMA2-Chat-13B	33.7	32.5	35.7	33.1
CogVLM [46]	Vicuna-7B	33.4	32.6	34.1	33.5
InternLM-Xcomposer-VL2 [11]	InternLM2-7B	51.5	49.4	47.1	<u>58.0</u>
InternLM-Xcomposer-VL2-4bit [11]	InternLM2-7B	37.6	37.4	38.8	36.7
LLaVA-Next [31]	Vicuna-7B	36.8	36.4	34.0	39.9
Yi-VL [20]	Yi-6B	34.8	32.4	34.6	37.5
SPHINX-v2-1k [15]	LLaMA2-13B	48.0	41.7	41.9	**60.5**
mPLUG-Owl2 [51]	LLaMA2-7B	33.4	33.5	32.6	34.0
SEED-X [18]	LLaMA2-13B	47.1	46.9	43.3	52.6
GPT-4V [1]	-	**53.8**	**54.8**	**49.4**	57.2
Gemini-Pro-Vision [42]	-	<u>52.8</u>	<u>52.1</u>	**49.4**	56.8
Claude-3-Opus [2]	-	44.2	43.7	43.9	45.1

Charts. To obtain chart data with rich text information, we manually gather 20 types of charts from the Internet. These types encompass Fishbone Diagram, Flow Chart, Food-chain, Gantt Chart, House Design Diagram, Kanban Board, Knowledge Graph, Matrix Diagram, Mockups, Organizational Chart, Pert Chart, Sankey Diagram, Spider Diagram, Storyboard, Swot Analysis, Timeline, Tree Diagram, Venn Diagram, Wireframes, and Word Cloud. Specifically, we employ GPT-4V [1] to generate corresponding questions and use human annotators to enhance the quality of the questions and corresponding options.

Maps. As for map data rich in text information, we manually collect 25 types of maps from the Internet. These types include Astronomical Map, Concept Map, Crime Map, Demographic Map, Environmental Map, Flow Map, Geologic Map, Heat Map, Historical Map, Infographics, Land Use Map, Migration Map, Mindmap, Nautical Map, Political Map, Resource Map, Road Map, Route Map, Site Map, Thematic Maps, Topographic Map, Transit Map, User Journey Map, Water Cycle, and Weather Map. Specifically, we utilize GPT-4V [1] to generate corresponding questions and employ human annotators to enhance the quality of the questions and corresponding options.

Webs. For web data rich in text information, we man-

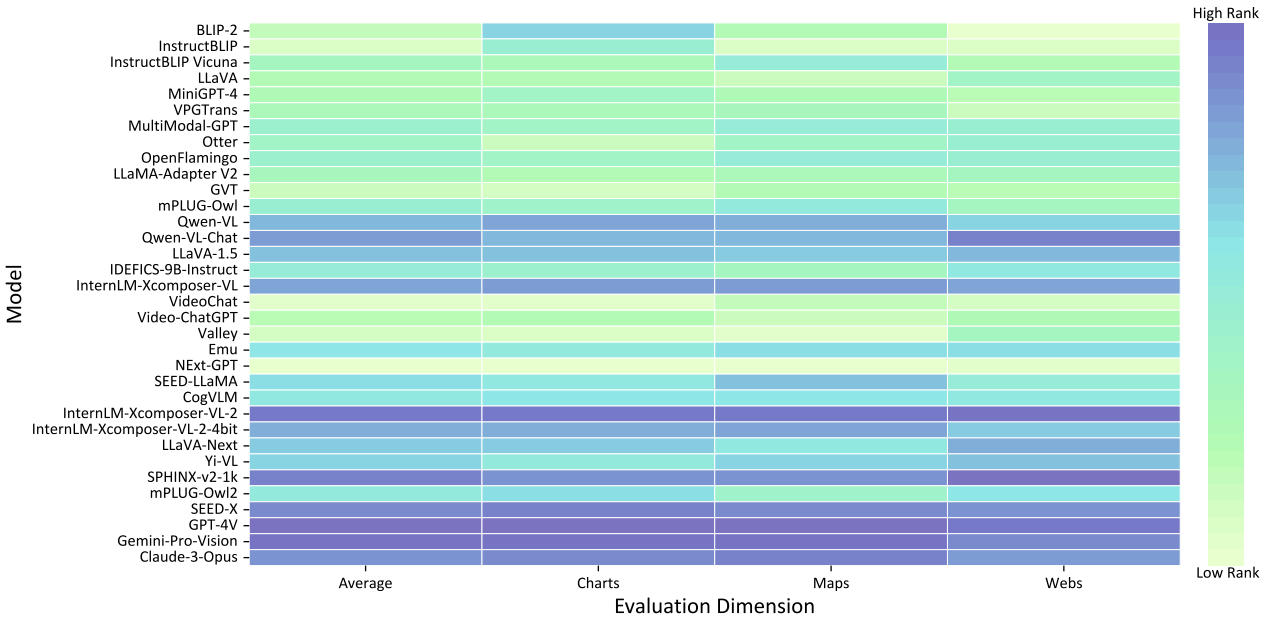


Figure 4. Illustration of each model's performance across different categories in SEED-Bench-2-Plus with text-rich scenarios, where darker colors represent higher ranks.

Table 3. Evaluation results of various MLLMs in different types of "Charts" (Part 1) in SEED-Bench-2-plus. The best (second best) is in bold (underline).

