SEED-Bench-H: Hierarchical Benchmarking Multimodal Large Language Models

Bohao Li[1,2], Yuying Ge[2], Yixiao Ge[2,3], Ying Shan[2,3], Ruimao Zhang[1*]

[1*]The Chinese University of HongKong, Shenzhen(CUHK-SZ).
[2]Tencent AI Lab.
[3]ARC Lab, Tencent PCG.

Abstract

Multimodal large language models (MLLMs), which build upon the foundation of powerful large language models (LLMs), have recently showcased remarkable abilities in generating not only text but also images, given interleaved multimodal inputs (acting like a combination of GPT-4V and DALL-E 3). However, existing MLLM benchmarks are limited to assessing only the comprehension ability of single image-text inputs, falling short of keeping pace with the advancements in MLLMs. A comprehensive benchmark is crucial for examining the progress and identifying the limitations of current MLLMs. In this work, we classify the capabilities of MLLMs into hierarchical levels from $\boldsymbol{L_0}$ to $\boldsymbol{L_4}$, based on the modalities they can accept and generate, and introduce SEED-Bench-H, a comprehensive benchmark that evaluates the **hierarchical** capabilities of MLLMs. Specifically, SEED-Bench-H consists of 28K multiple-choice questions with precise human annotations, spanning 34 dimensions, including the evaluation of both text and image generation. Multiple-choice questions with ground truth options derived from human annotations enable an objective and efficient assessment of model performance, eliminating the need for human or GPT intervention during evaluation. We further assess the performance of 30 prominent open-source MLLMs and 3 closed-source MLLMs, summarizing valuable observations. By uncovering the limitations of existing MLLMs through extensive evaluations, we aim for SEED-Bench-H to provide insights that will drive future research toward the goal of General Artificial Intelligence. Dataset and evaluation code are available at https://github.com/AILab-CVC/SEED-Bench.

---

*Correspondence Author.

†This paper is a comprehensive integration of previous SEED-Bench series (SEED-Bench [1], SEED-Bench-2 [2], SEED-Bench-2-Plus [3]), with additional evaluation dimensions.

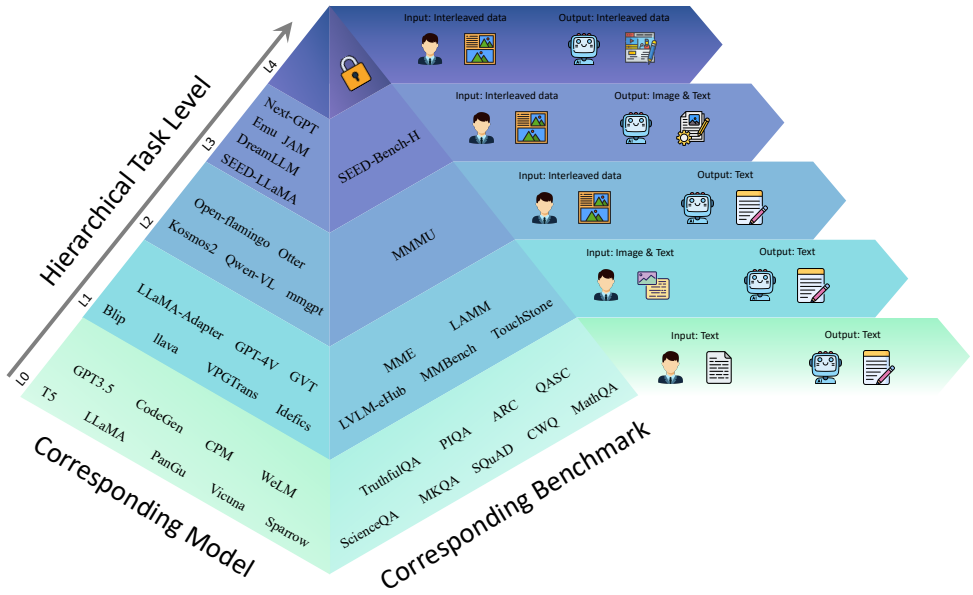

Fig. 1 Overview of **hierarchical capability levels** of MLLMs from $\boldsymbol{L_0}$ to $\boldsymbol{L_4}$, where higher level encompasses lower capability tiers. $\boldsymbol{L_4}$ entails producing open-form interleaved image and text output given arbitrary interleaved image-text inputs, which indicates that the model can handle any form of visual text input and output any form of visual text. Models and corresponding evaluation benchmarks at each pyramid tier are illustrated. SEED-Bench-H covers the assessment of MLLMs up to $\boldsymbol{L_3}$.

Keywords: Benchmark, Multimodal, Large Language Model

1 Introduction

In recent years, Large Language Models (LLMs) [4–8] have displayed impressive capabilities in understanding, reasoning, and generating texts across a wide range of open-ended tasks. Building on the strong generality of LLMs, Multimodal Large Language Models (MLLMs) [9–17, 17–26] have demonstrated exceptional capabilities in comprehending multimodal data through predicting open-form texts. Recent work [27–32] further empower LLMs with the ability to generate images beyond texts (acting like a combination of GPT-4V [33] and DALL-E 3 [34]), as they argue that the emergence of multimodal capabilities requires that text and image can be represented and processed interchangeably in a unified autoregressive Transformer. However, despite the extensive capabilities of MLLMs, existing MLLM benchmarks [35–39] primarily focus on evaluating single image-text comprehension, thus failing to fully showcase the progress and limitations of current MLLMs. The lag of benchmarks behind the rapid development of MLLMs hinders the exploration and evolution of models.

In this work, we categorize the capabilities of MLLMs into hierarchical levels ranging from L_0 to L_4 based on the modalities they can accept and generate, as depicted

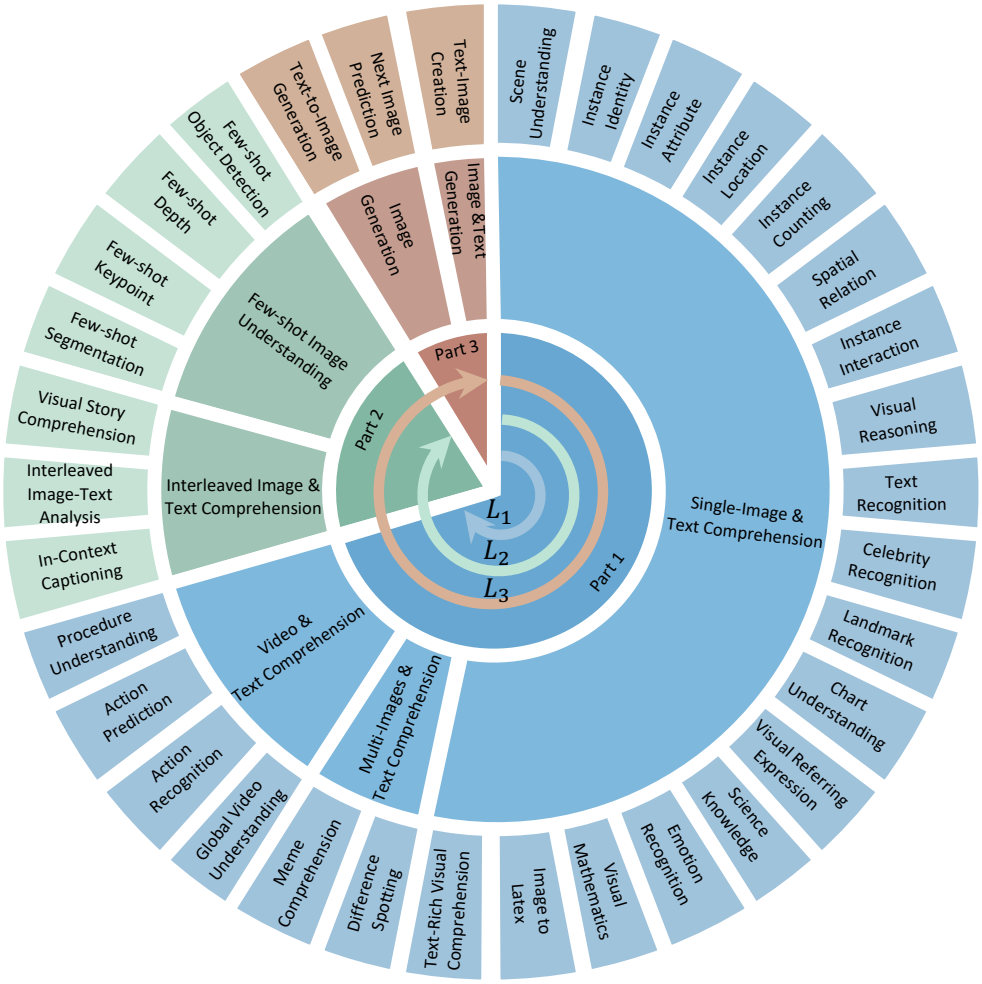

Fig. 2 Overview of 34 evaluation dimensions in SEED-Bench-H, which consists of three parts, with part-1 constituting L_1, part-1&2 constituting L_2, and part-1&2&3 constituting L_3.

in Fig. 1. Building upon LLMs, the lowest-tier capability L_0 involves generating texts given text inputs, while the highest-tier capability L_4 entails producing open-form interleaved image and text output given arbitrary interleaved image-text inputs. Achieving the capability L_4 is a critical milestone on the path towards General Artificial Intelligence (AGI), as a human-level AI should be able to effortlessly digest and create multimodal content. In the capability pyramid, higher levels inherently include the capabilities of lower tiers. This hierarchical categorization not only clearly illustrates the current progress of MLLMs, but also provides a well-defined roadmap for future research.

We propose SEED-Bench-H, a comprehensive benchmark that evaluates the **hierarchical** capabilities of MLLMs up to L_3, including the generation of both texts

Table 1 Comparisons between existing MLLM benchmarks. "H/G Evaluation" denotes whether human or GPT is used for evaluation.

Benchmark	Visual Modality	Evaluation Level	Customized Question	#Answer Annotation	Answer Type	H/G Evaluation	#Models
LLaVA-Bench [11]	Image	L_1	✓	150	free-form	GPT	4
OCR-Bench [40]	Image	L_1	✗	-	free-form	N/A	6
MME [35]	Image	L_1	✓	2194	Y/N	N/A	10
M3Exam [41]	Image	L_1	✓	12317	A/B/C/D	N/A	7
LAMM [36]	Image(s) & Point cloud	L_1	✗	-	free-form	GPT	4
LVLM-eHub [37]	Image	L_1	✗	-	free-form	Human	8
MMBench [38]	Image(s)	L_1	✓	2974	free-form	GPT	14
VisIT-Bench [42]	Images	L_1	✓	592	free-form	Human/GPT	14
MM-VET [43]	Image	L_1	✓	200	free-form	GPT	9
Touchstone [39]	Image(s)	L_1	✓	908	free-form	GPT	7
SciGraphQA [44]	Image	L_1	✓	3K	free-form	N/A	4
Q-bench [45]	Image	L_1	✓	3489	Y/N & free-form	N/A	15
MathVista [46]	Image	L_1	✓	735	A/B/C/D & free-form	N/A	12
CONTEXTUAL [47]	Image(s)	L_1	✓	506	free-form	GPT	13
MMMU [48]	Image(s)	L_2	✓	11.5K	A/B/C/D & free-form	N/A	23
SEED-Bench-H	Image(s) & Video	L_3	✓	28479	A/B/C/D	N/A	33

and images given interleaved image-text inputs. As shown in Fig. 1, SEED-Bench-H comprises three parts, where part-1 constitutes capability level L_1 for images and text comprehension, part-1&2 constitutes capability level L_2 for interleaved image-text comprehension, and part-1&2&3 constitute capability level L_3 for image and text generation. To the best of our knowledge, SEED-Bench-H is the first benchmark that provides hierarchical evaluations of MLLMs, effectively showcasing the range of model capabilities.