Model	Language Model	Fishbone Diagram	Flow Chart	Foodchain	Gantt Chart	House Design Diagram	Kanban Board	Knowledge Graph	Matrix Diagram	Mockups	Organizational Chart
BLIP-2 [25]	Flan-T5-XL	37.8	36.2	34.9	32.6	25.0	48.8	39.5	30.0	43.8	39.1
InstructBLIP [9]	Flan-T5-XL	37.8	38.3	41.9	34.9	15.9	36.6	31.6	30.0	46.9	41.3
InstructBLIP Vicuna [9]	Vicuna-7B	26.7	40.4	27.9	25.6	18.2	26.8	52.6	22.5	46.9	34.8
LLaVA [30]	LLaMA-7B	15.6	40.4	41.9	39.5	29.6	22.0	55.3	15.0	53.1	30.4
MiniGPT-4 [58]	Vicuna-7B	28.9	36.2	27.9	32.6	25.0	36.6	36.8	20.0	50.0	41.3
VPGTrans [55]	LLaMA-7B	26.7	34.0	20.9	27.9	15.9	46.3	55.3	17.5	37.5	37.0
MultiModal-GPT [19]	Vicuna-7B	26.7	46.8	30.2	37.2	25.0	36.6	55.3	17.5	37.5	39.1
Otter [24]	LLaMA-7B	28.9	44.7	27.9	32.6	20.5	31.7	55.3	15.0	37.5	37.0
OpenFlamingo [3]	LLaMA-7B	26.7	46.8	30.2	37.2	25.0	36.6	55.3	17.5	37.5	39.1
LLaMA-Adapter V2 [14]	LLaMA-7B	24.4	38.3	23.3	39.5	13.6	43.9	44.7	20.0	50.0	34.8
GVT [45]	Vicuna-7B	24.4	38.3	23.3	32.6	22.7	29.3	44.7	17.5	43.8	34.8
mPLUG-Owl [50]	LLaMA-7B	20.0	46.8	30.2	34.9	22.7	36.6	47.4	17.5	40.6	39.1
Qwen-VL [4]	Qwen-7B	37.8	44.7	39.5	34.9	34.1	41.5	44.7	30.0	56.3	47.8
Qwen-VL-Chat [4]	Qwen-7B	33.3	44.7	41.9	34.9	36.4	29.3	50.0	27.5	56.3	52.2
LLaVA-1.5 [29]	Vicuna-7B	37.8	51.1	34.9	32.6	20.5	31.7	63.2	22.5	62.5	39.1
IDEFICS-9B-Instruct [21]	LLaMA-7B	24.4	44.7	30.2	32.6	15.9	36.6	50.0	22.5	43.8	34.8
InternLM-Xcomposer-VL [56]	InternLM-7B	37.8	53.2	58.1	27.9	38.6	24.4	42.1	37.5	43.8	37.0
VideoChat [26]	Vicuna-7B	24.4	31.9	32.6	23.3	20.5	31.7	39.5	17.5	37.5	43.5
Video-ChatGPT [36]	LLaMA-7B	31.1	36.2	27.9	32.6	20.5	26.8	36.8	15.0	53.1	32.6
Valley [35]	LLaMA-13B	22.2	38.3	23.3	34.9	20.5	36.6	44.7	17.5	53.1	39.1
Emu [41]	LLaMA-13B	31.1	38.3	37.2	39.5	25.0	26.8	57.9	12.5	56.3	37.0
NExt-GPT [48]	Vicuna-7B	37.8	29.8	30.2	32.6	25.0	29.3	29.0	20.0	53.1	19.6
SEED-LLaMA [17]	LLaMA2-Chat-13B	28.9	42.6	39.5	34.9	15.9	31.7	52.6	15.0	59.4	37.0
CogVLM [46]	Vicuna-7B	28.9	44.7	34.9	39.5	9.1	34.2	52.6	25.0	56.3	28.3
InternLM-Xcomposer-VL2 [11]	InternLM2-7B	42.2	53.2	65.1	**44.2**	40.9	31.7	60.5	37.5	59.4	34.8
InternLM-Xcomposer-VL2-4bit [11]	InternLM2-7B	24.4	51.1	46.5	34.9	34.1	36.6	50.0	35.0	40.6	21.7
LLaVA-Next [31]	Vicuna-7B	28.9	46.8	34.9	34.9	31.8	36.6	47.4	30.0	46.9	39.1
Yi-VL [20]	Yi-6B	28.9	36.2	37.2	25.6	27.3	12.2	47.4	40.0	43.8	41.3
SPHINX-v2-1k [15]	LLaMA2-7B	42.2	53.2	34.9	20.9	43.2	29.3	50.0	40.0	**65.6**	**60.9**
mPLUG-Owl2 [51]	LLaMA2-7B	31.1	44.7	30.2	34.9	25.0	41.5	52.6	20.0	43.8	41.3
SEED-X [18]	LLaMA2-13B	37.8	51.1	53.5	30.2	34.1	36.6	55.3	40.0	**65.6**	47.8
GPT-4V [1]	-	**51.1**	66.0	67.4	30.2	45.5	40.0	**76.3**	**45.0**	61.3	42.2
Gemini-Pro-Vision [42]	-	34.2	**66.7**	**70.0**	27.5	**46.5**	**57.5**	65.0	27.6	51.6	52.3
Claude-3-Opus [2]	-	35.6	59.6	67.4	32.6	38.6	22.0	57.9	37.5	48.4	19.6

Table 4. Evaluation results of various MLLMs in different types of "Charts" (Part 2) in SEED-Bench-2-plus. The best (second best) is in bold (underline).

Model	Language Model	Pert Chart	Sankey Diagram	Spider Diagram	Storyboard	Swot Analysis	Timeline	Tree Diagram	Venn Diagramm	Wireframes	Word Cloud
BLIP-2 [25]	Flan-T5-XL	23.3	40.0	22.2	6.7	26.3	50.0	32.4	37.0	33.3	43.2
InstructBLIP [9]	Flan-T5-XL	20.9	36.7	22.2	17.8	23.7	36.7	29.4	30.4	25.0	40.9
InstructBLIP Vicuna [9]	Vicuna-7B	23.3	20.0	31.1	13.3	31.6	46.7	17.7	23.9	41.7	36.4
LLaVA [30]	LLaMA-7B	27.9	20.0	24.4	11.1	31.6	43.3	29.4	17.4	36.1	25.0
MiniGPT-4 [58]	Vicuna-7B	34.9	26.7	20.0	11.1	29.0	36.7	29.4	21.7	38.9	34.1
VPGTrans [55]	LLaMA-7B	34.9	13.3	31.1	17.8	26.3	46.7	14.7	26.1	41.7	31.8
MultiModal-GPT [19]	Vicuna-7B	27.9	16.7	20.0	11.1	23.7	40.0	23.5	30.4	36.1	29.6
Otter [24]	LLaMA-7B	32.6	20.0	22.2	6.7	23.7	46.7	26.5	26.1	33.3	27.3
OpenFlamingo [3]	LLaMA-7B	27.9	16.7	20.0	11.1	23.7	40.0	23.5	30.4	36.1	29.6
LLaMA-Adapter V2 [14]	LLaMA-7B	25.6	16.7	24.4	8.9	18.4	43.3	32.4	15.2	52.8	38.6
GVT [45]	Vicuna-7B	25.6	20.0	17.8	22.2	36.8	40.0	41.2	17.4	36.1	27.3
mPLUG-Owl [50]	LLaMA-7B	34.9	6.7	26.7	15.6	29.0	43.3	26.5	26.1	36.1	31.8
Qwen-VL [4]	Qwen-7B	27.9	26.7	20.0	42.2	42.1	50.0	44.1	23.9	52.8	31.8
Qwen-VL-Chat [4]	Qwen-7B	37.2	20.0	24.4	55.6	34.2	46.7	41.2	15.2	41.7	27.3
LLaVA-1.5 [29]	Vicuna-7B	27.9	20.0	24.4	40.0	39.5	53.3	29.4	26.1	44.4	38.6
IDEFICS-9B-Instruct [21]	LLaMA-7B	27.9	23.3	22.2	22.2	26.3	50.0	26.5	26.1	36.1	31.8
InternLM-Xcomposer-VL [56]	InternLM-7B	37.2	30.0	37.8	31.1	36.8	50.0	32.4	39.1	52.8	50.0
VideoChat [26]	Vicuna-7B	25.6	23.3	26.7	8.9	29.0	33.3	26.5	15.2	36.1	34.1
Video-ChatGPT [36]	LLaMA-7B	23.3	20.0	26.7	28.9	23.7	40.0	38.2	19.6	41.7	31.8
Valley [35]	LLaMA-13B	23.3	23.3	22.2	20.0	26.3	26.7	26.5	23.9	36.1	29.6
Emu [41]	LLaMA-13B	27.9	16.7	22.2	8.9	36.8	36.7	35.3	26.1	44.4	38.6
NExt-GPT [48]	Vicuna-7B	20.9	23.3	20.0	13.3	34.2	13.3	26.5	21.7	25.0	25.0
SEED-LLaMA [17]	LLaMA2-Chat-13B	34.9	13.3	33.3	17.8	36.8	40.0	23.5	32.6	33.3	29.6
CogVLM [46]	Vicuna-7B	25.6	30.0	22.2	24.4	39.5	46.7	20.6	23.9	38.9	38.6
InternLM-Xcomposer-VL2 [11]	InternLM2-7B	32.6	33.3	51.1	**68.9**	52.6	53.3	52.9	56.5	66.7	52.3
InternLM-Xcomposer-VL2-4bit [11]	InternLM2-7B	27.9	30.0	51.1	40.0	44.7	36.7	32.4	23.9	38.9	47.7
LLaVA-Next [31]	Vicuna-7B	25.6	30.0	26.7	42.2	34.2	53.3	44.1	21.7	38.9	43.2
Yi-VL [20]	Yi-6B	23.3	16.7	26.7	33.3	42.1	26.7	29.4	23.9	47.2	38.6
SPHINX-v2-1k [15]	LLaMA2-7B	39.5	16.7	35.6	53.3	42.1	56.7	32.4	32.6	50.0	36.4
mPLUG-Owl2 [51]	LLaMA2-7B	20.9	30.0	31.1	26.7	36.8	36.7	26.5	23.9	41.7	34.1
SEED-X [18]	LLaMA2-13B	37.2	**50.0**	46.7	53.3	65.8	53.3	38.2	34.8	**69.4**	50.0
GPT-4V [1]	-	39.5	40.0	**62.2**	65.9	65.8	**66.7**	**61.8**	**60.9**	55.6	**55.8**
Gemini-Pro-Vision [42]	-	**41.5**	48.2	57.5	56.1	**67.7**	62.1	46.4	58.5	58.6	50.0
Claude-3-Opus [2]	-	37.2	33.3	55.6	53.5	60.5	43.3	50.0	41.3	36.1	45.5

ually gather 18 types of screenshots from various websites on the Internet. These types include Airbnb, Amazon, Ebay, Github, Google Scholar Page, Homepage, Indeed, Linkedin, Papers, Poster, Project Page, Stackoverflow, Steam, Steam Game Setting, Tripadvisor, Twitter, Wikipedia, and Yelp. Specifically, we utilize GPT-4V [1] to generate corresponding questions and employ human annotators to enhance the quality of the questions and corresponding options.

3.3. Evaluation Strategy

Different from MMBench [32], which employs ChatGPT to match a model's prediction to one of the choices in a multiple-choice question (achieving only an 87.0% alignment rate), we adopt the answer ranking strategy [7, 9, 28] to evaluate existing MLLMs with multiple-choice questions. Specifically, for each choice of a question, we compute the likelihood that an MLLM generates the content of this choice given the question. We select the choice with the highest likelihood as the model's prediction. Our evaluation strategy does not depend on the instruction-following capabilities of models to output "A", "B", "C" or "D". Also, this evaluation strategy eliminates the impact of the order of multiple-choice options on the model's performance.