Specifically, SEED-Bench-H consists of 28K multiple-choice questions with ground truth answers derived from human annotation (12× larger than MME [35] and 9× larger than MMBench [38] as shown in Tab. 1). SEED-Bench-H spans 34 evaluation dimensions in 3 parts, enabling a comprehensive assessment of MLLMs' performance across diverse aspects, such as single image, multiple image, video, interleaved image and text, few-shot image, image generation, and image and text generation, as illustrated in Figure 2. We employ three approaches for the generation of multiple-choice questions, including (1) a sophisticated pipeline utilizing foundation models, (2) the adaptation of existing datasets, and (3) a combination of human creation and GPT assistance. We further incorporate an automated filtering mechanism and manual verification process to ensure the quality of questions and the accuracy of ground truth answers. Different from existing MLLM benchmarks [11, 36–39, 42, 43] that employ human annotators or GPT to evaluate open-form output, resulting in compromised efficiency, increased subjectivity, and reduced assessment accuracy, SEED-Bench-H provides multiple-choice questions, which restricts the model's output to A/B/C/D options. This approach facilitates the convenient calculation of accuracy, serving as an objective metric for evaluation.

Based on SEED-Bench-H, we conduct a comprehensive evaluation of 30 prominent open-source MLLMs and 3 significant closed-source MLLMs. Our evaluation results yield the following three key findings: (1) Existing MLLMs have not yet reached the ceiling level of capability L_1 for the comprehension of fixed-form images and texts, with even the top-ranked model achieving only a 60% accuracy rate. MLLMs, in particular, exhibit poor performance in certain dimensions, such as understanding charts and visual mathematics. (2) MLLMs achieve less satisfactory performance at capability L_2 than that at L_1, which indicates that it is more challenging for MLLMs to comprehend free-form interleaved image-text inputs, since most MLLMs are trained on structured image-caption pairs. (3) At present, only a few MLLMs can attain capability L_3,

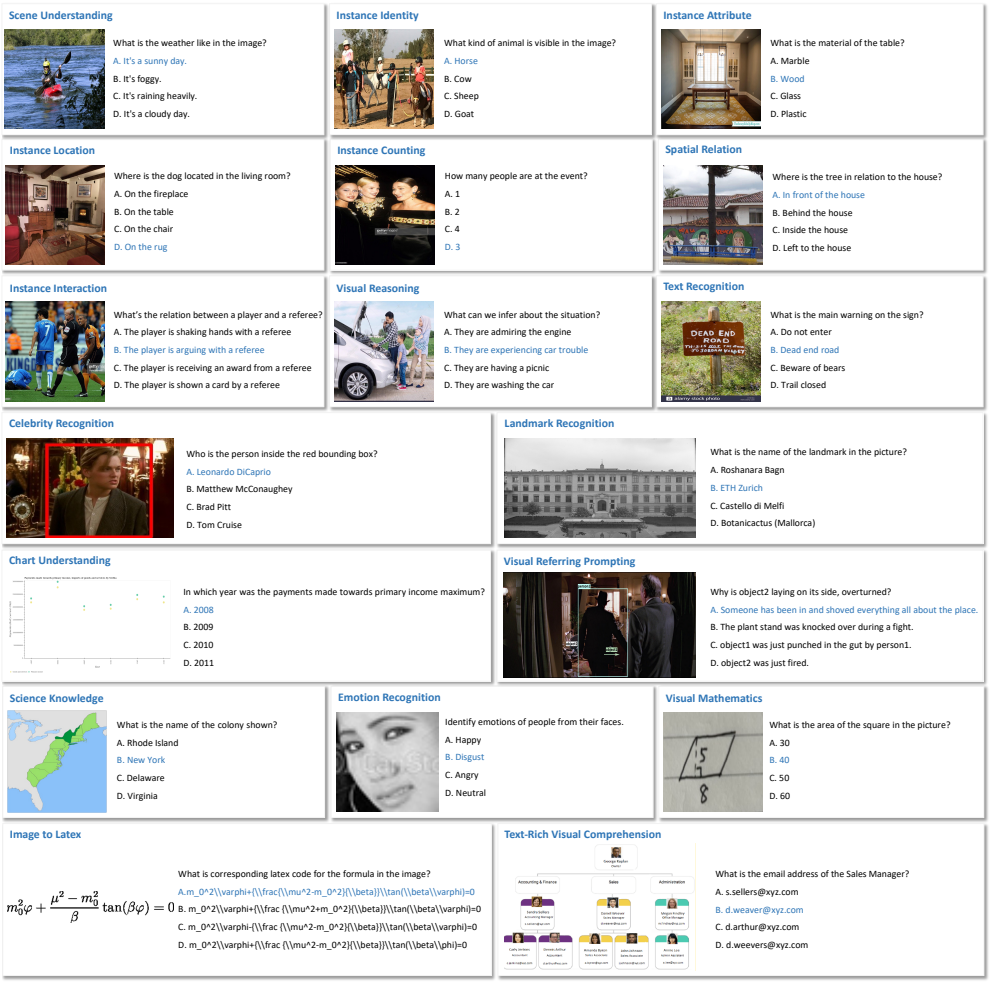

Fig. 3 Data samples from a subset of evaluation dimensions in part-1 with single image as input, which encompasses capability L_1 in SEED-Bench-H.

which requires models to output content in multiple modalities. A universal MLLM that unifies the generation of images and texts is currently underexplored. We have launched an evaluation platform and consistently maintain a leaderboard for assessing and comparing model performance.

2 Related Work

Multimodal Large Language Models. Following the remarkable success of Large Language Models (LLM) [4, 7, 8], recent studies have focused on generative Multimodal Large Language Models (MLLMs) [9–17, 21–24] to enhance multimodal comprehension by aligning visual features of pre-trained image encoders with LLMs

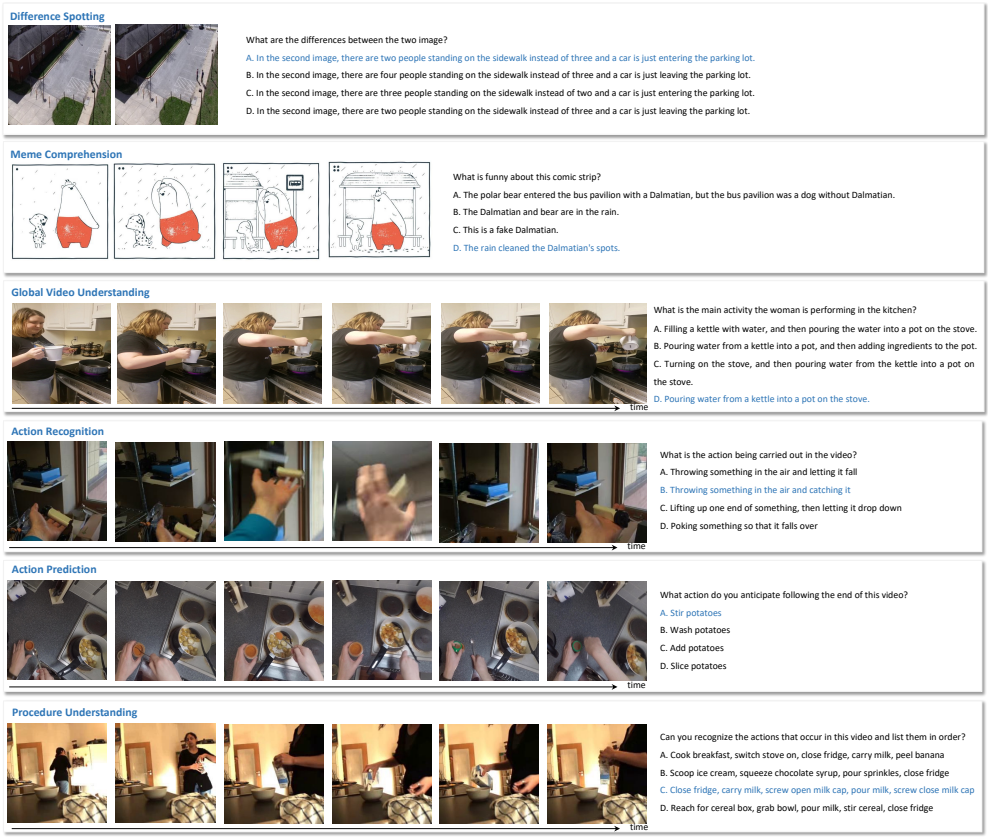

Fig. 4 Data samples from a subset of evaluation dimensions in part-1 with multiple images or videos as inputs, which encompasses capability L_1 in SEED-Bench-H.

on image-text datasets. Some research [18–20] has further considered video inputs and leveraged the vast capabilities of LLMs for video understanding tasks. Recent work [27–32, 49] has made significant strides in equipping MLLMs with the capacity to generate images beyond text. In SEED-Bench-H, we provide a comprehensive and objective evaluation of these models to thoroughly assess their hierarchical capabilities.

Benchmarks for Multimodal Large Language Models. With the rapid advancement of Multimodal Large Language Models (MLLMs), several concurrent works [35–39] have proposed various benchmarks for evaluating MLLMs. However, they remain limited to assessing only the model's ability to predict text given single image-text inputs, failing to keep pace with the advancements in multimodal model capabilities. For instance, GVT [50] 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 [37] 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

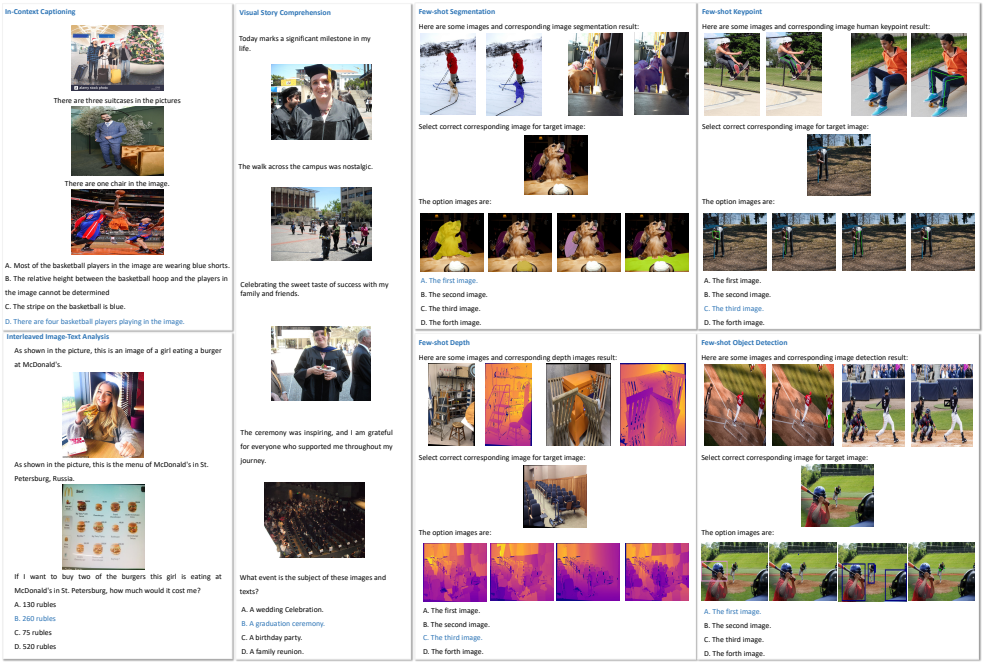

Fig. 5 Data samples of evaluation dimensions in part-2 with interleaved image-text or few-shot examples as inputs, which encompasses capability L_2 together with dimensions in L_1.

compare the predictions of models. The involvement of human annotators during evaluation introduces bias and incurs significant costs. LLaVA-Bench [11], LAMM [36], and Touchstone [39] utilize GPT to evaluate the answers' relevance and accuracy to the ground truth. The reliance on entity extraction and GPT metric can impact the accuracy and reliability of the evaluation. MME [35] and MMBench [38] aim to enhance the objective evaluation of MLLMs by constructing 2194 True/False Questions and 2974 Multiple Choice Questions across a variety of ability dimensions respectively. Considering the limited scale of these benchmarks, their evaluation results may exhibit instability. In this work, we introduce SEED-Bench-H to evaluate the hierarchical capabilities of MLLMs, including the generation of both text and images. It contains 28K human-annotated multiple-choice questions covering 34 evaluation dimensions.

3 SEED-Bench-H

3.1 Hierarchical Capability Levels

We categorize the capabilities of MLLMs into hierarchical levels from L_0 to L_4, based on input and output modalities, where higher level encompasses lower capability tiers, as illustrated in Fig. 1. SEED-Bench-H covers the assessment of MLLMs up to L_3. The detailed categorization of capability level is illustrated as below,

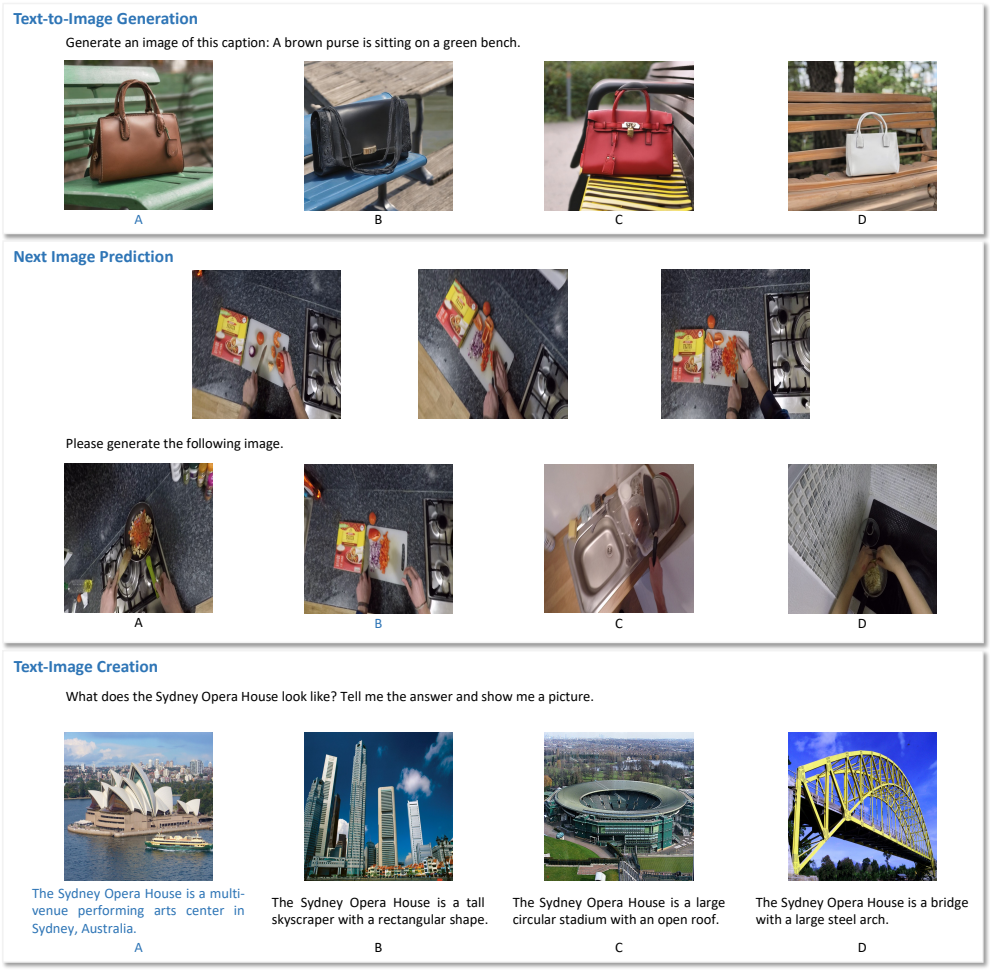

Fig. 6 Data samples of evaluation dimensions in part-3 with images and texts as outputs, which encompasses capability L_3 together with dimensions in L_2.

Level L_0: Building upon LLMs, the most fundamental capability of MLLMs generating text based on provided text inputs, which does not necessitate specific evaluation within the MLLM benchmark.

Level L_1: MLLMs at this capability level should possess the ability to comprehend multimodal inputs in a fixed format, *i.e.*, image or multiple images (video input can be regarded as multiple images) and then texts. Current MLLM benchmarks only evaluate this capability level with single image and text as inputs.

Level L_2: MLLMs at this capability level should be able to understand multimodal inputs with open-form interleaved image-text data, which aligns with the multimodal inputs encountered in real-life scenarios.

Level L_3: Besides the inherent ability of LLMs to generate texts, MLLMs at this capability level should also be proficient in producing images, as advanced MLLMs are expected to process and represent multimodal content on both input and output sides. **Level** L_4: MLLMs at the highest capability level should possess the ability to process and generate interleaved image-text content in an open-form format, which is an essential step towards achieving general artificial intelligence. We will incorporate evaluations of this capability level in our future work.

3.2 Evaluation Dimensions

As shown in Fig. 2, SEED-Bench-H comprises a total of 34 evaluation dimensions, which constitute three capabilities levels, from L_1 to L_3. Since the higher level encompasses lower capability tiers, we further divide the evaluation dimensions of L_3 into three non-overlapping parts: part-1 forms level L_1, part-2 combined with part-1 constitutes level L_2, part-3, part-2 and part-1 form level L_3 together. We introduce the dimensions of each part in detail below.

3.2.1 Part-1

The dimensions of part-1 evaluate MLLMs' comprehension of multimodal inputs in a fixed format, and can be further grouped into three sub-parts based on the types of visual inputs: (1) Single-Image & Text, (2) Multiple-Images & Text, (3) Video & Text. The dimensions within Single-Image & Text Comprehension, Multiple-Images & Text Comprehension, and Video & Text Comprehension are visually represented in Fig. 7.

3.2.1.1 Single-Image & Text Comprehension

This subpart comprises various evaluation dimensions, including Scene Understanding, Instance Identity, Instance Attribute, Instance Location, Instance Counting, Spatial Relation, Instance Interaction, Visual Reasoning, Text Recognition, Celebrity Recognition, Landmark Recognition, Chart Understanding, Visual Referring Expression, Science Knowledge, Emotion Recognition, Visual Mathematics, Image to Latex, and Text-Rich Visual Comprehension. These dimensions assess MLLMs' comprehension of image-text pairs from a broad range of aspects, encompassing global/object-level understanding, recognition/reasoning, and various specialized domains.

- Scene Understanding: This dimension focuses on the global information in an image and requires a holistic understanding to answer questions about the overall scene.
- Instance Identity: This dimension involves identifying specific instances in an image, including the existence or category of particular objects, evaluating a model's object recognition capabilities.
- Instance Attribute: This dimension pertains to an instance's attributes, such as color, shape, or material, assessing a model's understanding of an object's visual appearance.
- Instance Location: This dimension concerns the absolute position of a specified instance, requiring a model to accurately localize the object referred to in the question.

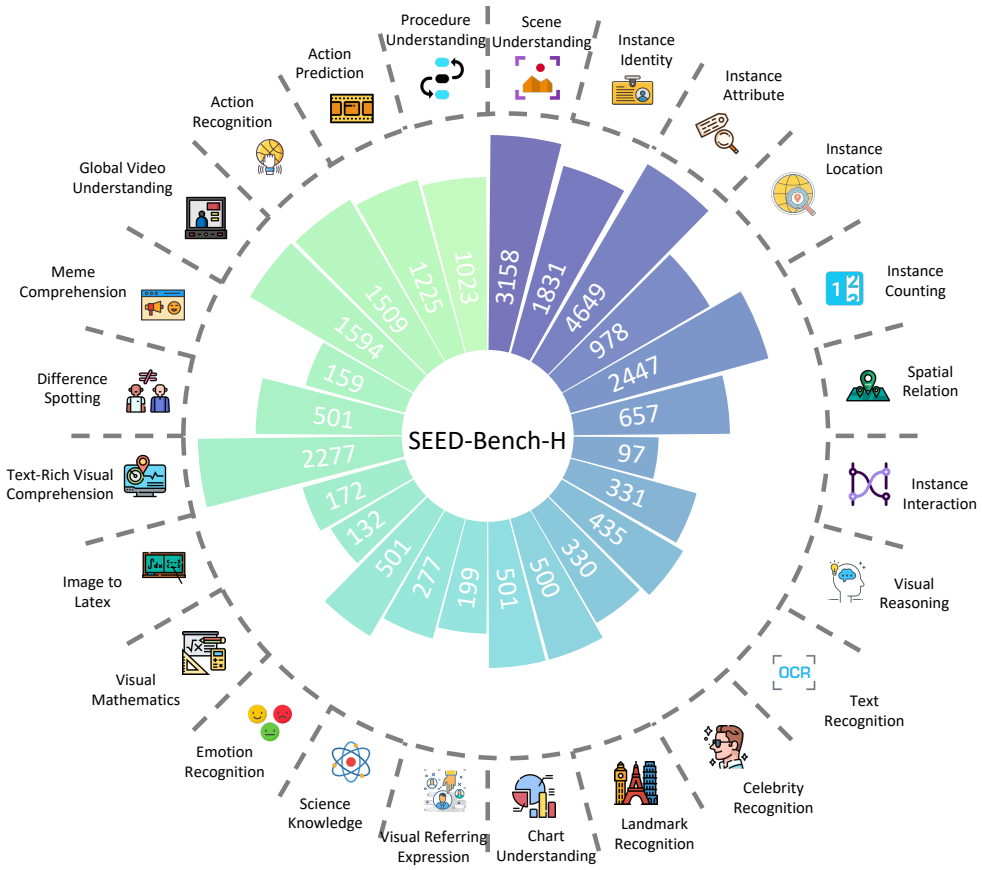

Fig. 7 Overview of 24 evaluation dimensions in SEED-Bench-H capability L_1. The number in the bar denotes the number of multiple-choice questions in each dimension.

- Instance Counting: This dimension requires the model to count the number of specific objects in the image, demonstrating an understanding of all objects and successfully counting the instances of the referred object.
- Spatial Relation: This dimension requires a model to ground two mentioned objects and recognize their relative spatial relation within the image.
- Instance Interaction: This dimension involves recognizing the state relation or interaction relations between two humans or objects.
- Visual Reasoning: This dimension evaluates a model's ability to reason based on visual information, necessitating a comprehensive understanding of the image and the application of commonsense knowledge to answer questions correctly.
- Text Recognition: In this dimension, the model should answer questions about textual elements in the image.

- Celebrity Recognition: This dimension focuses on identifying well-known public figures in images, evaluating a model's ability to recognize celebrity faces and names and understand their relevance in the given context.
- Landmark Recognition: In this dimension, the model is required to recognize and identify famous landmarks or locations in the image, understanding visual features and contextual information associated with these landmarks.
- Chart Understanding: This dimension requires the model to interpret and extract information from various chart types, such as line graphs, evaluating its ability to understand visual data representations and derive meaningful insights.
- Visual Referring Expression: In this dimension, the model is required to answer relevant questions based on the visual content of the image, assessing its ability to understand the scene and engage in meaningful visual dialogue.
- Science Knowledge: This dimension evaluates a model's ability to integrate multiple knowledge sources and apply commonsense reasoning to answer image-related questions, requiring an understanding of context, background information, and relationships between objects and events in the scene.
- Emotion Recognition: This dimension focuses on recognizing and interpreting emotions expressed by human faces in images, evaluating the model's ability to understand facial expressions and associate them with corresponding emotional states.
- Visual Mathematics: In this dimension, the model is required to solve mathematical problems or equations based on the visual content of the image, assessing its ability to understand and apply mathematical concepts and operations to real-world scenarios.
- Image to Latex: In this dimension, the model is tasked with providing the corresponding LaTeX code based on the given equation image. This task tests the model's ability to recognize and interpret mathematical symbols from an equation image and translate them into accurate LaTeX code, demonstrating its proficiency in optical character recognition, mathematical notation recognition, and LaTeX syntax recognition. Additionally, the model must showcase attention to detail, problem-solving skills, and resilience in handling potential image issues.
- Text-Rich Visual Comprehension: In this dimension, we take into account three dimensions that cover different types of text-rich data, namely, charts, maps, and webs to assess the capabilities of MLLMs in managing text-rich data. For charts data, 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. For maps data, 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. For webs data, 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.

3.2.1.2 Multiple-Images & Text Comprehension

This subpart includes Difference Spotting and Meme Comprehension, which evaluate MLLMs' capability to extract information and discern differences given multiple images.

- Difference Spotting: In this dimension, the model is required to identify differences between two images, assessing its ability to recognize subtle variations in visual elements and understand the significance of these differences.
- Meme Comprehension: This dimension requires the model to comprehend and interpret internet memes, which often involve humor, sarcasm, or cultural references. It evaluates the model's ability to recognize visual and textual meme elements and understand their intended meaning and context.