4. Evaluation Results

4.1. Models

We evaluate a total of 31 open-source MLLMs including BLIP-2 [25], InstructBLIP [9], InstructBLIP Vicuna [9], LLaVA [30], MiniGPT-4 [58], VPGTrans [55], MultiModal-GPT [19], Otter [24], OpenFlamingo [3], LLaMA-Adapter V2 [14], GVT [45], mPLUG-Owl [50], Qwen-VL [4], Qwen-VL-Chat [4], LLaVA1.5 [29], IDEFICS-9B-Instruct [21], InternLM-Xcomposer-VL [56], VideoChat [26], Video-ChatGPT [36], Valley [35], Emu [41], NExt-GPT [48], SEED-LLaMA [17], CogVLM [46], InternLM-Xcomposer-VL-2 [11], InternLM-Xcomposer-VL-2-4bit [11], LLaVA-Next [31], Yi-VL [20], SPHINX-v2-1k [15], mPLUG-Owl2 [51], SEED-X [18] (We evaluate the general instruction-tuned model SEED-X-I), and 3 closed-source MLLMs including GPT-4V [1], Gemini-Pro-Vision [42], and Claude-3-Opus [2], based on their official implementations. It is

Table 5. Evaluation results of various MLLMs in different types of "Maps" (Part 1) in SEED-Bench-2-plus. The best (second best) is in bold (underline).

Model	Language Model	Astronomical Maps	Concept Maps	Crime Maps	Demographic Maps	Environmental Maps	Flow Maps	Geologic Maps	Heat Maps	Historical Maps	Infographic	Land Use Maps	Migration Maps	Mindmap
BLIP-2 [25]	Flan-T5-XL	20.0	39.6	34.6	31.6	14.7	23.3	18.5	26.1	26.7	43.1	27.8	23.7	22.0
InstructBLIP [9]	Flan-T5-XL	16.7	31.3	34.6	21.1	14.7	20.0	18.5	30.4	26.7	40.0	27.8	31.6	22.0
InstructBLIP Vicuna [9]	Vicuna-7B	20.0	37.5	38.5	42.1	29.4	33.3	14.8	34.8	40.0	32.3	11.1	44.7	29.3
LLaVA [30]	LLaMA-7B	10.0	29.2	15.4	42.1	23.5	30.0	18.5	30.4	40.0	35.4	16.7	31.6	26.8
MiniGPT-4 [58]	Vicuna-7B	13.3	22.9	26.9	36.8	26.5	16.7	18.5	34.8	53.3	33.9	19.4	31.6	29.3
VPGTrans [55]	LLaMA-7B	30.0	31.3	30.8	31.6	17.7	26.7	11.1	34.8	40.0	36.9	27.8	34.2	26.8
MultiModal-GPT [19]	Vicuna-7B	26.7	39.6	34.6	42.1	26.5	30.0	14.8	34.8	53.3	38.5	16.7	44.7	24.4
Otter [24]	LLaMA-7B	26.7	37.5	34.6	52.6	29.4	30.0	11.1	30.4	53.3	38.5	16.7	42.1	24.4
OpenFlamingo [3]	LLaMA-7B	26.7	39.6	34.6	42.1	26.5	30.0	14.8	34.8	53.3	38.5	16.7	44.7	24.4
LLaMA-Adapter V2 [14]	LLaMA-7B	16.7	33.3	42.3	47.4	29.4	23.3	14.8	17.4	40.0	33.9	22.2	47.4	26.8
GVT [45]	Vicuna-7B	16.7	31.3	30.8	36.8	29.4	33.3	29.6	34.8	33.3	29.2	16.7	42.1	34.2
mPLUG-Owl [50]	LLaMA-7B	23.3	41.7	46.2	36.8	23.5	26.7	18.5	39.1	73.3	36.9	19.4	42.1	29.3
Qwen-VL [4]	Qwen-7B	13.3	31.3	34.6	31.6	29.4	40.0	14.8	34.8	40.0	53.9	33.3	39.5	26.8
Qwen-VL-Chat [4]	Qwen-7B	20.0	43.8	30.8	26.3	26.5	36.7	22.2	17.4	40.0	55.4	36.1	29.0	26.8
LLaVA-1.5 [29]	Vicuna-7B	20.0	35.4	46.2	36.8	26.5	26.7	25.9	26.1	46.7	38.5	16.7	39.5	39.0
IDEFICS-9B-Instruct [21]	LLaMA-7B	16.7	29.2	38.5	31.6	23.5	40.0	22.2	26.1	46.7	41.5	22.2	**50.0**	22.0
InternLM-Xcomposer-VL [56]	InternLM-7B	26.7	33.3	42.3	47.4	29.4	33.3	22.2	21.7	40.0	55.4	47.2	42.1	46.3
VideoChat [26]	Vicuna-7B	16.7	33.3	19.2	36.8	23.5	23.3	22.2	30.4	33.3	30.8	22.2	36.8	24.4
Video-ChatGPT [36]	LLaMA-7B	20.0	27.1	30.8	47.4	23.5	33.3	25.9	30.4	26.7	27.7	11.1	29.0	31.7
Valley [35]	LLaMA-13B	23.3	25.0	34.6	26.3	23.5	36.7	14.8	30.4	46.7	32.3	19.4	42.1	17.1
Emu [41]	LLaMA-13B	20.0	37.5	46.2	47.4	26.5	23.3	22.2	39.1	46.7	38.5	22.2	36.8	26.8
NExt-GPT [48]	Vicuna-7B	30.0	25.0	15.4	42.1	23.5	23.3	33.3	30.4	13.3	16.9	25.0	39.5	24.4
SEED-LLaMA [17]	LLaMA2-Chat-13B	36.7	37.5	23.1	42.1	35.3	20.0	22.2	43.5	53.3	36.9	22.2	44.7	31.7
CogVLM [46]	Vicuna-7B	13.3	35.4	42.3	31.6	32.4	40.0	18.5	26.1	33.3	43.1	22.2	36.8	36.6
InternLM-Xcomposer-VL2 [11]	InternLM2-7B	**60.0**	56.3	**65.4**	**57.9**	35.3	50.0	25.9	**52.2**	46.7	64.6	30.6	34.2	46.3
InternLM-Xcomposer-VL2-4bit [11]	InternLM2-7B	40.0	50.0	23.1	36.8	32.4	40.0	25.9	43.5	60.0	55.4	38.9	36.8	43.9
LLaVA-Next [31]	Vicuna-7B	23.3	35.4	34.6	31.6	29.4	33.3	22.2	34.8	33.3	40.0	5.6	29.0	26.8
Yi-VL [20]	Yi-6B	20.0	54.2	23.1	26.3	17.7	26.7	14.8	30.4	40.0	44.6	22.2	31.6	41.5
SPHINX-v2-1k [15]	LLaMA2-7B	20.0	43.8	53.9	42.1	38.2	43.3	37.0	39.1	33.3	52.3	16.7	47.4	39.0
mPLUG-Owl2 [51]	LLaMA2-7B	13.3	31.3	30.8	36.8	32.4	33.3	40.7	30.4	33.3	32.3	16.7	42.1	24.4
SEED-X [18]	LLaMA2-13B	43.3	47.9	30.8	36.8	32.4	43.3	29.6	34.8	60.0	58.5	38.9	39.5	**48.8**
GPT-4V [1]	-	55.2	**62.5**	30.4	52.9	**41.9**	48.3	42.3	28.6	66.7	67.2	32.3	46.0	41.5
Gemini-Pro-Vision [42]	-	46.2	59.0	36.4	43.8	33.3	**53.6**	**50.0**	42.9	**76.9**	**75.0**	**52.2**	42.4	48.6
Claude-3-Opus [2]	-	56.7	37.5	34.6	31.6	32.4	43.3	33.3	39.1	66.7	67.7	44.4	36.8	43.9