3.2.1.3 Video & Text Comprehension

This subpart comprises Global Video Understanding, Action Recognition, Action Prediction, and Procedure Understanding, which assess MLLMs' abilities for fine-grained action recognition, temporal relationship understanding, and temporal reasoning.

- Global Video Understanding: In this dimension, the model is required to answer questions from different aspects of a video's content, involving the understanding of key events, actions, and objects in the video, as well as recognizing their importance and relevance in the overall context of the video.
- Action Recognition: This dimension requires the model to recognize actions shown in videos, evaluating its ability to capture temporal dynamics, physical motions, human actions, and dynamic interactions between objects.
- Action Prediction: This dimension aims to predict future actions through preceding video segments, requiring an understanding of contextual information from videos and temporal reasoning.
- Procedure Understanding: This dimension necessitates that the model captures key actions and performs temporal ordering on them, evaluating its ability for temporally fine-grained understanding and procedure reasoning.

3.2.2 Part-2

Part-2 evaluates MLLMs' comprehension of arbitrary interleaved image-text inputs, which includes 2 subparts: (1) Interleaved Image & Text Comprehension, and (2) Few-shot Image Understanding.

3.2.2.1 Interleaved Image & Text Comprehension

This subpart encompasses In-Context Captioning, where two examples of image-caption pairs and an image are provided, and the model is expected to describe the specific aspect of the image. It also includes Interleaved Image-Text Analysis, where the model answers questions based on images and texts with varying quantities and positions, and Visual Story Understanding, where the model is tasked with extracting

information from stories composed of images and texts, and answering corresponding questions based on the acquired information.

- In-Context Captioning: This dimension highlights a model's ability to learn and adapt its understanding based on the provided image context. It assesses the model's capacity to integrate new information, identify patterns, and generate predictions for the target image.
- Interleaved Image-Text Analysis: In this dimension, the model is required to process and understand data presented in an interleaved or mixed format, such as images combined with text. It assesses the model's ability to integrate multiple information modalities and derive meaningful insights from the combined data.
- Visual Story Understanding: In this dimension, the model is tasked with extracting information from stories that consist of images and texts, and answering corresponding questions based on the acquired information.

3.2.2.2 Few-shot Image Understanding

In this subpart, we propose four dimensions to evaluate the MLLM's ability to extract information from given few-shot image examples. These include Few-shot Segmentation, where the model is tasked with selecting the correct target category object segmentation results from given examples; Few-shot Keypoint, where the model selects the correct human keypoint image based on a given example; Few-shot Depth, where the model is expected to select suitable depth maps following a given depth image example; and Few-shot Object Detection, where the model should possess the ability to accurately locate and identify the objects within the given image and select the correct detection result based on the provided example.

- Few-shot Segmentation: In this dimension, the model is tasked with selecting the correct target category object segmentation results from given examples. The model must accurately identify and classify different segments within the image and apply few-shot learning to generalize from limited examples, answering questions based on object boundaries and their relationships.
- Few-shot Keypoint: This dimension requires the model to select the correct human keypoint image based on a given example. The model is expected to use few-shot learning to apply this knowledge to new images.
- Few-shot Depth: This dimension aims for the model to select suitable depth maps following a given depth image example. The model is required to answer questions involving the three-dimensional structure or layout of the scene from limited examples.
- Few-shot Object Detection: This dimension necessitates that the model possesses the ability to accurately locate and identify the objects within the given image and select the correct detection result based on the provided example.

3.2.3 Part-3

The dimensions of Part-3 evaluate MLLMs' capability to generate images in addition to texts. This part can be divided into two subparts: (1) Image Generation, and (2) Image & Text Generation.

3.2.3.1 Image Generation

This subpart includes Text-to-Image Generation, where the model is expected to generate an image based on a caption prompt, and Next Image Generation, where the model is required to generate a subsequent image based on previous images. To evaluate an MLLM's ability in image generation, we introduce two tasks: Text-to-Image Generation and Next Image Prediction. These tasks assess the MLLM's generation ability from text and multiple images.

- Text-to-Image Generation: This dimension evaluates a model's ability to generate realistic and visually coherent images based on a given prompt. It requires the model to understand visual elements, relationships, and composition rules necessary for creating a plausible image.
- Next Image Prediction: In this dimension, the model is required to generate images that depict specific actions or events, such as a person running or a car driving. It assesses the model's ability to understand action dynamics and accurately represent them in a static visual format.

3.2.3.2 Image & Text Generation

To evaluate an MLLM's comprehensive ability in generation, we introduce the Text-Image Creation task. This task involves providing a question and requiring the MLLM to generate the corresponding image and text as a description.

- Text-Image Creation: Given a question, the model is required to provide a text-based answer and subsequently generate a corresponding image as an illustration.

3.3 Data Source

To create a benchmark with various evaluation dimensions, we need to collect data containing images with abundant visual information and videos with rich temporal dynamics, enabling us to construct diverse and challenging multiple-choice questions.

3.3.1 Part-1

To build questions for spatial understanding (dimensions 1-9), we utilize the CC3M [51] dataset with filtered samples to build questions. Specifically, considering the noisy original captions of CC3M, we generate captions for each image with Tag2Text [52]. We filter out images with no more than 5 nouns in their captions to ensure information richness in the remaining images for constructing questions. For limited data on text recognition, we use data from IC03 [53], IC13 [54], IIIT5k [55], and SVT [56] datasets to enlarge this dimension.

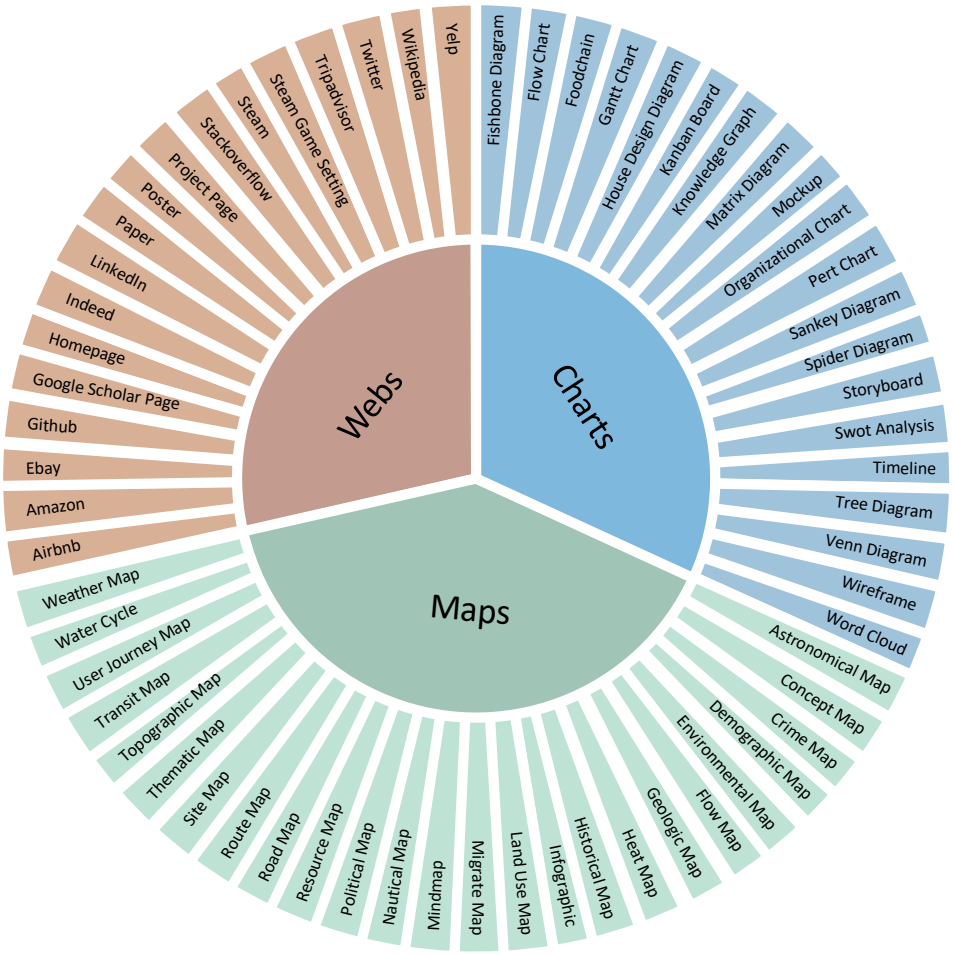

Fig. 8 Overview of 63 data types of charts, maps, and webs of text-rich visual comprehension data.

For celebrity and landmark recognition, we use data from MME [35], MMBench [38], and Google Landmark Dataset v2 [57] train set. We employ GPT-4 to generate confusing options for celebrity recognition in MME and randomly selecting other landmark names to generate confusing options in landmark recognition.

We utilize PlotQA [58] test set for chart understanding, VCR [59] valid dataset for visual referring expression, ScienceQA [60] test set for science knowledge, FER2013 [61] test dataset for emotion recognition, and MME [35] dataset for visual mathematics.

For image-to-latex, we use the Im2Latex [62] dataset and GPT-4 [5] for generating interference options. For text-rich visual comprehension, we manually collect various types of charts, maps, and webs data from the internet, using GPT-4V to

generate questions and human annotators to enhance quality. Detailed text-rich data composition is illustrated in Fig. 8.

For difference spotting, we use the SD part of the MIMICIT [63] dataset, and for meme comprehension, we generate questions manually.

As global video understanding, we select the Charades [64] test dataset as the video source, as the videos in the dataset contain rich information. For each video , we use Tag2Text [52] to generate each second caption and GRIT [65] to generate each 5-second dense caption containing each object's location. We then use GPT-4 to integrate captions and generate corresponding questions based on these captions. After generation, we use GPT-4 to filter out questions that can be answered using only a single frame.

For action recognition and prediction, we adopt the Something-Something-v2 [66] and Epic-Kitchen 100 [67] datasets, with human annotators filtering questions. We also utilize the Breakfast [68] dataset for procedure understanding tasks.

3.3.2 Part-2

For in-context captioning, we use the ground-truth caption generated by instance attribute dimension and instance counting dimension. For each caption in the instance attribute, we use GPT-4 to classify. And we generate questions for interleaved image-text analysis data manually.

For visual story understanding, we introduce VIST [69] as the data source and employ GPT-4V [33] to generate corresponding QAs for a series of images. Specifically, we randomly select 2-4 images from 5-grams of photos to generate questions and utilize human annotation to enhance the quality of the annotations.

We use the MSCOCO [70] dataset to construct few-shot segmentation, keypoint, and object detection data. For few-shot depth data, we use the Middlebury stereo [71] dataset.

3.3.3 Part-3

For text-to-image generation, we modify target categories or attributes in the prompt of the CC-500 [72] and ABC-6k [72] datasets using GPT-4, and employ Stable-Diffusion-XL [73] for image generation, with human annotators filtering unqualified data. For the next image prediction dimension, we use the Epic-Kitchen 100 [67] dataset and start-end frames in action prediction dimensions.

What's more, we generate questions for text-image creation tasks by humans. This method allows us to ensure the quality and relevance of the questions, providing a more accurate assessment of the model's capabilities. By incorporating human-generated questions, we aim to create a comprehensive and challenging evaluation framework that truly tests the model's understanding of the intricate relationship between text and images.

3.4 Construction of Multiple-choice Questions

We employ three approaches to construct multiple-choice question covering 34 evaluation dimensions: (1) an automatic pipeline to generate questions for specific evaluation

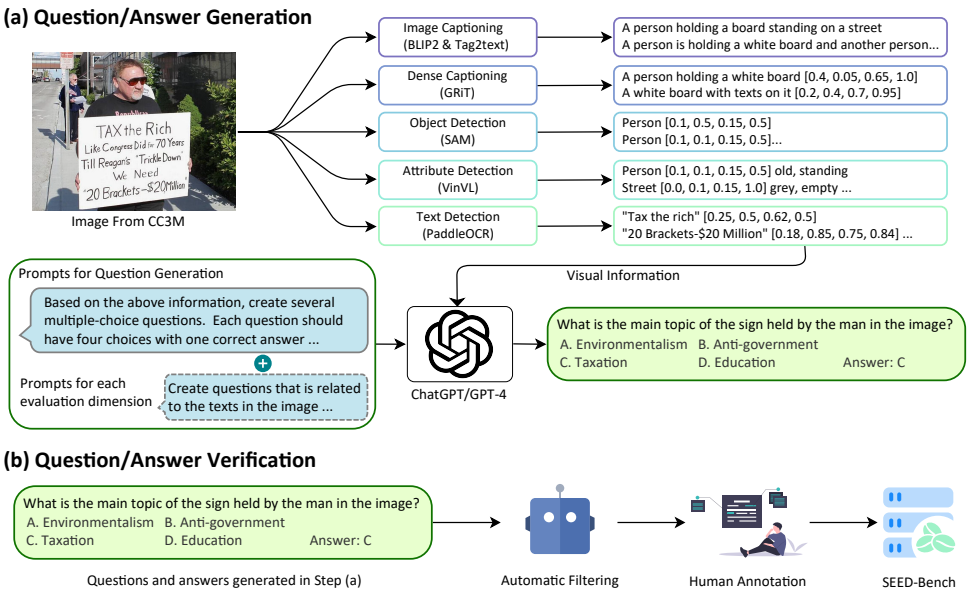

Fig. 9 Overview of automatic pipeline in SEED-Bench-H for generating multiple-choice questions. (a) We first leverage various foundation models to extract visual information including image-level captions, instance-level descriptions and textual elements. Based on specially designed prompts corresponding to specific evaluation dimension, ChatGPT/GPT-4 subsequently generates questions and four candidate options with one groundtruth answer. (b) We further filter out questions by utilizing LLMs and employ human annotators to select the correct option and classify each question into one evaluation dimension.

dimension, (2) tailor of existing datasets for the format of multiple-choice questions, (3) human creation combined with GPT.

3.4.1 Automatic pipeline.

As shown in Fig. 9, our pipeline for generating multiple-choice questions involves question/answer generation and verification.

3.4.1.1 Visual Information Extraction.

For constructing questions related to spatial understanding, we interpret the rich information in each image with texts using multiple pretrained models, so that ChatGPT/GPT-4 can understand the image and create questions accordingly. The extraction of visual information for images includes the following parts:

- **Image Captions.** Image captions contain the overall description of an image. We employ BLIP2 [74] and Tag2Text [52] to create captions for each image. The former creates captions for the whole image while the latter generates captions based on descriptions of each instance. The two models complement each other to depict the image content within a single sentence.

- **Instance Descriptions.** Besides captions which may ignore specific details in the image, we also extract visual information from images using instance-level descriptions, including object detection, attribute detection, and dense captions. Specifically, we use SAM [75] to segment each instance in the image and obtain their bounding boxes according to the segmentation results. The object labels are obtained using Tag2Text [52]. Besides, we also utilize attribute detector [76] to obtain the attributes of each instance in the image. Finally, we employ GRiT [65] to generate dense captions, which describe each detected instance in the image with a short sentence. These instance-level descriptions are complementary to the image captions, further enriching the visual information of each image.
- **Textual Elements.** Besides objects, the texts in the image also contain important information describing the image. We employ PaddleOCR [77] for detecting textual elements.

3.4.1.2 Question-Answer Generation.

After extracting visual information from the image, we task ChatGPT/GPT-4 with generating multiple-choice questions based on the extracted information or video annotations. For each of the spatial understanding evaluation, we carefully design prompts and ask ChatGPT/GPT-4 to create multiple choice questions with four candidate options based on the extracted visual information. We create questions with ChatGPT for all evaluation dimensions, except for the reasoning dimension, where we use GPT-4 [5] due to its exceptional reasoning capability. For each question, we ask ChatGPT/GPT-4 to create four choices with one correct option and three distractors. We try to make the multiple-choice questions challenging by encouraging the three wrong choices to be similar to the correct one.

3.4.1.3 Automatic Filtering.

Our benchmark aims at evaluating the multimodal vision-language understanding capability of MLLMs. However, we observe that some generated questions can be correctly answered by LLMs without seeing the image. We argue that such questions are not helpful to evaluate the visual comprehension capability of MLLMs. To this end, we feed the generated questions (without image) into three powerful LLMs, including Vicuna-7B [7], Flan-T5-XXL [4] and LLaMA-7B [8] and ask them to answer the questions. We empirically found that 5.52% of the generated questions can be correctly answered by all of the three LLMs. We filter out these questions from our benchmark.

3.4.1.4 Human Annotation.

To ensure the accuracy and objectiveness of SEED-Bench-H, we further employ human annotators to verify the generated question/answer pairs. Human annotators are asked to choose the correct answer for each multiple-choice question and categorize each question into one of the evaluation dimension. If one question can not be answered based on the visual input or does not have any correct choice or has multiple correct choices, it will be discarded by human annotators.

Table 2 Evaluation results of various MLLMs in different capability levels of SEED-Bench-H. $\bar{T}$ denotes the averaged accuracy across corresponding dimensions, and $R_{\bar{T}}$ denotes the rank based on the the averaged accuracy. The evaluation dimensions of part-2, together with L_1, encompass L_2, while the evaluation dimensions of part-3, together with L_2, encompass L_3. The best (second best) is in bold (underline).

Model	Language Model	L_1 (Part-1)		Part-2		L_2		Part-3		L_3	
		$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$
BLIP-2 [9]	Flan-T5-XL	39.7	15	31.3	20	37.8	14	-	-	-	-
InstructBLIP [13]	Flan-T5-XL	40.9	13	32.0	12	38.9	12	-	-	-	-
InstructBLIP Vicuna [13]	Vicuna-7B	41.1	12	30.5	25	38.7	13	-	-	-	-
LLaVA [11]	LLaMA-7B	38.2	19	31.8	13	36.7	19	-	-	-	-
MiniGPT-4 [10]	Vicuna-7B	38.7	18	31.0	21	37.0	18	-	-	-	-
VPGTrans [78]	LLaMA-7B	35.9	27	29.1	28	34.3	25	-	-	-	-
MultiModal-GPT [15]	Vicuna-7B	37.4	22	31.4	18	36.0	20	-	-	-	-
Otter [14]	LLaMA-7B	36.5	25	32.4	10	35.6	22	-	-	-	-
OpenFlamingo [79]	LLaMA-7B	37.3	23	31.6	16	36.0	21	-	-	-	-
LLaMA-Adapter V2 [80]	LLaMA-7B	37.5	21	-	-	-	-	-	-	-	-
GVT [50]	Vicuna-7B	34.3	29	34.2	6	34.3	26	-	-	-	-
mPLUG-Owl [12]	LLaMA-7B	39.2	17	31.8	14	37.5	16	-	-	-	-
Qwen-VL [21]	Qwen-7B	43.0	8	32.6	9	40.6	7	-	-	-	-
Qwen-VL-Chat [21]	Qwen-7B	43.1	7	30.7	23	40.3	9	-	-	-	-
LLaVA-1.5 [22]	Vicuna-7B	46.7	3	31.7	15	43.3	3	-	-	-	-
IDEFICS-9B-Instruct [23]	LLaMA-7B	38.2	20	32.8	7	37.0	17	-	-	-	-
InternLM-Xcomposer-VL [24]	InternLM-7B	<u>57.0</u>	2	34.4	4	<u>51.9</u>	2	-	-	-	-
VideoChat [18]	Vicuna-7B	36.8	24	31.3	19	35.6	23	-	-	-	-
Video-ChatGPT [19]	LLaMA-7B	36.4	26	31.6	17	35.3	24	-	-	-	-
Valley [20]	LLaMA-13B	34.3	28	30.5	24	33.5	27	-	-	-	-
CogVLM [81]	Vicuna-7B	42.1	9	32.7	8	40.0	10	-	-	-	-
InternLM-Xcomposer-VL2 [82]	InternLM2-7B	-	-	-	-	-	-	-	-	-	-
InternLM-Xcomposer-VL2-4bit [82]	InternLM2-7B	44.9	4	30.5	26	41.7	4	-	-	-	-
LLaVA-Next [83]	Vicuna-7B	39.8	16	-	-	-	-	-	-	-	-
Yi-VL [84]	Yi-6B	41.8	11	32.4	11	39.6	11	-	-	-	-
SPHINX-v2-1k [85]	LLaMA2-13B	-	-	-	-	-	-	-	-	-	-
mPLUG-Owl2 [86]	LLaMA2-7B	39.7	14	30.9	22	37.7	15	-	-	-	-
GPT-4V [33]	-	**69.2**	1	**44.2**	1	**67.1**	1	-	-	-	-
Gemini-Pro-Vision [87]	-	-	-	-	-	-	-	-	-	-	-
Claude-3-Opus [88]	-	43.5	5	34.3	5	41.5	6	-	-	-	-
Emu [27]	LLaMA-13B	42.0	10	34.6	3	40.4	8	<u>41.4</u>	2	<u>42.3</u>	2
NExt-GPT [31]	Vicuna-7B	29.9	30	30.2	27	30.0	28	33.9	3	31.4	3
SEED-LLaMA [30]	LLaMA2-Chat-13B	43.4	6	<u>35.6</u>	2	41.6	5	**52.3**	1	**44.8**	1

3.4.2 Tailoring existing datasets.

For existing datasets with annotated label, we first prompt ChatGPT/GPT-4 to generate questions based on provided information. We then construct distracting choices either from the annotated labels of other samples or by utilizing ChatGPT to generate three distractors. For distractors generated by ChatGPT, we additionally utilize human annotators to filter out options that are too similar to the groundtruth answer.

3.4.3 Human creation combined with GPT.

For evaluation dimensions lacking suitable data, *e.g. Interleaved Image-Text Analysis* and *Text-Image Creation*, we employ human annotators to meticulously design questions, retrieve corresponding images, and construct distracting choices with the assistance of ChatGPT.

Table 3 Evaluation results of various MLLMs in different subpart levels of SEED-Bench-H. $\bar{T}$ denotes the averaged accuracy across corresponding dimensions, and $R_{\bar{T}}$ denotes the rank based on the the averaged accuracy. The evaluation dimensions of part-2, together with L_1, encompass L_2, while the evaluation dimensions of part-3, together with L_2, encompass L_3. 'Inter. Image & Text' denotes 'Interleaved image-text' and 'Image Gen.' denotes 'Image Generation'. The best (second best) is in bold (underline).