Table 6. Evaluation results of various MLLMs in different types of "Maps" (Part 2) in SEED-Bench-2-plus. The best (second best) is in bold (underline).

Model	Language Model	Nautical Maps	Political Maps	Resource Maps	Road Maps	Route Maps	Site Maps	Thematic Maps	Topographic Maps	Transit Maps	User Journey Maps	Water Cycle	Weather Maps
BLIP-2 [25]	Flan-T5-XL	16.7	50.0	39.1	36.4	37.5	38.2	12.5	44.0	34.5	23.7	39.0	22.6
InstructBLIP [9]	Flan-T5-XL	25.0	38.5	21.7	43.2	27.5	38.2	12.5	48.0	34.5	26.3	39.0	19.4
InstructBLIP Vicuna [9]	Vicuna-7B	16.7	53.9	13.0	**59.1**	27.5	32.4	15.6	52.0	37.9	29.0	36.6	25.8
LLaVA [30]	LLaMA-7B	25.0	42.3	30.4	45.5	25.0	29.4	15.6	32.0	48.3	26.3	31.7	25.8
MiniGPT-4 [58]	Vicuna-7B	25.0	46.2	26.1	47.7	32.5	29.4	28.1	44.0	41.4	26.3	26.8	22.6
VPGTrans [55]	LLaMA-7B	25.0	50.0	39.1	43.2	20.0	26.5	34.4	36.0	24.1	23.7	46.3	32.3
MultiModal-GPT [19]	Vicuna-7B	25.0	50.0	21.7	47.7	25.0	29.4	18.8	52.0	27.6	31.6	43.9	19.4
Otter [24]	LLaMA-7B	25.0	50.0	34.8	45.5	27.5	29.4	25.0	52.0	27.6	21.1	41.5	16.1
OpenFlamingo [3]	LLaMA-7B	25.0	50.0	21.7	47.7	25.0	29.4	18.8	52.0	27.6	31.6	43.9	19.4
LLaMA-Adapter V2 [14]	LLaMA-7B	33.3	46.2	21.7	40.9	25.0	23.5	21.9	52.0	31.0	31.6	31.7	19.4
GVT [45]	Vicuna-7B	25.0	34.6	21.7	45.5	30.0	23.5	25.0	36.0	41.4	18.4	29.3	25.8
mPLUG-Owl [50]	LLaMA-7B	25.0	46.2	30.4	43.2	30.0	29.4	18.8	48.0	34.5	31.6	41.5	12.9
Qwen-VL [4]	Qwen-7B	41.7	38.5	26.1	36.4	37.5	44.1	34.4	44.0	27.6	21.1	53.7	48.4
Qwen-VL-Chat [4]	Qwen-7B	**58.3**	38.5	34.8	29.6	27.5	**50.0**	28.1	52.0	24.1	23.7	48.8	25.8
LLaVA-1.5 [29]	Vicuna-7B	33.3	53.9	17.4	45.5	30.0	29.4	21.9	**64.0**	27.6	36.8	46.3	19.4
IDEFICS-9B-Instruct [21]	LLaMA-7B	33.3	42.3	34.8	52.3	22.5	20.6	21.9	36.0	31.0	23.7	41.5	16.1
InternLM-Xcomposer-VL [56]	InternLM-7B	16.7	73.1	34.8	43.2	17.5	38.2	34.4	44.0	27.6	44.7	56.1	29.0
VideoChat [26]	Vicuna-7B	25.0	50.0	26.1	45.5	27.5	29.4	12.5	32.0	48.3	34.2	31.7	16.1
Video-ChatGPT [36]	LLaMA-7B	25.0	38.5	30.4	40.9	32.5	26.5	28.1	32.0	41.4	21.1	29.3	19.4
Valley [35]	LLaMA-13B	33.3	34.6	26.1	29.6	27.5	20.6	12.5	32.0	27.6	26.3	34.2	12.9
Emu [41]	LLaMA-13B	25.0	42.3	21.7	52.3	30.0	32.4	26.5	18.8	**60.0**	37.9	34.2	41.5
NExt-GPT [48]	Vicuna-7B	25.0	30.8	30.4	20.5	30.0	26.5	23.5	21.9	24.0	**48.3**	29.0	34.2
SEED-LLaMA [17]	LLaMA2-Chat-13B	33.3	57.7	30.4	43.2	25.0	44.1	41.2	18.8	52.0	27.6	39.5	48.8
CogVLM [46]	Vicuna-7B	25.0	69.2	17.4	43.2	25.0	23.5	25.0	60.0	20.7	34.2	43.9	19.4
InternLM-Xcomposer-VL2 [11]	InternLM2-7B	25.0	65.4	43.5	47.7	40.0	32.4	40.6	52.0	48.3	34.2	68.3	38.7
InternLM-Xcomposer-VL2-4bit [11]	InternLM2-7B	41.7	53.9	30.4	36.4	32.5	32.4	15.6	32.0	34.5	39.5	51.2	25.8
LLaVA-Next [31]	Vicuna-7B	41.7	38.5	26.1	50.0	32.5	35.3	25.0	48.0	48.3	34.2	41.5	22.6
Yi-VL [20]	Yi-6B	50.0	50.0	39.1	43.2	22.5	47.1	25.0	52.0	24.1	31.6	34.2	41.9
SPHINX-v2-1k [15]	LLaMA2-7B	**58.3**	65.4	17.4	43.2	42.5	**50.0**	18.8	56.0	37.9	23.7	68.3	25.8
mPLUG-Owl2 [51]	LLaMA2-7B	33.3	46.2	21.7	43.2	35.0	26.5	25.0	44.0	34.5	29.0	46.3	25.8
SEED-X [18]	LLaMA2-13B	41.7	61.5	17.4	40.9	35.0	44.1	**50.0**	31.3	**60.0**	**48.3**	34.2	53.7
GPT-4V [1]	-	54.6	**76.9**	45.5	45.5	**48.6**	44.1	37.5	52.2	41.4	43.2	**73.7**	45.2
Gemini-Pro-Vision [42]	-	57.1	47.6	38.1	54.1	33.3	43.5	32.1	50.0	30.0	45.7	72.5	44.4
Claude-3-Opus [2]	-	33.3	73.1	**47.8**	40.9	32.5	41.2	38.2	21.9	56.0	34.5	44.7	**68.3**

Table 7. Evaluation results of various MLLMs in different types of "Webs" (Part 1) in SEED-Bench-2-plus. The best (second best) is in bold (underline).

Model	Language Model	Airbnb	Amazon	Ebay	Github	Google Scholar Page	Homepage	Indeed	Linkedin	Papers
BLIP-2 [25]	Flan-T5-XL	14.3	20.0	18.2	20.0	13.9	40.0	34.2	23.8	16.7
InstructBLIP [9]	Flan-T5-XL	16.3	20.0	25.0	30.0	16.7	35.0	34.2	28.6	19.4
InstructBLIP Vicuna [9]	Vicuna-7B	12.2	24.4	22.7	30.0	19.4	42.5	34.2	23.8	36.1
LLaVA [30]	LLaMA-7B	16.3	17.8	25.0	30.0	27.8	42.5	42.1	23.8	41.7
MiniGPT-4 [58]	Vicuna-7B	16.3	22.2	25.0	30.0	27.8	47.5	31.6	16.7	33.3
VPGTrans [55]	LLaMA-7B	10.2	24.4	31.8	35.0	27.8	40.0	31.6	26.2	36.1
MultiModal-GPT [19]	Vicuna-7B	24.5	22.2	27.3	35.0	27.8	47.5	29.0	31.0	36.1
Otter [24]	LLaMA-7B	22.5	26.7	25.0	35.0	38.9	52.5	26.3	26.2	36.1
OpenFlamingo [3]	LLaMA-7B	24.5	22.2	27.3	35.0	27.8	47.5	29.0	31.0	36.1
LLaMA-Adapter V2 [14]	LLaMA-7B	20.4	26.7	22.7	30.0	30.6	42.5	34.2	16.7	30.6
GVT [45]	Vicuna-7B	18.4	20.0	20.5	30.0	33.3	42.5	39.5	21.4	38.9
mPLUG-Owl [50]	LLaMA-7B	16.3	24.4	25.0	35.0	30.6	45.0	31.6	26.2	30.6
Qwen-VL [4]	Qwen-7B	71.4	**51.1**	63.6	60.0	47.2	47.5	65.8	52.4	41.7
Qwen-VL-Chat [4]	Qwen-7B	**77.6**	46.7	63.6	**65.0**	47.2	45.0	68.4	57.1	47.2
LLaVA-1.5 [29]	Vicuna-7B	26.5	28.9	43.2	45.0	25.0	37.5	42.1	42.9	47.2
IDEFICS-9B-Instruct [21]	LLaMA-7B	32.7	24.4	22.7	35.0	27.8	50.0	29.0	26.2	36.1
InternLM-Xcomposer-VL [56]	InternLM-7B	34.7	42.2	47.7	60.0	25.0	52.5	57.9	42.9	44.4
VideoChat [26]	Vicuna-7B	10.2	22.2	25.0	35.0	22.2	45.0	31.6	14.3	33.3
Video-ChatGPT [36]	LLaMA-7B	14.3	26.7	22.7	30.0	27.8	45.0	26.3	21.4	41.7
Valley [35]	LLaMA-13B	28.6	20.0	22.7	25.0	33.3	42.5	26.3	26.2	41.7
Emu [41]	LLaMA-13B	35.5	12.2	24.4	36.4	35.0	27.8	47.5	36.8	33.3
NExt-GPT [48]	Vicuna-7B	16.1	18.4	26.7	27.3	30.0	25.0	25.0	23.7	16.7
SEED-LLaMA [17]	LLaMA2-Chat-13B	22.6	22.5	20.0	34.1	40.0	22.2	47.5	34.2	31.0
CogVLM [46]	Vicuna-7B	24.5	22.2	36.4	45.0	30.6	42.5	42.1	26.2	47.2
InternLM-Xcomposer-VL2 [11]	InternLM2-7B	53.1	**51.1**	50.0	65.0	38.9	**70.0**	81.6	59.5	47.2
InternLM-Xcomposer-VL2-4bit [11]	InternLM2-7B	38.8	26.7	34.1	35.0	30.6	37.5	55.3	45.2	36.1
LLaVA-Next [31]	Vicuna-7B	28.6	26.7	40.9	55.0	22.2	45.0	52.6	40.5	44.4
Yi-VL [20]	Yi-6B	40.8	28.9	27.3	45.0	30.6	35.0	42.1	38.1	41.7
SPHINX-v2-1k [15]	LLaMA2-7B	69.4	35.6	**65.9**	**65.0**	47.2	65.0	**81.6**	57.1	44.4
mPLUG-Owl2 [51]	LLaMA2-7B	28.6	26.7	31.8	50.0	19.4	42.5	36.8	33.3	41.7
SEED-X [18]	LLaMA2-13B	45.2	51.0	37.8	54.6	50.0	33.3	45.0	**73.7**	54.8
GPT-4V [1]	-	58.3	44.7	47.6	55.0	51.4	43.6	66.7	62.5	**62.9**
Gemini-Pro-Vision [42]	-	60.5	48.3	55.3	-	40.0	47.4	60.0	63.6	30.0
Claude-3-Opus [2]	-	23.3	34.7	37.8	20.5	**60.0**	33.3	57.5	57.9	40.5

important to note that for Gemini-Pro-Vision, we only report task performance when the model responds to over half of the valid data in the task.

4.2. Main Results

The evaluation results for various MLLMs across different categories of SEED-Bench-2-Plus are presented in Table 2. The detailed results of 63 types are presented in Table 3, Table 4, Table 5, Table 6, Table 7, and Table 8, respectively. Notably, GPT-4V surpasses a significant number of MLLMs, demonstrating superior performance across most evaluation types.

To provide a more comprehensive overview of model capabilities across different categories, we have visualized the ranking of each model within each category in Figure 4. Here, darker hues represent higher ranks. The top-performing MLLM, GPT-4V, exhibits competitive results across different categories.

4.3. Observations

Through a comprehensive and objective evaluation of various MLLMs across different text-rich scenarios in SEED-Bench-2-Plus, we have gleaned valuable insights that can guide future research efforts.

Comprehending text-rich data proves to be more complex. The majority of MLLMs achieve lower results on text-rich data, with the average accuracy rate being less than 40%. Considering the potential of MLLMs as multimodal agents, particularly website agents, a crucial capability is to analyze a variety of websites and generate accurate responses. The unsatisfactory results indicate that significant advancements are required before MLLMs can be effectively utilized as multimodal agents. This underlines the complexity of comprehending text-rich data and highlights the need for further research and development in this area to enhance the proficiency of MLLMs in text-rich scenarios.

Maps are more difficult for text-rich comprehension.

Table 8. Evaluation results of various MLLMs in different types of "Webs" (Part 2) in SEED-Bench-2-plus. The best (second best) is in bold (underline).

Model	Language Model	Poster	Project Page	Stackoverflow	Steam	Steam Game Setting	Tripadvisior	Twitter	Wikipedia	Yelp
BLIP-2 [25]	Flan-T5-XL	41.5	23.1	30.3	23.5	17.7	21.4	42.1	34.7	21.7
InstructBLIP [9]	Flan-T5-XL	48.8	15.4	27.3	17.7	14.7	21.4	36.8	38.8	26.1
InstructBLIP Vicuna [9]	Vicuna-7B	41.5	38.5	27.3	29.4	32.4	23.8	42.1	53.1	19.6
LLaVA [30]	LLaMA-7B	41.5	15.4	42.4	35.3	26.5	21.4	42.1	40.8	28.3
MiniGPT-4 [58]	Vicuna-7B	43.9	23.1	30.3	29.4	41.2	28.6	36.8	44.9	19.6
VPGTrans [55]	LLaMA-7B	43.9	23.1	33.3	32.4	29.4	16.7	15.8	38.8	30.4
MultiModal-GPT [19]	Vicuna-7B	39.0	30.8	39.4	29.4	29.4	23.8	36.8	49.0	19.6
Otter [24]	LLaMA-7B	39.0	23.1	42.4	26.5	26.5	26.2	36.8	40.8	21.7
OpenFlamingo [3]	LLaMA-7B	39.0	30.8	39.4	29.4	29.4	23.8	36.8	49.0	19.6
LLaMA-Adapter V2 [14]	LLaMA-7B	46.3	30.8	42.4	38.2	32.4	16.7	47.4	36.7	30.4
GVT [45]	Vicuna-7B	31.7	30.8	30.3	35.3	29.4	28.6	36.8	40.8	17.4
mPLUG-Owl [50]	LLaMA-7B	41.5	30.8	36.4	29.4	35.3	23.8	36.8	44.9	28.3
Qwen-VL [4]	Qwen-7B	61.0	76.9	69.7	35.3	64.7	59.5	73.7	40.8	54.4
Qwen-VL-Chat [4]	Qwen-7B	48.8	69.2	**72.7**	38.2	**67.7**	64.3	73.7	44.9	56.5
LLaVA-1.5 [29]	Vicuna-7B	58.5	53.9	51.5	35.3	58.8	31.0	52.6	42.9	21.7
IDEFICS-9B-Instruct [21]	LLaMA-7B	36.6	23.1	39.4	41.2	35.3	31.0	36.8	46.9	28.3
InternLM-Xcomposer-VL [56]	InternLM-7B	53.7	38.5	45.5	38.2	35.3	31.0	63.2	36.7	34.8
VideoChat [26]	Vicuna-7B	36.6	30.8	33.3	29.4	35.3	31.0	52.6	40.8	10.9
Video-ChatGPT [36]	LLaMA-7B	41.5	38.5	36.4	29.4	32.4	31.0	42.1	38.8	21.7
Valley [35]	LLaMA-13B	29.3	38.5	45.5	32.4	32.4	23.8	52.6	36.7	23.9
Emu [41]	LLaMA-13B	38.9	51.2	38.5	33.3	24.2	21.4	42.1	59.2	23.9
NExt-GPT [48]	Vicuna-7B	27.8	31.7	38.5	36.4	21.2	23.8	36.8	26.5	21.7
SEED-LLaMA [17]	LLaMA2-Chat-13B	38.9	46.3	15.4	45.5	30.3	26.2	31.6	44.9	26.1
CogVLM [46]	Vicuna-7B	46.3	30.8	33.3	29.4	50.0	23.8	47.4	32.7	26.1
InternLM-Xcomposer-VL2 [11]	InternLM2-7B	63.4	**84.6**	63.6	50.0	41.2	57.1	**84.2**	51.0	52.2
InternLM-Xcomposer-VL2-4bit [11]	InternLM2-7B	43.9	38.5	27.3	50.0	38.2	23.8	68.4	32.7	21.7
LLaVA-Next [31]	Vicuna-7B	51.2	61.5	48.5	33.3	55.9	33.3	63.2	42.9	28.3
Yi-VL [20]	Yi-6B	51.2	30.8	42.4	35.3	35.3	38.1	63.2	36.7	26.1
SPHINX-v2-1k [15]	LLaMA2-7B	58.5	76.9	57.6	63.6	64.7	**69.1**	79.0	57.1	**58.7**
mPLUG-Owl2 [51]	LLaMA2-7B	41.5	15.4	33.3	50.0	38.2	14.3	57.9	42.9	23.9
SEED-X [18]	LLaMA2-13B	36.1	73.2	69.2	42.4	57.6	54.8	79.0	46.9	56.5
GPT-4V [1]	-	68.3	83.3	51.5	**70.6**	42.4	40.0	83.3	**65.3**	47.6
Gemini-Pro-Vision [42]	-	**82.4**	75.0	70.4	36.8	57.7	44.4	81.8	35.3	54.8
Claude-3-Opus [2]	-	41.7	58.5	61.5	39.4	57.6	40.5	42.1	63.3	39.1

Maps are inherently complex and multidimensional, frequently featuring multiple overlapping layers of information. They not only display geographical details but also contain various symbols, colors, and texts to represent different types of information, such as topographical features, political boundaries, and points of interest. This complexity can pose challenges for models to accurately interpret and understand the full context of maps, compared with charts and webs.

The performance of MLLMs varies significantly across different data types. Different data types, such as knowledge graph and matrix diagram, which both fall under the category of charts, differ considerably in their complexity and structure. Some data types may require more sophisticated understanding and processing capabilities than others due to their inherent characteristics. As such, it is reasonable to observe a performance discrepancy of models across different data types. This finding underscores the need for MLLMs to be robust and adaptable, capable of handling a diverse range of data types in text-rich scenarios.

5. Conclusion

In this study, we present SEED-Bench-2-Plus, an extensive benchmark specifically designed for evaluating Multimodal Large Language Models (MLLMs) in text-rich scenarios. SEED-Bench-2-Plus comprises 2K multiple-choice questions, each annotated by humans, and span across 3 brod categories across 63 data types. We carry out a comprehensive evaluation of 31 noteworthy open-source MLLMs and 3 closed-source MLLMs, examining and comparing their performances to derive valuable insights that can guide future research. Serving as a valuable supplement to SEED-Bench-2 [22], both the dataset and the evaluation code of SEED-Bench-2-Plus are made publicly available. Additionally, we consistently maintain a leaderboard to facilitate the advancements in the field of text-rich visual comprehension with MLLMs.

References

[1] Gpt-4v(ision) system card. 2023. 1, 2, 5, 6, 7, 8, 9, 10

[2] Anthropic news: Claude 3 family, 2024. 2, 5, 6, 7, 8, 9, 10

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 5, 6, 7, 8, 9, 10

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[5] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023. 2, 4

[6] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 2

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 7

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 1, 2

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[10] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 2

[11] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 5, 6, 7, 8, 9, 10

[12] FastChat. Vicuna. https://github.com/lm-sys/FastChat, 2023. 1, 2

[13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2, 4

[14] 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. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 5, 6, 7, 8, 9, 10

[15] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024. 5, 6, 7, 8, 9, 10

[16] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 2

[17] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[18] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 5, 6, 7, 8, 9, 10

[19] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[20] Hugging Face. Yi-VL-6B. https://huggingface.co/01-ai/Yi-VL-6B, 2024. 5, 6, 7, 8, 9, 10

[21] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[22] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023. 1, 2, 10

[23] 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. 2

[24] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023. 2, 5, 6, 7, 8, 9, 10

[26] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[27] Yu Lili, Shi Bowen, Pasunuru Ram, Miller Benjamin, Golovneva Olga, Wang Tianlu, Babu Arun, Tang Binh, Karrer Brian, Sheynin Shelly, Ross Candace, Polyak Adam, Howes Russ, Sharma Vasu, Xu Jacob, Singer Uriel, Li (AI) Daniel, Ghosh Gargi, Taigman Yaniv, Fazel-Zarandi Maryam, Celikyilmaz Asli, Zettlemoyer Luke, and Aghajanyan Armen. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. 2023. 2

[28] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021. 7

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 4, 5, 6, 7, 8, 9, 10

[31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. 2024. 5, 6, 7, 8, 9, 10

[32] 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. 2, 4, 7

[33] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 2

[34] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 2

[35] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[36] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[37] OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2022. 1

[38] OpenAI. Gpt-4 technical report, 2023. 1

[39] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1, 2

[40] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 1, 2

[41] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 2, 5, 6, 7, 8, 9, 10

[42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[44] Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. *arXiv preprint arXiv:2401.13311*, 2024. 2

[45] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. What makes for good visual tokenizers for large language models? *arXiv preprint arXiv:2305.12223*, 2023. 4, 5, 6, 7, 8, 9, 10

[46] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 5, 6, 7, 8, 9, 10

[47] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 2

[48] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 2, 5, 6, 7, 8, 9, 10

[49] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 2, 4

[50] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[51] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 5, 6, 7, 8, 9, 10

[52] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023. 2, 4

[53] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2

[54] 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. 2, 4

[55] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. abs/23045.01278, 2023. 5, 6, 7, 8, 9, 10

[56] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 1, 2, 5, 6, 7, 8, 9, 10

[57] Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179*, 2023. 2

[58] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 5, 6, 7, 8, 9, 10