Model	Language Model	Single Image		Multi Images		Videos		Inter. Image & Text		Few-shot Image		Image Gen.		Image & Text Gen.	
		$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$	$\bar{T}$	$R_{\bar{T}}$
BLIP-2 [9]	Flan-T5-XL	42.1	17	28.2	30	34.8	10	41.0	16	24.1	23	-	-	-	-
InstructBLIP [13]	Flan-T5-XL	43.4	14	29.0	29	35.7	7	42.7	12	23.9	26	-	-	-	-
InstructBLIP Vicuna [13]	Vicuna-7B	42.8	15	46.0	10	31.5	18	39.0	23	24.2	20	-	-	-	-
LLaVA [11]	LLaMA-7B	39.8	22	38.5	20	31.0	22	37.6	27	27.5	2	-	-	-	-
MiniGPT-4 [10]	Vicuna-7B	40.5	21	32.9	25	33.4	12	39.6	22	24.6	12	-	-	-	-
VPGTrans [78]	LLaMA-7B	36.8	29	34.3	23	32.6	16	34.5	29	25.0	9	-	-	-	-
MultiModal-GPT [15]	Vicuna-7B	37.5	27	48.3	7	31.5	17	41.0	15	24.3	18	-	-	-	-
Otter [14]	LLaMA-7B	37.6	25	37.0	21	31.4	19	43.6	9	24.1	22	-	-	-	-
OpenFlamingo [79]	LLaMA-7B	37.5	26	47.4	8	31.3	21	41.1	14	24.5	14	-	-	-	-
LLaMA-Adapter V2 [80]	LLaMA-7B	38.9	23	40.7	18	29.8	27	-	-	-	-	-	-	-	-
GVT [50]	Vicuna-7B	33.4	32	50.4	6	30.1	25	43.1	10	27.5	2	-	-	-	-
mPLUG-Owl [12]	LLaMA-7B	40.8	19	44.2	12	29.3	28	39.7	21	25.8	5	-	-	-	-
Qwen-VL [21]	Qwen-7B	45.6	9	39.3	19	32.8	15	42.1	13	25.5	6	-	-	-	-
Qwen-VL-Chat [21]	Qwen-7B	45.7	8	40.8	17	32.8	14	40.2	19	23.6	28	-	-	-	-
LLaVA-1.5 [22]	Vicuna-7B	49.5	6	43.0	14	35.7	6	40.4	18	25.2	8	-	-	-	-
IDEFICS-9B-Instruct [23]	LLaMA-7B	38.1	24	52.5	4	31.3	20	44.3	6	24.2	19	-	-	-	-
InternLM-Xcomposer-VL [24]	InternLM-7B	60.6	4	52.2	5	42.8	2	47.2	5	24.8	10	-	-	-	-
VideoChat [18]	Vicuna-7B	37.1	28	41.0	16	33.4	13	41.0	16	24.1	21	-	-	-	-
Video-ChatGPT [19]	LLaMA-7B	35.8	30	53.8	3	30.2	24	39.0	24	26.1	4	-	-	-	-
Valley [20]	LLaMA-13B	34.6	31	44.7	11	28.0	29	37.7	26	25.2	7	-	-	-	-
CogVLM [81]	Vicuna-7B	43.7	13	41.3	15	35.4	8	44.1	7	24.1	23	-	-	-	-
InternLM-Xcomposer-VL2 [82]	InternLM2-7B	63.9	2	-	-	-	-	-	-	-	-	-	-	-	-
InternLM-Xcomposer-VL2-4bit [82]	InternLM2-7B	49.5	7	32.4	27	30.7	23	38.5	25	24.5	15	-	-	-	-
LLaVA-Next [83]	Vicuna-7B	42.5	16	47.1	9	22.9	30	43.9	8	-	-	-	-	-	-
Yi-VL [84]	Yi-6B	44.2	12	32.8	26	35.2	9	43.0	11	24.4	17	-	-	-	-
SPHINX-v2-1k [85]	LLaMA2-13B	54.2	5	-	-	-	-	-	-	-	-	-	-	-	-
mPLUG-Owl2 [86]	LLaMA2-7B	40.6	20	43.4	13	34.1	11	40.1	20	24.1	23	-	-	-	-
GPT-4V [33]	-	**69.5**	1	**78.6**	1	**61.3**	1	**58.3**	1	**27.8**	1	-	-	-	-
Gemini-Pro-Vision [87]	-	62.5	3	-	-	-	-	-	-	-	-	-	-	-	-
Claude-3-Opus [88]	-	41.8	18	66.6	2	39.8	3	47.4	4	24.6	13	-	-	-	-
Emu [27]	LLaMA-13B	44.3	11	33.2	24	36.1	5	49.0	3	23.8	27	45.0	2	34.2	3
NExt-GPT [31]	Vicuna-7B	29.8	33	31.6	28	29.9	26	37.5	28	24.7	11	32.4	3	36.7	2
SEED-LLaMA [30]	LLaMA2-Chat-13B	45.5	10	35.5	22	37.6	4	50.6	2	24.4	16	**45.5**	1	**65.8**	1

3.5 Evaluation Strategy

3.5.1 Evaluation of text output.

Different from MMBench [38] that employs ChatGPT to match a model's prediction to one of the choices in a multiple-choice question (achieves only 87.0% alignment rate), we adopt the answer ranking strategy [13, 89, 90] for evaluating 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 rely on the instruction-following capabilities of models to output "A" or "B" or "C" or "D". Furthermore, this evaluation strategy eliminates the impact of the order of multiple-choice options on the model's performance.

3.5.2 Evaluation of image output.

Since not all MLLMs with image generation capabilities employ visual autoregression, adopting an answer ranking strategy for image evaluation is impractical. Instead, we calculate the CLIP similarity score [91] between the generated image and each candidate image option, selecting the highest-scoring option as the final prediction of the given multiple-choice question.

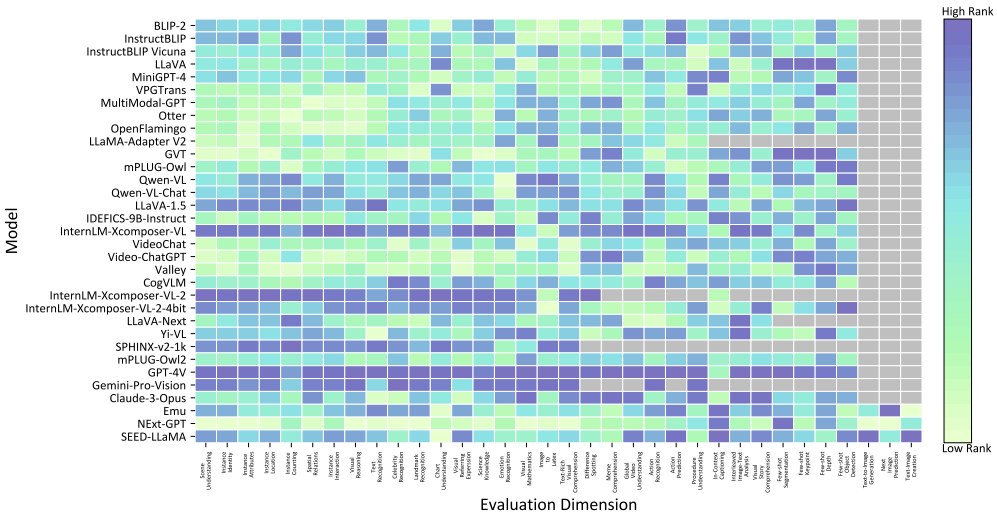

Fig. 10 Illustration of each model's performance across different evaluation dimensions, where darker colors represent higher ranks. Gray indicates that the model has not yet reached the capability level required for evaluating that dimension or don't obtain valid data in that dimension.

3.5.3 Evaluation of text and image output.

We first employ an answer ranking strategy to select the most likely text prediction. If it matches the ground truth, we evaluate the image output using the CLIP similarity score [91] between the generated image and each candidate. The model is deemed correct only if both text and image predictions match the ground truth.

4 Results

4.1 Models

We evaluate a total of 30 open-source MLLMs including BLIP-2 [9], InstructBLIP [13], InstructBLIP Vicuna [13], LLaVA [11], MiniGPT-4 [10], VPGTrans [78], MultiModal-GPT [15], Otter [14], OpenFlamingo [79], LLaMA-Adapter V2 [80], GVT [50], mPLUG-Owl [12], Qwen-VL [21], Qwen-VL-Chat [21], LLaVA1.5 [22], IDEFICS-9B-Instruct [23], InternLM-Xcomposer-VL [24], VideoChat [18], Video-ChatGPT [19], Valley [20], CogVLM [81], InternLM-Xcomposer-VL2 [82], InternLM-Xcomposer-VL2-4bit [82], LLaVA-Next [83], Yi-VL [84], SPHINX-v2-1k [85], mplug-Owl2 [86], Emu [27], NExt-GPT [31], and SEED-LLaMA [30] based on their official implementations, and 3 closed-souce MLLMs including GPT-4V [33], Gemini [87], and Claude-3-Opus [88] based on their official implementations. For each model, we first determine its capability level and then evaluate the corresponding dimensions. It is important to note that we have confirmed with the authors that LLaMA-Adapter V2 [80] is unable to handle interleaved image-text data and few-shot data and SPHINX-v2-1k [85] only handle single image data. For InternLM-Xcomposer-VL2 [82], we employ an A100 (80GB) for evaluation; however, it runs out of memory

Table 4 Evaluation results of various MLLMs in 'Single-Image & Text Comprehension' part of SEED-Bench-H. The best (second best) is in bold (underline).

Model	Language Model	Scene Understanding	Instance Identity	Instance Attribute	Instance Location	Instance Counting	Spatial Relation	Instance Interaction	Visual Reasoning	Text Recognition	Celebrity Recognition	Landmark Recognition	Chart Understanding	Visual Referring Expression	Science Knowledge	Emotion Recognition	Visual Mathematics	Image to Latex	Text-Rich Visual Comprehension
BLIP-2 [9]	Flan-T5-XL	58.5	48.6	49.0	39.1	43.4	36.2	48.5	52.9	60.7	51.8	51.4	19.2	43.2	52.4	29.3	22.0	21.5	30.1
InstructBLIP [13]	Flan-T5-XL	58.9	49.7	61.7	35.1	58.1	34.9	47.4	55.9	61.4	48.5	45.4	26.4	41.7	47.7	34.5	21.2	23.3	29.3
InstructBLIP Vicuna [13]	Vicuna-7B	53.6	43.9	49.0	37.8	56.5	35.8	43.3	56.2	57.2	60.3	44.4	27.9	39.2	39.4	23.0	26.5	44.8	31.0
LLaVA [11]	LLaMA-7B	53.8	47.5	38.3	34.2	42.0	34.7	40.2	52.9	46.4	51.8	45.6	30.3	40.2	37.6	34.3	20.5	35.5	30.0
MiniGPT-4 [10]	Vicuna-7B	56.3	49.2	45.8	37.9	45.3	32.6	47.4	57.1	41.8	52.5	45.2	20.2	41.2	43.3	24.2	25.0	31.4	30.3
VPGTrans [78]	LLaMA-7B	46.9	38.6	33.6	35.6	27.5	34.4	33.0	50.8	47.6	52.4	38.2	30.1	34.7	36.1	31.5	27.3	33.7	30.3
MultiModal-GPT [15]	Vicuna-7B	46.9	42.5	32.0	32.3	27.7	29.7	29.9	48.3	35.2	60.9	50.4	24.2	42.2	37.6	32.1	27.3	43.6	31.7
Otter [14]	LLaMA-7B	45.9	39.7	31.9	31.6	26.4	32.0	33.0	49.2	39.3	59.7	53.0	23.6	41.2	36.1	37.3	22.0	43.6	31.2
OpenFlamingo [79]	LLaMA-7B	46.7	42.3	31.7	33.4	27.4	29.8	29.9	47.7	35.6	60.3	49.8	24.2	42.2	39.0	32.1	27.3	43.6	31.7
LLaMA-Adapter V2 [80]	LLaMA-7B	45.2	38.5	29.3	33.0	29.7	35.5	39.2	52.0	48.7	58.5	46.4	24.2	41.2	40.1	39.7	23.5	45.4	30.6
GVT [90]	Vicuna-7B	41.7	35.5	31.8	29.5	36.2	32.0	32.0	51.1	35.2	39.4	36.4	25.0	36.2	31.1	20.6	22.7	35.5	30.0
mPLUG-Owl [12]	LLaMA-7B	49.7	45.3	32.5	36.7	27.3	32.7	44.3	54.7	49.2	70.9	49.6	23.2	44.2	44.0	32.5	23.5	41.9	31.8
Qwen-VL [21]	Qwen-7B	55.3	44.9	56.2	45.7	58.5	35.2	54.6	53.8	47.6	61.8	56.8	28.1	42.2	44.0	18.4	31.8	49.4	37.0
Qwen-VL-Chat [21]	Qwen-7B	56.5	47.6	54.8	46.9	54.2	40.3	55.7	55.0	47.4	62.4	55.6	25.2	43.7	41.2	20.6	28.8	44.8	42.2
LLaVA-1.5 [22]	Vicuna-7B	63.7	62.4	66.7	51.3	60.2	38.5	47.4	59.8	69.0	60.6	49.8	25.0	45.7	56.7	31.1	24.2	42.4	36.6
IDEFICS-9B-Instruct [23]	LLaMA-7B	48.2	38.2	37.8	32.9	29.0	32.4	37.1	54.1	45.7	52.4	52.8	22.6	42.7	33.2	26.6	21.2	47.7	32.0
InternLM-Xcomposer-VL [24]	InternLM-7B	74.8	70.5	67.6	60.5	55.3	53.4	**76.3**	76.1	61.4	86.1	78.0	27.2	60.3	**84.8**	68.8	25.8	24.4	40.4
VideoChat [18]	Vicuna-7B	44.3	40.7	32.2	36.9	32.9	32.6	42.3	51.1	45.8	35.2	46.8	20.6	43.2	39.4	34.3	19.7	41.3	28.6
Video-ChatGPT [19]	LLaMA-7B	44.1	37.0	35.8	30.7	44.2	31.1	29.9	49.9	39.8	49.7	40.6	22.0	33.2	37.2	22.4	25.0	43.0	29.7
Valley [20]	LLaMA-13B	45.3	36.4	33.7	30.6	27.1	31.5	35.1	52.0	35.2	44.9	43.4	23.8	33.2	37.2	26.0	22.7	35.5	29.0
CogVLM [81]	Vicuna-7B	51.7	43.5	38.9	33.8	29.4	33.6	45.4	53.5	51.5	88.2	67.2	24.2	46.7	49.8	25.8	26.5	43.6	33.4
InternLM-Xcomposer-VL2 [82]	InernLM2-7B	**77.5**	73.5	**74.8**	**65.4**	**65.8**	**57.5**	71.1	75.8	61.2	78.5	87.4	**45.1**	61.8	78.4	66.7	30.3	29.1	50.9
InternLM-Xcomposer-VL2-4bit [82]	InernLM2-7B	68.4	58.1	57.3	44.4	38.7	37.6	58.8	69.5	62.3	66.4	66.6	29.7	52.3	59.6	47.1	27.3	8.8	37.7
LLaVA-Next [83]	Vicuna-7B	69.6	66.5	52.9	42.1	61.1	38.4	42.3	51.7	44.6	48.5	41.0	27.2	41.7	47.3	32.7	22.7	38.4	36.5
Yi-VL [84]	Yi-6B	56.9	49.0	51.7	38.2	47.9	39.9	53.6	52.6	27.1	65.5	49.4	24.4	40.2	46.9	44.1	33.3	41.3	34.6
SPHINX-v2-1k [85]	LLaMA2-13B	67.6	60.6	68.7	53.2	**65.8**	45.4	62.9	66.2	68.3	73.9	55.3	30.7	47.7	52.7	25.8	25.8	49.4	47.0
mPLUG-Owl2 [86]	LLaMA2-7B	50.4	42.8	41.3	37.9	38.7	35.2	47.4	51.1	43.5	62.4	43.8	24.4	39.7	43.7	25.6	28.8	40.7	33.3
GPT-4V [33]	-	**77.5**	**73.9**	70.6	61.8	56.8	56.9	74.2	**78.5**	**82.3**	**97.4**	**45.1**	**71.9**	66.1	**71.1**	**43.9**	**77.9**	**53.5**
Gemini-Pro-Vision [87]	-	73.4	70.0	63.6	54.2	47.8	46.6	70.2	77.4	50.6*	**92.7**	86.2*	37.0*	42.9*	77.3*	66.7*	34.2*	62.2	52.8
Claude-3-Opus [88]	-	51.2	42.5	39.6	39.9	35.3	40.9	42.3	58.0	33.3	48.2	55.2	24.8	34.2	45.9	43.1	37.9	36.6	44.1
Emu [27]	LLaMA-13B	59.0	50.0	43.7	37.1	44.3	33.6	49.5	58.3	61.4	68.8	61.6	19.0	45.7	41.5	24.2	26.4	40.7	33.5
NExt-GPT [31]	Vicuna-7B	36.4	35.1	25.6	29.9	36.1	30.9	39.2	41.7	31.0	30.9	27.4	21.2	34.2	31.8	24.4	17.4	16.9	26.3
SEED-LLaMA [30]	LLaMA2-Chat-13B	64.0	55.0	51.3	45.4	43.3	37.9	56.7	59.2	57.0	55.5	52.8	18.8	49.3	44.8	28.8	24.4	41.9	33.8

when processing more than three image. Consequently, we do not provide InternLM-Xcomposer-VL2 [82] performance results for 'Meme Comprehension', 'Global Video Understanding', 'Action Recognition', 'Action Prediction', 'Procedure Understanding', 'Interleaved Image-Text Analysis', 'Visual Story Comprehension', 'Few-shot Segmentation', 'Few-shot Keypoint', 'Few-shot Depth', and 'Few-shot Object Detection'. In the case of LLaVA-Next [83], the prompt in the task for few-shot image understanding is too long, leading to a 'nan' loss output. As a result, we do not include LLaVA-Next [83] performance in few-shot image understanding. Lastly, for Gemini-Pro-Vision [87], we only report task performance when the model responds valid data in the task and if the number of valid data is not over half of the task, we use * to illustrate. Some MLLMs can reach the capability level L_3, but they are not available as open-source.

4.2 Main Results

The evaluation results of various MLLMs in different capability levels of SEED-Bench-H are listed in Tab. 2. GPT-4V outperforms a large number of MLLMs, achieving the best performance based on the averaged accuracy in capability level L_1 and L_2, and SEED-LLaMA ranks top-1 in capability level L_3 with two competitor.

For each subpart performance, we listed each model evaluation result in Tab 3, and for each task performance, we illustrated in Tab 4, Tab 5, and Tab 6. To better showcase the capabilities of models across different evaluation dimensions, we further visualize the ranking of each model within each evaluation dimension in Fig. 10, where darker colors represent higher ranks and grey color indicates that the model has not yet reached the capability level required for evaluating that dimension. The champion GPT-4V achieves competitive results in a large number of evaluation dimensions of capability level L_1 and L_2. Although NExt-GPT reaches the capability level L_3, it performs poorly in multiple evaluation dimensions at levels L_1 and L_2.

Table 5 Evaluation results of various MLLMs in 'Multi-Images & Text Comprehension' part, 'Video & Text Comprehension' part of SEED-Bench-H. The best (second best) is in bold (underline).

Model	Language Model	Multi-Images & Text Comprehension		Video & Text Comprehension			
		Difference Spotting	Meme Comprehension	Global Video Understanding	Action Recognition	Action Prediction	Procedure Understanding
BLIP-2 [9]	Flan-T5-XL	17.8	38.6	42.5	37.7	36.2	22.9
InstructBLIP [13]	Flan-T5-XL	22.8	35.2	41.5	36.1	40.5	24.5
InstructBLIP Vicuna [13]	Vicuna-7B	36.5	55.4	40.4	38.6	31.2	15.6
LLaVA [11]	LLaMA-7B	27.0	50.0	44.1	36.2	25.1	18.6
MiniGPT-4 [10]	Vicuna-7B	19.0	46.7	39.0	38.7	27.4	28.6
VPGTrans [78]	LLaMA-7B	24.6	44.0	37.8	38.2	20.9	33.5
MultiModal-GPT [15]	Vicuna-7B	40.1	56.5	37.6	38.7	25.3	24.4
Otter [14]	LLaMA-7B	27.4	46.7	36.6	37.9	26.0	24.8
OpenFlamingo [79]	LLaMA-7B	39.9	54.9	37.6	38.4	25.2	24.1
LLaMA-Adapter V2 [80]	LLaMA-7B	29.1	52.2	41.9	38.2	18.8	20.3
GVT [50]	Vicuna-7B	41.5	59.2	40.4	29.7	26.3	24.1
mPLUG-Owl [12]	LLaMA-7B	33.5	54.9	42.0	37.8	18.3	19.3
Qwen-VL-Chat [21]	Qwen-7B	34.3	47.2	39.7	42.8	29.6	19.1
LLaVA-1.5 [22]	Vicuna-7B	35.7	50.3	46.1	39.4	29.4	28.1
IDEFICS-9B-Instruct [23]	LLaMA-7B	56.5	48.4	42.7	38.6	23.6	20.5
InternLM-Xcomposer-VL [24]	InternLM-7B	47.7	56.6	58.6	**49.9**	37.6	24.9
VideoChat [18]	Vicuna-7B	30.3	51.6	41.5	34.0	30.6	27.4
Video-ChatGPT [19]	LLaMA-7B	46.1	61.4	42.6	32.2	27.0	19.0
Valley [20]	LLaMA-13B	37.1	52.2	31.5	32.1	21.9	26.5
CogVLM [81]	Vicuna-7B	33.5	49.1	38.9	43.8	32.7	26.3
InternLM-Xcomposer-VL2 [82]	InernLM2-7B	**71.5**	-	-	-	-	-
InternLM-Xcomposer-VL2-4bit [82]	InernLM2-7B	27.7	37.1	35.5	33.0	26.8	27.5
LLaVA-Next [83]	Vicuna-7B	38.9	55.4	17.8	25.1	22.8	26.1
Yi-VL [84]	Yi-6B	22.2	43.4	34.6	32.4	34.6	37.5
SPHINX-v2-1k [85]	LLaMA2-13B	-	-	-	-	-	-
mPLUG-Owl2 [86]	LLaMA2-7B	36.5	50.3	42.4	36.9	30.4	26.9
GPT-4V [33]	-	67.9	**89.3**	**64.5**	**65.7**	**51.7**	**63.7**
Gemini-Pro-Vision [87]	-	-	-	-	50.7	-	41.9
Claude-3-Opus [88]	-	62.1	71.1	48.8	37.1	32.1	41.3
Emu [27]	LLaMA-13B	29.3	37.1	41.9	42.7	37.9	21.8
NExt-GPT [31]	Vicuna-7B	24.2	39.0	35.5	33.8	25.6	24.5
SEED-LLaMA [30]	LLaMA2-Chat-13B	29.5	41.5	46.7	39.4	43.9	20.3

Table 6 Evaluation results of various MLLMs in 'Interleaved Image & Text Comprehension' part, 'Image Generation' part, 'Image & Text Generation' part of SEED-Bench-H. The best (second best) is in bold (underline).

Model	Language Model	part 2							part 3		
		Interleaved Image & Text Comprehension			Few-shot Image & Understanding				Image Generation		Image & Text Generation
		In-Context Captioning	Interleaved Image-Text Analysis	Visual Story Comprehension	Few-shot Segmentation	Few-shot Keypoint	Few-shot Depth	Few-shot Object Detection	Text-to-Image Generation	Next Image Prediction	Text-Image Creation
BLIP-2 [9]	Flan-T5-XL	40.0	30.6	52.3	24.6	23.0	25.0	23.6	-	-	-
InstructBLIP [13]	Flan-T5-XL	36.7	34.7	56.8	24.6	22.4	25.0	23.6	-	-	-
InstructBLIP Vicuna [13]	Vicuna-7B	26.7	32.7	57.7	24.4	24.6	22.9	24.8	-	-	-
LLaVA [11]	LLaMA-7B	40.0	20.4	52.3	27.6	**28.4**	**29.2**	24.8	-	-	-
MiniGPT-4 [10]	Vicuna-7B	45.8	22.5	50.5	25.4	24.8	22.9	25.4	-	-	-
VPGTrans [78]	LLaMA-7B	19.2	28.6	55.9	24.6	23.8	27.1	24.4	-	-	-
MultiModal-GPT [15]	Vicuna-7B	39.2	30.6	53.2	24.4	25.4	22.9	24.4	-	-	-
Otter [14]	LLaMA-7B	42.5	30.6	57.7	24.2	22.8	25.0	24.2	-	-	-
OpenFlamingo [79]	LLaMA-7B	38.3	32.7	52.3	24.8	25.4	22.9	25.0	-	-	-
LLaMA-Adapter V2 [80]	LLaMA-7B	-	-	-	-	-	-	-	-	-	-
GVT [50]	Vicuna-7B	42.5	34.7	52.3	27.6	**28.4**	**29.2**	24.8	-	-	-
mPLUG-Owl [12]	LLaMA-7B	29.2	28.6	61.3	26.4	23.6	27.1	26.2	-	-	-
Qwen-VL [21]	Qwen-7B	45.8	26.5	54.1	27.0	25.8	22.9	26.2	-	-	-
Qwen-VL-Chat [21]	Qwen-7B	42.5	28.6	49.6	25.0	22.8	22.9	23.6	-	-	-
LLaVA-1.5 [22]	Vicuna-7B	39.2	22.5	59.5	24.0	25.2	25.0	26.4	-	-	-
IDEFICS-9B-Instruct [23]	LLaMA-7B	45.8	34.7	52.3	25.2	23.0	25.0	23.6	-	-	-
InternLM-Xcomposer-VL [24]	InternLM-7B	27.5	36.7	77.5	24.8	26.6	22.9	24.8	-	-	-
VideoChat [18]	Vicuna-7B	40.0	30.6	52.3	25.2	22.4	25.0	23.8	-	-	-
Video-ChatGPT [19]	LLaMA-7B	37.5	24.5	55.0	27.4	26.8	25.0	25.0	-	-	-
Valley [20]	LLaMA-13B	35.8	28.6	48.7	22.0	26.4	27.1	25.2	-	-	-
CogVLM [81]	Vicuna-7B	45.0	30.6	56.8	24.6	23.2	25.0	23.4	-	-	-
InternLM-Xcomposer-VL2 [82]	InernLM2-7B	22.5	-	-	-	-	-	-	-	-	-
InternLM-Xcomposer-VL2-4bit [82]	InernLM2-7B	16.7	28.6	70.3	20.2	23.8	25.0	**29.0**	-	-	-
LLaVA-Next [83]	Vicuna-7B	40.0	36.7	55.0	-	-	-	-	-	-	-
Yi-VL [84]	Yi-6B	40.0	36.7	52.3	24.0	21.8	27.1	24.6	-	-	-
SPHINX-v2-1k [85]	LLaMA2-13B	-	-	-	-	-	-	-	-	-	-
mPLUG-Owl2 [86]	LLaMA2-7B	36.7	28.6	55.0	24.6	23.0	25.0	23.6	-	-	-
GPT-4V [33]	-	29.2	**59.2**	**86.5**	**31.6**	27.2	27.1	25.4	-	-	-
Gemini-Pro-Vision [87]	-	-	-	-	-	-	-	-	-	-	-
Claude-3-Opus [88]	-	20.8	36.7	84.7	25.0	23.2	25.0	25.0	-	-	-
Emu [27]	LLaMA-13B	51.7	30.6	64.9	25.2	23.2	25.0	21.8	46.8	**43.2**	34.2
NExt-GPT [31]	Vicuna-7B	46.7	24.5	41.4	27.6	22.2	25.0	23.8	45.1	19.8	36.7
SEED-LLaMA [30]	LLaMA2-Chat-13B	**54.2**	32.7	64.9	25.2	24.0	22.9	25.4	**50.2**	40.7	**65.8**

4.3 Observations

Through the comprehension and objective evaluation of various MLLMs in different capability levels of SEED-Bench-H, we have uncovered insights that can inform future work.

Existing MLLMs have yet to reach the ceiling level of capability L_1. Even the top-ranked MLLM achieves only a 60% averaged accuracy in capability L_1, which evaluates the comprehension of multimodal inputs in a fixed format, *i.e.*, images or multiple images (videos) and then texts.

The comprehension of Interleaved Image-Text data is more difficult. The majority of MLLMs achieve worse results on part 2, which consists of multiple-choice questions with interleaved image-text inputs, than that on L_1 with fixed-form image and text as inputs.

Only a small number of MLLMs can reach the capability L_3. Only three open-source MLLMs possess the ability to generate images, besides the inherent ability of LLMs to output texts. A universal MLLM that unifies the generation of images and texts is currently underexplored.

It is challenging to address multimodal comprehension and generation simultaneously. Although NExt-GPT reaches the capability level L_3, which can generate both texts and images, it shows poor performance in capability L_1 for multimodal comprehension. Equipping MLLMs with image generation ability without compromising their inherent text output performance remains to be addressed.

All MLLMs struggle with understanding charts and visual mathematics. The top-performing MLLMs achieves only around 30% accuracy, which indicates that the understanding capabilities of MLLMs within specialized domains need enhancement.

MLLMs trained on Interleaved Image-Text data excel in similar-format questions. SEED-LLaMA, Emu, and IDEFICS-9B-Instruct achieve higher accuracy in part 2, which consists of multiple-choice questions with interleaved image-text inputs. These MLLMs are trained on interleaved image-text data besides structured image-caption pairs, which demonstra tes the importance of data for MLLM training.

VideoLLMs fail to achieve competitive performance on temporal understanding. Despite being instruction-tuned on video data, Video-ChatGPT and Valley underperform in temporal understanding compared to MLLMs pre-trained on image data. It indicates that current VideoLLMs have limited capabilities for fine-grained action recognition and temporal reasoning.

Few-shot image understanding remains a significant challenge. Even the top-ranked MLLMs, such as GPT-4V and Claude-3-Opus, struggle to complete few-shot image understanding tasks. The performance of all MLLMs is close to the worst result for a four-choice question (25%), indicating that there is substantial room for improvement in image reasoning for current MLLMs.

Comprehending text-rich data proves to be more complex. The majority of MLLMs achieve lower results on text-rich visual comprehension, with the average accuracy rate being less than 40%. As multimodal agents, specifically website agents,

need to analyze various websites and output correct actions, this suggests that significant advancements are required before MLLMs can be effectively employed as website agents.

5 Conclusion

In this work, we introduce SEED-Bench-H, a large-scale benchmark for evaluating Multimodal Large Language Models (MLLMs) in terms of hierarchical capabilities, including the generation of both texts and images. SEED-Bench-H consists of 28K multiple-choice questions with accurate human annotations, which covers 34 evaluation dimensions. We conduct a thorough evaluation of 30 prominent open-source MLLMs and 3 closed-source MLLMs, analyzing and comparing their performances to provide insights for future research. We have launched and maintain a leaderboard, offering a platform for the community to assess model performance. As of April 6, 2024, over 30 models have submitted their evaluation results to our leaderboard. This clearly demonstrates the positive impact our work has had in fostering growth and progress within the community. By providing a platform for comparison and collaboration, we continue to encourage innovation and drive advancements in the field.

References

[1] Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)

[2] Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench-2: Benchmarking multimodal large language models. arXiv preprint arXiv:2311.17092 (2023)

[3] Li, B., Ge, Y., Chen, Y., Ge, Y., Zhang, R., Shan, Y.: Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. arXiv preprint arXiv:2404.16790 (2024)

[4] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)

[5] OpenAI: GPT-4 Technical Report (2023)

[6] OpenAI: Introducing chatgpt. https://openai.com/blog/chatgpt (2022)

[7] FastChat: Vicuna. https://github.com/lm-sys/FastChat (2023)

[8] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[9] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ICML (2023)

[10] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)

[11] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)

[12] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)

[13] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)

[14] Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)

[15] Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., Chen, K.: MultiModal-GPT: A Vision and Language Model for Dialogue with Humans (2023)

[16] Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355 (2023)

[17] Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)

[18] Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)

[19] Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)

[20] Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M., Lu, P., Wang, T., Wei, Z.: Valley: Video assistant with large language model enhanced ability. arXiv preprint arXiv:2306.07207 (2023)

[21] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)

[22] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)

[23] Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents (2023)

[24] Zhang, P., Wang, X.D.B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Yan, H., et al.: Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112 (2023)

[25] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)

[26] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)

[27] Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222 (2023)

[28] Lili, Y., Bowen, S., Ram, P., Benjamin, M., Olga, G., Tianlu, W., Arun, B., Binh, T., Brian, K., Shelly, S., Candace, R., Adam, P., Russ, H., Vasu, S., Jacob, X., Uriel, S., Daniel, L.A., Gargi, G., Yaniv, T., Maryam, F.-Z., Asli, C., Luke, Z., Armen, A.: Scaling autoregressive multi-modal models: Pretraining and instruction tuning (2023)

[29] Ge, Y., Ge, Y., Zeng, Z., Wang, X., Shan, Y.: Planting a seed of vision in large language model. arXiv preprint arXiv:2307.08041 (2023)

[30] Ge, Y., Zhao, S., Zeng, Z., Ge, Y., Li, C., Wang, X., Shan, Y.: Making llama see and draw with seed tokenizer. arXiv preprint arXiv:2310.01218 (2023)

[31] Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.-S.: Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519 (2023)

[32] Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)

[33] Gpt-4v(ision) system card. (2023). https://api.semanticscholar.org/CorpusID:263218031

[34] Betker, J., Goh, G., Jing, L., TimBrooks, Wang, J., Li, L., LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, Ramesh, A.: Improving image generation with better captions. https://api.semanticscholar.org/CorpusID:264403242

[35] Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., Li, K., Sun, X., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)

[36] Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Sheng, L., Bai, L., Huang, X., Wang, Z., et al.: Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. arXiv preprint arXiv:2306.06687 (2023)

[37] Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:2306.09265 (2023)

[38] Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)

[39] Bai, S., Yang, S., Bai, J., Wang, P., Zhang, X., Lin, J., Wang, X., Zhou, C., Zhou, J.: Touchstone: Evaluating vision-language models by language models. arXiv preprint arXiv:2308.16890 (2023)

[40] Liu, Y., Li, Z., Li, H., Yu, W., Huang, M., Peng, D., Liu, M., Chen, M., Li, C., Jin, L., et al.: On the hidden mystery of ocr in large multimodal models. arXiv preprint arXiv:2305.07895 (2023)

[41] Zhang, W., Aljunied, S.M., Gao, C., Chia, Y.K., Bing, L.: M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. arXiv preprint arXiv:2306.05179 (2023)

[42] Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., Schimdt, L.: Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. arXiv preprint arXiv:2308.06595 (2023)

[43] Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)

[44] Li, S., Tajbakhsh, N.: Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. arXiv preprint arXiv:2308.03349 (2023)

[45] Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., et al.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. arXiv preprint arXiv:2309.14181 (2023)

[46] Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)

[47] Wadhawan, R., Bansal, H., Chang, K.-W., Peng, N.: Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. arXiv preprint arXiv:2401.13311 (2024)

[48] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502 (2023)

[49] Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024)

[50] Wang, G., Ge, Y., Ding, X., Kankanhalli, M., Shan, Y.: What makes for good visual tokenizers for large language models? arXiv preprint arXiv:2305.12223 (2023)

[51] Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)

[52] Huang, X., Zhang, Y., Ma, J., Tian, W., Feng, R., Zhang, Y., Li, Y., Guo, Y., Zhang, L.: Tag2text: Guiding vision-language model via image tagging. arXiv preprint arXiv:2303.05657 (2023)

[53] Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R., Ashida, K., Nagai, H., Okamoto, M., Yamamoto, H., *et al.*: Icdar 2003 robust reading competitions: entries, results, and future directions. International Journal of Document Analysis and Recognition (IJDAR) **7**, 105–122 (2005)

[54] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: Icdar 2013 robust reading competition. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 1484–1493 (2013). IEEE

[55] Mishra, A., Alahari, K., Jawahar, C.: Scene text recognition using higher order language priors. In: BMVC-British Machine Vision Conference (2012). BMVA

[56] Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: 2011 International Conference on Computer Vision, pp. 1457–1464 (2011). IEEE

[57] Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2575–2584 (2020)

[58] Methani, N., Ganguly, P., Khapra, M.M., Kumar, P.: Plotqa: Reasoning over scientific plots. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1527–1536 (2020)

[59] Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6720–6731 (2019)

[60] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)

[61] Dumitru, Goodfellow, I., Cukierski, W., Bengio, Y.: Challenges in Representation Learning: Facial Expression Recognition Challenge. https://kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge

[62] Deng, Y., Kanervisto, A., Ling, J., Rush, A.M.: Image-to-markup generation with coarse-to-fine attention. In: International Conference on Machine Learning, pp. 980–989 (2017). PMLR

[63] Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., Liu, Z.: Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425 (2023)

[64] Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 510–526 (2016). Springer

[65] Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., Wang, L.: Grit: A generative region-to-text transformer for object understanding. arXiv preprint arXiv:2212.00280 (2022)

[66] Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., *et al.*: The" something something" video database for learning and evaluating visual common sense. In: ICCV (2017)

[67] Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision. arXiv preprint arXiv:2006.13256 (2020)

[68] Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: CVPR (2014)

[69] Huang, T.-H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., *et al.*: Visual storytelling. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1233–1239 (2016)

[70] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)

[71] Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36, pp. 31–42 (2014). Springer

[72] Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv preprint arXiv:2212.05032 (2022)

[73] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

[74] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)

[75] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)

[76] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: CVPR (2021)

[77] https://github.com/PaddlePaddle/PaddleOCR: PaddleOCR

[78] Zhang, A., Fei, H., Yao, Y., Ji, W., Li, L., Liu, Z., Chua, T.-S.: Transfer visual prompt generator across llms **abs/23045.01278** (2023)

[79] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)

[80] Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P.,

He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)

[81] Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023)

[82] Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., et al.: Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420 (2024)

[83] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (2024)

[84] Hugging Face: Yi-VL-6B. https://huggingface.co/01-ai/Yi-VL-6B (2024)

[85] Gao, P., Zhang, R., Liu, C., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., et al.: Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. arXiv preprint arXiv:2402.05935 (2024)

[86] Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv preprint arXiv:2311.04257 (2023)

[87] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

[88] Anthropic News: Claude 3 Family. https://www.anthropic.com/news/claude-3-family

[89] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

[90] Lin, S., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958 (2021)

[91] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR