SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension

Bohao Li[1⋆] **Rui Wang**[1⋆] **Guangzhi Wang**[2⋆] **Yuying Ge**[1†] **Yixiao Ge**[1,2†] **Ying Shan**[1,2]

[1]Tencent AI Lab [2]ARC Lab, Tencent PCG

https://github.com/AILab-CVC/SEED-Bench

Abstract

Based on powerful Large Language Models (LLMs), recent generative Multimodal Large Language Models (MLLMs) have gained prominence as a pivotal research area, exhibiting remarkable capability for both comprehension and generation. In this work, we address the evaluation of generative comprehension in MLLMs as a preliminary step towards a comprehensive assessment of generative models, by introducing a benchmark named SEED-Bench. SEED-Bench consists of 19K multiple choice questions with accurate human annotations ($\times$6 larger than existing benchmarks), which spans 12 evaluation dimensions including the comprehension of both the image and video modality. We develop an advanced pipeline for generating multiple-choice questions that target specific evaluation dimensions, integrating both automatic filtering and manual verification processes. Multiple-choice questions with groundtruth options derived from human annotation enables an objective and efficient assessment of model performance, eliminating the need for human or GPT intervention during evaluation. We further evaluate the performance of 18 models across all 12 dimensions, covering both the spatial and temporal understanding. By revealing the limitations of existing MLLMs through evaluation results, we aim for SEED-Bench to provide insights for motivating future research. We will launch and consistently maintain a leaderboard to provide a platform for the community to assess and investigate model capability.

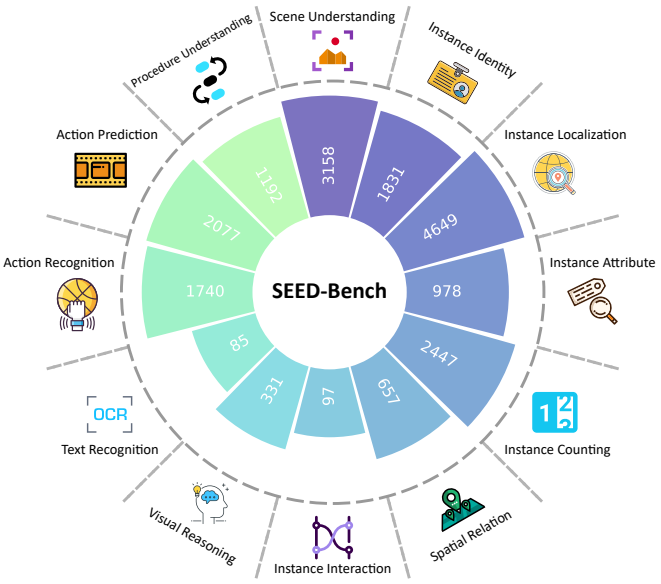
1 Introduction

In recent years, Large Language Models (LLMs) [1, 2, 3, 4, 5] have exhibited remarkable capabilities to understand, reason, and generate texts across a variety of open-ended tasks. Leveraging the strong generality of LLMs, generative Multimodal Large Language Models (MLLMs) [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21] have demonstrate enhanced abilities for multimodal comprehension and generation. However, current MLLMs mainly evaluate their performance with a limited number of qualitative examples, or by employing previous benchmarks that are not tailored for evaluating MLLMs with open-form output. For example, in VQAv2 [22], an answer is considered correct only if the model's output exactly matches the groundtruth answer, which typically consists of just one or two words. The lack of a comprehensive and objective benchmark to evaluate MLLMs poses a significant challenge for comparing and investigating the performance of various models.

Concurrent works [23, 24, 25, 26] have made efforts to develop benchmarks for specifically evaluating MLLMs as shown in Table 1. For example, LVLM-eHub [25] and LAMM [24] utilize exiting public datasets across various computer vision tasks as evaluation samples, and employ human annotators or GPT to assess the quality, relevance, and usefulness of model's predictions. However, the involvement

⋆Equal Contribution.

†Correspondence to yuyingge@tencent.com and yixiaoge@tencent.com.



Rank	Model	Accuracy(%)
1	InstructBLIP Vicuna	53.37
2	InstructBLIP	52.73
3	BLIP2	46.35
4	MiniGPT-4	42.84
5	VPGTrans	39.10
6	VideoChat	37.63
7	mPLUG-Owl	34.01
8	Otter	33.91
9	LLaVa	33.52
10	GVT	33.48
11	MultiModal-GPT	33.15
12	OpenFlamingo	33.14
13	LLaMA-Adapter V2	32.73
14	Video-ChatGPT	31.17
15	Valley	30.32
16	Vicuna	28.50
17	Flan-T5	27.65
18	LLaMA	26.75

: LLM : ImageLLM :VideoLLM

Figure 1: Left: Overview of 12 evaluation dimensions in SEED-Bench including both the spatial and temporal understanding, where the number in the bar denotes the number of human-annotated multiple-choice questions in each dimension. Right: the overall leaderboard displaying the averaged accuracy of 18 models across 12 evaluation dimensions.

of human and GPT during evaluation not only compromises efficiency, but also leads to increased subjectivity and reduced accuracy of the assessment. MME [23] and MMBench [26] further advance objective evaluation of MLLMs by constructing True/False Questions or Multiple-Choice Questions, which cover a variety of ability dimensions. Restricting the model's output to True/False or A/B/C/D options facilitates the convenient computation of accuracy, which serves as an objective metric for evaluation. However, the relatively small scale of these benchmarks (fewer than 3K samples) introduces instability in the evaluation statistics.

In this work, we focus on evaluating the generative comprehension capability of MLLMs as a preliminary step towards a comprehensive assessment of generative models, by introducing a benchmark named SEED-Bench*. SEED-Bench spans 12 evaluation dimensions across both image and video modalities as shown in Fig. 1. SEED-Bench consists of 19K multiple choice questions with groundtruth answers derived from human annotation ($\times$9 larger than MME and $\times$6 larger than MMBench) as shown in Fig. 2. We design a sophisticated pipeline for the generation of multiple-choice questions that are tailored to evaluate specific dimensions. We further incorporate automated filtering mechanism and manual verification process to ensure the quality of questions and the accuracy of groundtruth answers.

Specifically, for images, we utilize various foundation models to extract their visual information including image-level captions [6, 27], instance-level descriptions [28, 29, 30] and textual elements [31]. For videos, we leverage the original human annotations to provide visual information. We then feed the visual information to ChatGPT/GPT-4 with specially designed prompts corresponding to specific evaluation dimension. ChatGPT/GPT-4 subsequently generates questions as well as four candidate options with one groundtruth answer. We further filter out questions that can be answered without the visual input through utilizing multiple LLMs. Finally, we employ human annotators to choose the correct option of each multiple-choice question and classify each question into one evaluation dimension, resulting in a clean and high-quality benchmark containing 19K multiple-choice questions.

*In pursuit of Artificial General Intelligence (AGI), LLMs have witnessed substantial progress. We have made a bold assumption that the premise for the emergence of multimodal capabilities is to unify both comprehension and generation within an autoregressive generative model, where SEED [18] takes a modest step. Besides the exploration of models, it is essential to have appropriate evaluations that motivate research directions. Therefore, we concurrently propose SEED-Bench to evaluate the comprehension ability of generative models.

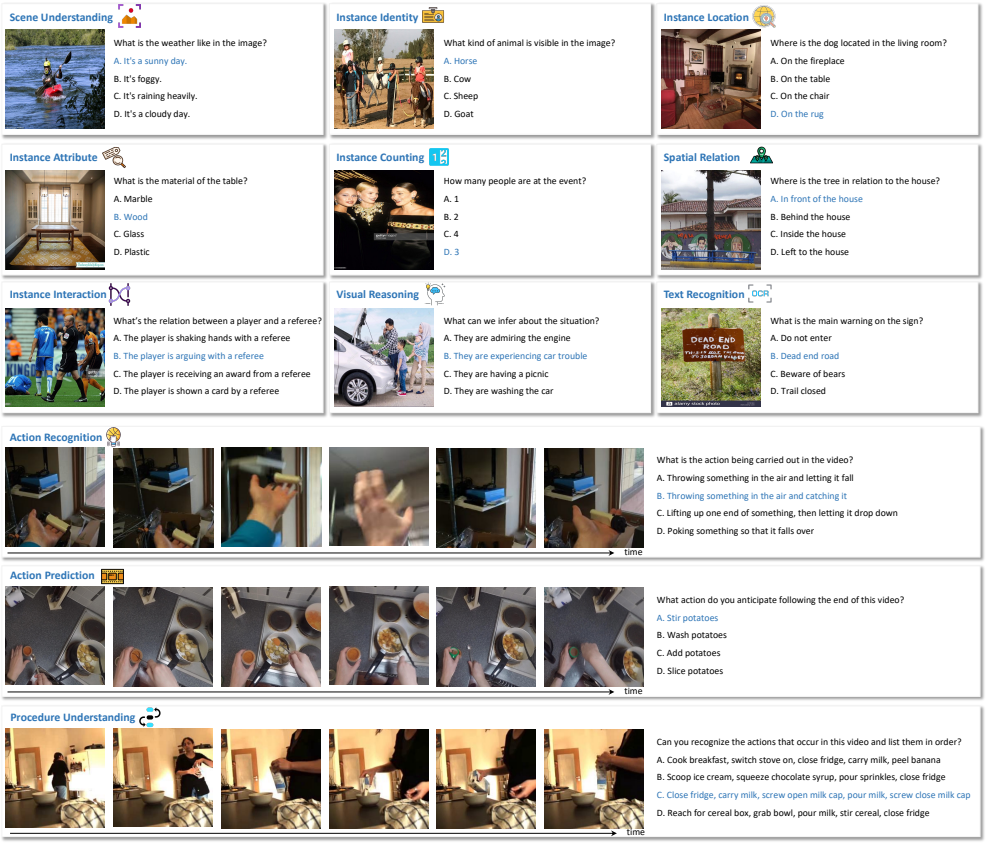


Figure 2: Data samples of SEED-Bench, which covers 12 evaluation dimensions including both the spatial and temporal understanding. Each evaluation dimension contains multiple-choice questions with groundtruth options derived from human annotation.

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

Benchmark	Visual Modality	Customized Question	#Answer Annotation	Answer Type	H/G Evaluation	#Models
MME [23]	Image	✓	2194	Y/N	N/A	10
LAMM [24]	Image & Point cloud	✗	-	free-form	GPT	4
LVLM-eHub [25]	Image	✗	-	free-form	Human	8
MMBench [26]	Image	✓	2974	free-form	GPT	14
Ours	Image & Video	✓	19242	A/B/C/D	N/A	18

Our pipeline supports the scalability of evaluation data across multiple domains, and we will continue to expand the benchmark with more evaluation dimensions.

Based on SEED-Bench, we comprehensively evaluate 18 models including LLMs, ImageLLMs and VideoLLMs across all 12 dimensions as shown in Fig. 1. Different from MMBench [26] 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 follow GPT-3 [32] to calculate log-likelihood for each candidate option and select the one with the highest value as the final prediction, without relying on the instruction-following capabilities of models to output "A" or "B" or "C" or "D". By analyzing the results across 12 dimensions, we conduct a comprehensive comparison of existing multimodal models in both spatial and temporal understanding capabilities. We observe that the majority of MLLMs still exhibit limited performance across all 12 evaluation dimensions, and surprisingly find that VideoLLMs fail to achieve competitive performance on temporal understanding compared with ImageLLMs. Through the evaluation results, we aim for SEED-Bench to provide insights for motivating future exploration of a more advanced MLLM. We will launch an evaluation platform and consistently maintain a leaderboard for assessing and comparing model performance.

2 Related Work

Multimodal Large Language Models. With the impressive success of Large language models (LLM) [1, 5, 4], recent studies work on generative Multimodal Large Language Models (MLLMs) [6, 7, 8, 9, 10, 11, 12, 13, 14, 18, 19, 20, 21] to improve multimodal comprehension and generation through utilizing the strong generality of LLMs. Some work [15, 16, 17] further considers video inputs and leverage the vast capabilities of LLMs for video understanding tasks. In SEED-Bench, we provide a comprehensive quantitative evaluations of these models to thoroughly assess and compare their performance in generative comprehension.

Benchmarks for Multimodal Large Language Models. With the rapid development of Multimodal Large Language Models (MLLMs), some concurrent works [23, 24, 25, 26] propose various benchmarks for evaluating MLLMs. For example, GVT [33] 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). But its evaluation is constrained to limited aspects of visual understanding. LVLM-eHub [25] 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 predictions of models. The involvement of human annotators during evaluation not only introduces bias but also incurs significant costs. LAMM [24] evaluates image and point cloud tasks by using entity extraction to obtain key answers from open-form predictions and utilizing GPT to evaluate the answers' relevance and accuracy to the groundtruth. The reliance on entity extraction and GPT metric can impact the accuracy and reliability of the evaluation. MME [23] and MMBench [26] aim to enhance the objective evaluation of MLLMs by constructing 2914 True/False Questions and 2974 Multiple Choice Questions across a variety of ability dimensions respectively. Considering the relatively small scale of these benchmarks, their evaluation results may exhibit instability. In this work, we introduce SEED-Bench to provide objective and comprehension evaluation of MLLMs, which contains 19K multiple-choice questions covering 12 evaluation dimensions including both spatial and temporal understanding.

3 SEED-Bench

Our benchmark contains 19K multiple-choice questions with accurate human annotations spanning 12 evaluation dimensions including both the spatial and temporal understanding. In this section, we first present the evaluation dimensions of SEED-Bench in Sec. 3.1. We introduce the data source in Sec. 3.2 and our pipeline for constructing multiple-choice questions in Sec. 3.3. We finally describe the evaluation strategy for MLLMs to answer multiple-choice questions in Sec. 3.4.

3.1 Evaluation Dimensions

In order to comprehensively assess the visual understanding capability of MLLMs, SEED-Bench incorporates 12 evaluation dimensions including both the spatial and temporal comprehension as shown in Table 2.

Spatial Understanding. For the evaluation of spatial comprehension, we consider 9 dimensions covering image-level and instance-level perception and reasoning.

- Scene Understanding. This dimension focuses on the global information in the image. Questions can be answered through a holistic understanding of the image.
- Instance Identity. This dimension involves the identification of a certain instance in the image, including the existence or category of a certain object in the image. It evaluates a model's object recognition capability.
- Instance Location. This dimension concerns the absolute position of one specified instance. It requires a model to correctly localize the object referred to in the question.
- Instance Attributes. This dimension is related to the attributes of an instance, such as color, shape or material. It assesses a model's understanding of an object's visual appearance.
- Instances Counting. This dimension requires the model to count the number of a specific object in the image. This requires the model to understand all objects, and successfully count the referred object's instances.

Table 2: Evaluation dimensions of SEED-Bench including both the spatial and temporal understanding. We omit the image in the sample questions.

	Evaluation Dimensions	Sample Questions
Spatial Understanding	1. Scene Understanding	What is the weather like in the image? A. It's a sunny day B. It's foggy C. It's raining heavily D. It's a cloudy day
	2. Instance Identity	What kind of animal is visible in the image? A. Horse B. Cow C. Sheep D. Goat
	3. Instance Location	Where is the dog located in the living room? A. On the fireplace B. On the table C. On the chair D. On the rug
	4. Instance Attribute	What is the material of the table? A. Marble B. Wood C. Glass D. Plastic
	5. Instance Counting	How many people are there in the image? A. 1 B. 2 C. 4 D. 3
	6. Spatial Relation	What is the tree in relateion to the house? A. In front of the house B. Behind the house C. Inside the house D. Left to the house
	7. Instance Interaction	What is the relation between a player and a referee? A. The player is shaking hands with a referee B. The player is arguing with a referee C. The player is receiving an award from a referee D. The player is shown a card by a referee
	8. Visual Reasoning	what can we infer about the situation? A. They are admiring the engine B. They are experiencing car trouble C. They are having a picnic D. They are washing the car
	9. Text Recognition	What is the main warning on the sign? A. Do not enter B. Dead end road C. Beware of bears D. Trail closed
Temporal Understanding	10. Action Recognition	What is the action being carried out in the video? A. Throwing something in the air and letting it fall B. Throwing something in the air and catching it C. Lifting up one end of something, then letting it drop down D. Poking something so that it falls over
	11. Action Prediction	What action do you anticipate following the end of this video? A. Stir potatoes B. Wash potatoes C. Add potatoes D. Slice potatoes
	12. Procedure Understanding	Can you recognize the actions in this video and list them in order? A. Cook breakfast, switch stove on, close fridge, carry milk, peel banana B. Scoop ice cream, squeeze chocolate syrup, pour sprinkles, close fridge C. Close fridge, carry milk, screw open milk cap, pour milk, screw close milk cap D. Reach for cereal box, grab bowl, pour milk, stir cereal, close fridge

- Spatial Relation. This dimension asks an model to ground the two mentioned objects, and recognize their relative spatial relation within the image.
- Instance Interaction. This dimension requires the model to recognize the state relation or interaction relations between two humans or objects.
- Visual Reasoning. This dimension evaluates if a model is able to reason based on the visual information. This requires the model to fully understand the image and utilize its commonsense knowledge to correctly answer the questions.
- Text Understanding. For this dimension, the model should answer question about the textual elements in the image.

Temporal Understanding. For the evaluation of temporal comprehension, we consider 3 dimensions focusing on the recognition, prediction and procedure understanding of actions.

- Action Recognition. In this dimension, the model is required to recognize the action shown in the videos. Not only the ability of capture temporal dynamics, but also the knowledge of physical motions, human actions and dynamic interaction between objects is evaluated.
- Action Prediction. The target of this dimension is to predict the future action through the preceding video segment, which requires the understanding of contextual information from videos and temporal reasoning.
- Procedure Understanding. This dimension requires the model to capture all the key actions and perform temporal ordering on them. We aims to evaluate the ability of temporally fine-grained understanding and procedure reasoning.

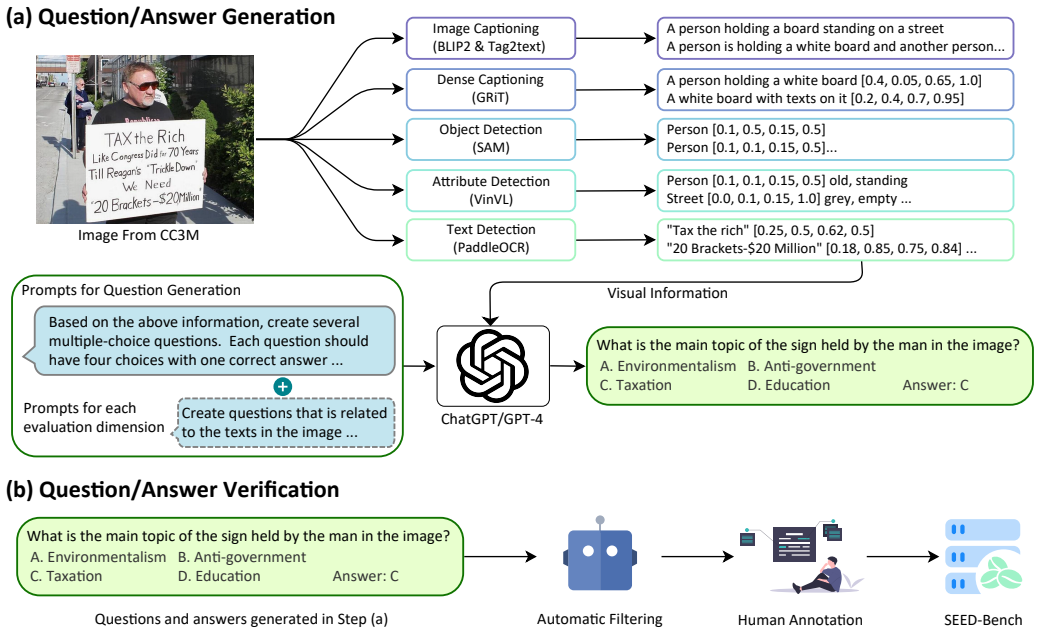


Figure 3: Overview of SEED-Bench pipeline for generating multiple-choice questions of images. (a) We first leverage various foundation models to extract visual information including image-level captions, instance-level descriptions and textual elements. Based on specially designed prompts corresponding 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.

3.2 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, so that we can construct diverse challenging multiple-choice questions. In SEED-Bench, we use CC3M [34] dataset with filtered samples to build questions for spatial understanding. Specifically, considering the noisy original captions of CC3M, we generate captions for each image with Tag2Text [27]. We filter out those images with no more than 5 nouns in their captions, so as to ensure the information richness in the remaining images for constructing questions.

We further adopt Something-Something-v2 (SSV2) [35], Epic-kitchen 100 [36] and Breakfast [37] dataset to build questions for temporal understanding. SSV2 is an action recognition dataset including 174 fine-grained categories of basic actions with everyday objects and we adopt 1740 videos from its validation set. We also select 138 long videos from Epic-kitchen 100 dataset with temporally annotated action labels. Moreover, videos and fine-grained action segmentation annotations in Breakfast dataset [37] are utilized for the procedure understanding task.

3.3 Multiple-Choice Questions

As shown in Fig. 3, our pipeline for generating multiple-choice questions involves question/answer generation and verification. For generating question/answer pairs, 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. For verifying question/answer pairs, we filter out questions that can be answered correctly by multiple LLMs without resorting to visual information. We further employ human annotators to select the correct option and classify each question into one evaluation dimension.

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. For constructing questions related to temporal understanding, considering that extracting reliable temporal information from videos (especially fine-grained actions and long-term temporal context) is extremely difficult given existing foundation models, we utilize the ground-truth annotations of video datasets. We will explore how to generate questions based on automatically extracted video information in the future. 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 [38] and Tag2Text [27] 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 [29] to segment each instance in the image and obtain their bounding boxes according to the segmentation results. The object labels are obtained using Tag2Text [27]. Besides, we also utilize attribute detector [30] to obtain the attributes of each instance in the image. Finally, we employ GRiT [28] 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 [31] for detecting textual elements.

Question-Answer Generation. After extracting visual information from the image and video, 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 [2] 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. The detailed prompts of generating multiple-choice questions for different evaluation dimensions are listed in Fig. 4. For generating questions related to temporal understanding, we utilize the ground-truth annotations of selected videos as the answer of multi-choice questions and employ ChatGPT to generate three distractors.

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 [4], Flan-T5-XXL [1] and LLaMA-7B [5] 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.

Human Annotation. To ensure the accuracy and objectiveness of SEED-Bench, 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. This results in a clean, high-quality and well-categorized benchmark for evaluation with a total of 19K multiple-choice questions. The statistics of the number of multiple-choice questions in each evaluation dimension is shown in Fig. 1. We can observe a minimum number of questions in text recognition with 85 samples, and a maximum number in instance localization with 4649 samples. We will maintain an even distribution among multiple-choice questions associated with different evaluation dimensions in the future.

Default Instruction:

"You are an AI visual assistant that can analyze a single image. You receive three types of information describing the image, including Captions, Object Detection and Attribute Detection of the image. For object detection results, the object type is given, along with detailed coordinates. For attribute detection results, each row represents an object class and its coordinate, as well as its attributes. All coordinates are in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y. Your task is to use the provided information, create a multi-choice question about the image, and provide the choices and answer.

Instead of directly mentioning the bounding box coordinates, utilize this data to explain the scene using natural language. Include details like object counts, position of the objects, relative position between the objects.

When using the information from the caption and coordinates, directly explain the scene, and do not mention that the information source is the caption or the bounding box. Always answer as if you are directly looking at the image.

Create several questions, each with 4 choices. Make the question challenging by not including the visual content details in the question so that the user needs to reason about that first. Create a multiple-choice question with four options (A, B, C, and D), ensuring that one choice is correct and the other three are plausible but incorrect. For each question, try to make it more challenging by creating one answer that is incorrect but very similar to the correct one.

Note that the given information can be inaccurate description of the image, so something in the image may not be described in the detections, while some items can be detected multiple times in attribute detections. Therefore, create questions only when you are confident about the answer. Don't explain your choice."

Scene Understanding Instruction:

"Create complex questions about the major content of the image. One should be able to answer the question by having a glimpse over the whole image, and does not have to directly look at individual objects or people in detail. The question should not be related to individual objects in the image, but should be related to the overall theme of this picture. "

Instance Identity Instruction:

"Create complex questions about the identity of objects appeared in the image, such as its type/class or its existence. For example, you may ask "What an object is?" or "Does some object appear in the image?". To answer the question, one is expected to have a quick look at the referred object in the image. "

Instance Localization Instruction:

"Create complex questions about the location of a certain object in the image. The question should be created based on the coordinates of the objects. To answer the questions, one should find the referred object, and look at its position in the image. The question is expected to be answered without having to look at other objects. "

Instance Attribute Instruction:

"Create complex questions about the attribute of a certain object, such as its color, shape or fine-grained type. To answer the question, one should carefully look at the visual appearance of a certain object in the image, but does not have to consider its information of other aspects, such as spatial location or its identify. "

Instance Counting Instruction:

"Create questions that involve the number of appearance of a certain object. Start with "How many ....". The choices of the question should be numbers. To answer the question, one should find and count all of the mentioned objects in the image. "

Spatial Relation Instruction:

"Create questions about spatial relations between two objects. The questions should be mainly based on the coordinates of the two objects. To answer the questions, one should find the two mentioned objects, and find their relative spatial relation to answer the question. "

Instance Interaction Instruction:

"Create questions about the relations and connections between two objects, such as "What a person is doing to an object" and "What is the relation between two objects". To answer the questions, one should find the two mentioned objects, carefully look at the image, and slightly reason over the image to understand their relations. "

Visual Reasoning Instruction:

"Create complex questions beyond describing the scene. To answer such questions, one should first understanding the visual content, then based on the background knowledge or reasoning, either explain why the things are happening that way, or provide guides and help to user's request. Make the question challenging by not including the visual content details in the question so that the user needs to reason about that first. "

Text Recognition Instruction:

"Create questions that is related to the texts in the image. Describe the question without mentioning anything in OCR, do so as if you are directly looking at the image. "

Figure 4: Prompts of generating multiple-choice questions for different evaluation dimensions.

Table 3: Evaluation results of different models on SEED-Bench, where "Spatial" shows the averaged performance on nine dimensions for evaluating spatial understanding, and "Temporal" shows the averaged performance on three dimensions for evaluating temporal understanding.

Model Type	Model	Language Model	Spatial		Temporal		Overall	
			Acc	Rank	Acc	Rank	Acc	Rank
LLM	Flan-T5 [1]	Flan-T5-XL	27.32	17	28.56	11	27.65	17
	Vicuna [4]	Vicuna-7B	28.16	16	29.46	8	28.50	16
	LLaMA [5]	LLaMA-7B	26.56	18	27.27	13	26.75	18
ImageLLM	BLIP-2 [6]	Flan-T5-XL	49.74	3	36.71	3	46.35	3
	InstructBLIP [10]	Flan-T5-XL	57.80	2	**38.31**	1	52.73	2
	InstructBLIP Vicuna [10]	Vicuna-7B	**58.76**	1	38.05	2	**53.37**	1
	LLaVA [8]	LLaMA-7B	36.96	8	23.75	16	33.52	9
	MiniGPT-4 [7]	Flan-T5-XL	47.40	4	29.89	7	42.84	4
	VPGTrans [40]	LLaMA-7B	41.81	5	31.40	5	39.10	5
	MultiModal-GPT [12]	LLaMA-7B	34.54	12	29.21	10	33.15	11
	Otter [11]	LLaMA-7B	35.16	11	30.35	6	33.91	8
	OpenFlamingo [41]	LLaMA-7B	34.51	13	29.25	9	33.14	12
	LLaMA-Adapter V2 [42]	LLaMA-7B	35.19	10	25.75	14	32.73	13
	GVT [33]	Vicuna-7B	35.49	9	27.77	12	33.48	10
	mPLUG-Owl [9]	LLaMA-7B	37.88	7	23.02	18	34.01	7
VideoLLM	VideoChat [15]	Vicuna-7B	39.02	6	33.68	4	37.63	6
	Video-ChatGPT [16]	LLaMA-7B	33.88	14	23.46	17	31.17	14
	Valley [17]	LLaMA-13B	32.04	15	25.41	15	30.32	15

3.4 Evaluation Strategy

Different from MMBench [26] 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 [10, 32, 39] 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 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.

4 Evaluation Results

4.1 Models

Based on our SEED-Bench, we evaluate 18 models including 3 LLMs, *i.e.*, Flan-T5 [1], Vicuna [4], LLaMA [5], 12 ImageLLMs, *i.e.*, OpenFlamingo [41], BLIP-2 [6], MiniGPT-4 [7], LLaVa [8], mPLUG-Owl [9], InstructBLIP [10], Otter [11], MultimodalGPT [12], GVT [33], PandaGPT [13], VPGTrans [40], LLaMA-Adapter V2 [42], and 3 VideoLLMs, *i.e.*, VideoChat [15], Video-ChatGPT [16] and Valley [17]. Each model is evaluated with all the 12 dimensions including both the spatial and temporal understanding. For ImageLLMs, besides the evaluation of spatial understanding, we aim to investigate their capability to perform temporal reasoning among multiple frames. For VideoLLMs, we seek to explore whether their spatial understanding abilities have degraded by taking a single image as the input.

4.2 Results

The evaluation results of different models on SEED-Bench are listed in Table. 1, where the accuracy refers to the proportion of correctly answered multiple-choice questions relative to the total number of questions. We are surprised to observe that InstructBLIP [10] not only achieves the best performance based on the averaged results across nine dimensions for evaluating spatial understanding, but also surpasses VideoLLMs in terms of the averaged results across three dimensions for evaluating temporal understanding. We display leaderboards of various evaluation dimensions on SEED-Bench in Fig. 5 to provide a comprehensive assessment of different models. The overall leaderboard based on the

Rank	Model	Accuracy(%)
1	InstructBLIP	60.29
2	InstructBLIP Vicuna	60.20
3	BLIP2	59.12
4	MiniGPT-4	56.27
5	VPGTrans	51.87
6	mPLUG-Owl	49.68
7	VideoChat	47.12
8	LLaMA-Adapter V2	45.22
9	Otter	44.90
10	OpenFlamingo	43.86
11	MultiModal-GPT	43.64
12	LLaVa	42.69
13	GVT	41.74
14	Valley	39.33
15	Video-ChatGPT	37.24
16	LLaMA	26.28
17	Vicuna	23.38
18	Flan-T5	23.04

(1) Scene Understanding

Rank	Model	Accuracy(%)
1	InstructBLIP Vicuna	58.93
2	InstructBLIP	58.49
3	BLIP2	53.90
4	MiniGPT-4	49.15
5	mPLUG-Owl	45.33
6	VPGTrans	44.13
7	VideoChat	43.80
8	Otter	38.56
9	LLaMA-Adapter V2	38.50
10	OpenFlamingo	38.12
11	MultiModal-GPT	37.85
12	GVT	35.50
13	LLaVa	34.90
14	Valley	32.88
15	Video-ChatGPT	31.40
16	Vicuna	30.67
17	Flan-T5	29.00
18	LLaMA	27.40

(2) Instance Identity

Rank	Model	Accuracy(%)
1	InstructBLIP Vicuna	65.63
2	InstructBLIP	63.37
3	BLIP2	49.19
4	MiniGPT-4	45.82
5	VPGTrans	39.90
6	VideoChat	34.85
7	LLaVa	33.45
8	Video-ChatGPT	33.23
9	Flan-T5	32.76
10	mPLUG-Owl	32.52
11	Otter	32.24
12	GVT	31.79
13	Valley	31.62
14	MultiModal-GPT	31.45
15	OpenFlamingo	31.28
16	Vicuna	29.69
17	LLaMA-Adapter V2	29.30
18	LLaMA	26.16

(3) Instance Location

Rank	Model	Accuracy(%)
1	InstructBLIP Vicuna	43.56
2	BLIP2	42.33
3	InstructBLIP	40.59
4	VideoChat	39.98
5	MiniGPT-4	37.93
6	mPLUG-Owl	36.71
7	VPGTrans	36.09
8	LLaMA-Adapter V2	33.03
9	Flan-T5	31.75
10	Vicuna	30.91
11	Otter	30.88
12	MultiModal-GPT	30.78
13	OpenFlamingo	30.06
14	GVT	29.45
15	LLaVa	28.43
16	Video-ChatGPT	28.43
17	LLaMA	28.25
18	Valley	27.91

(4) Instance Attributes

Rank	Model	Accuracy(%)
1	InstructBLIP	58.44
2	InstructBLIP Vicuna	57.05
3	MiniGPT-4	45.32
4	BLIP2	43.15
5	LLaVa	41.85
6	GVT	36.17
7	Video-ChatGPT	35.51
8	VPGTrans	33.71
9	VideoChat	32.82
10	Vicuna	30.83
11	LLaMA-Adapter V2	29.67
12	MultiModal-GPT	27.34
13	OpenFlamingo	27.30
14	mPLUG-Owl	27.26
15	Otter	26.28
16	LLaMA	25.07
17	Valley	24.23
18	Flan-T5	20.54

(5) Instance Counting

Rank	Model	Accuracy(%)
1	InstructBLIP Vicuna	40.33
2	InstructBLIP	38.66
3	BLIP2	36.68
4	VPGTrans	36.38
5	LLaMA-Adapter V2	35.46
6	VideoChat	34.55
7	mPLUG-Owl	32.72
8	MiniGPT-4	32.57
9	GVT	31.96
10	Otter	31.81
11	Flan-T5	31.75
12	LLaVa	30.75
13	OpenFlamingo	30.59
14	MultiModal-GPT	30.14
15	Valley	30.14
16	Video-ChatGPT	29.53
17	LLaMA	28.77
18	Vicuna	28.57

(6) Spatial Relations

Rank	Model	Accuracy(%)
1	BLIP2	55.67
2	InstructBLIP Vicuna	52.58
3	InstructBLIP	51.55
4	MiniGPT-4	47.42
5	mPLUG-Owl	44.33
6	VideoChat	42.27
7	LLaMA-Adapter V2	39.18
8	Flan-T5	32.98
9	VPGTrans	31.96
10	GVT	31.96
11	Otter	31.96
12	OpenFlamingo	29.90
13	MultiModal-GPT	29.90
14	Vicuna	29.79
15	LLaVa	27.84
16	Valley	27.84
17	Video-ChatGPT	23.71
18	LLaMA	19.15

(7) Instance Interaction

Rank	Model	Accuracy(%)
1	MiniGPT-4	57.10
2	mPLUG-Owl	54.68
3	VPGTrans	53.17
4	LLaMA-Adapter V2	51.96
5	Otter	51.36
6	MultiModal-GPT	51.36
7	GVT	51.06
8	VideoChat	50.45
9	OpenFlamingo	50.15
10	InstructBLIP Vicuna	47.73
11	LLaVa	46.83
12	InstructBLIP	45.92
13	BLIP2	45.62
14	Valley	43.81
15	Video-ChatGPT	42.30
16	LLaMA	37.01
17	Vicuna	18.51
18	Flan-T5	18.15

(8) Visual Reasoning

Rank	Model	Accuracy(%)
1	InstructBLIP Vicuna	43.53
2	LLaVa	37.65
3	Otter	31.76
4	VPGTrans	30.59
5	GVT	27.06
6	InstructBLIP	25.88
7	BLIP2	25.88
8	Video-ChatGPT	25.88
9	LLaMA-Adapter V2	24.71
10	OpenFlamingo	20.00
11	Flan-T5	19.40
12	mPLUG-Owl	18.82
13	MultiModal-GPT	18.82
14	VideoChat	17.65
15	Vicuna	13.43
16	MiniGPT-4	11.76
17	Valley	11.76
18	LLaMA	8.96

(9) Text Recognition

Rank	Model	Accuracy(%)
1	VPGTrans	39.54
2	LLaMA-Adapter V2	38.56
3	MiniGPT-4	38.22
4	Otter	37.93
5	OpenFlamingo	37.24
6	MultiModal-GPT	36.90
7	VideoChat	34.89
8	InstructBLIP Vicuna	34.48
9	GVT	33.91
10	InstructBLIP	33.10
11	LLaMA	32.99
12	BLIP2	32.59
13	Valley	31.26
14	LLaVa	29.71
15	Video-ChatGPT	27.59
16	Vicuna	27.30
17	mPLUG-Owl	26.72
18	Flan-T5	23.16

(10) Action Recognition

Rank	Model	Accuracy(%)
1	InstructBLIP Vicuna	49.64
2	InstructBLIP	49.11
3	BLIP2	47.47
4	VideoChat	36.35
5	Flan-T5	34.91
6	Vicuna	34.52
7	Otter	27.15
8	MultiModal-GPT	25.76
9	OpenFlamingo	25.42
10	GVT	25.37
11	MiniGPT-4	24.51
12	VPGTrans	24.31
13	Valley	23.21
14	LLaMA	23.11
15	LLaVa	21.43
16	Video-ChatGPT	21.33
17	LLaMA-Adapter V2	18.54
18	mPLUG-Owl	17.91

(11) Action Prediction

Rank	Model	Accuracy(%)
1	VPGTrans	31.88
2	VideoChat	27.27
3	InstructBLIP	27.10
4	MiniGPT-4	27.10
5	mPLUG-Owl	26.51
6	LLaMA	26.17
7	Flan-T5	25.42
8	Otter	24.83
9	OpenFlamingo	24.24
10	BLIP2	23.99
11	MultiModal-GPT	23.99
12	Vicuna	23.83
13	InstructBLIP Vicuna	23.07
14	GVT	22.99
15	Video-ChatGPT	21.14
16	Valley	20.72
17	LLaMA-Adapter V2	19.63
18	LLaVa	19.13

(12) Procedure Understanding

Figure 5: Leaderboards of different evaluation dimensions on SEED-Bench.

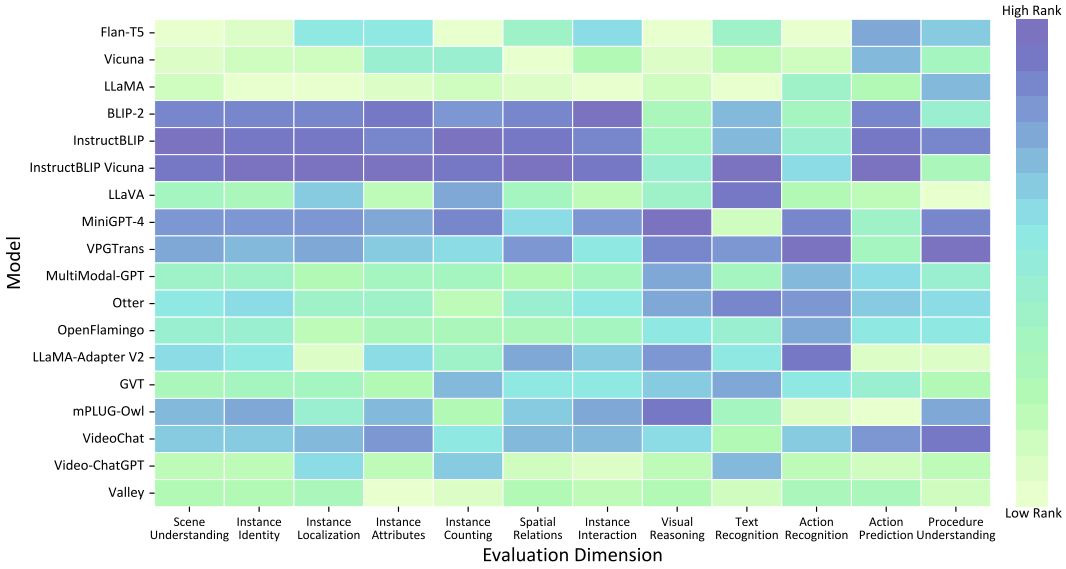


Figure 6: Illustration of each model's performance across different evaluation dimensions, where darker colors represent higher ranks.

averaged results across all the evaluation dimensions are shown in Fig. 1. To better showcase the the capabilities of models across different evaluation dimensions, we further visualize the ranking of each model within each evaluation dimension in Fig. 6, where darker colors represent higher ranks. We can observe that the BLIP series [6, 10] model achieves competitive results in multiple evaluation dimensions, but they are not good at visual reasoning and action recognition. VideoLLM Valley [17] achieves suboptimal performance in the majority of evaluation dimensions. LLaVa [8] exhibits unparalleled capabilities in the evaluation of text recognition compared to other evaluation dimensions. In terms of specific evaluation dimension, MiniGPT-4 [7] model and mPLUG-Owl [9] model performs better in visual reasoning, while VPGTrans [40] model excels in action recognition and procedure understanding. LLaMA Adapter V2 [42] model shows more proficiency in action recognition. What's more, Multimodal GPT [12], Otter [11], Openflamingo [41], GVT [33], and the three VideoLLMs [15, 16, 17] exhibit balanced strength across various evaluation dimensions.

4.3 Analysis

Through the comprehension and objective evaluation of various models on SEED-Bench, we have observed a number of findings that can bring insights for future work.

Most MLLMs still exhibit limited performance across all 12 evaluation dimensions. As shown in Fig. 1, 5, most MLLMs (except BLIP series models) can not reach 50% accuracy on both average performance and the performance on more than three single evaluation dimension. In some specific evaluation dimension (*e.g.*, visual reasoning), it seems that most MLLMs achieve high accuracy. However, when comparing the performance of MLLMs to LLMs, we observe that the performance improvement of most MLLMs is still relatively limited.

MLLMs achieve relatively high performance on global image comprehension On the evaluation of scene understanding and visual reasoning, the accuracy of most MLLMs is higher than 40%, and all MLLMs outperforms LLMs. This shows that MLLMs are more proficient in global understanding and reasoning of images, compared with other evaluation dimensions that require fine-grained instance-level comprehension.

InstructBLIP achieves top performance on 8 of 12 evaluation dimensions. We can observe that InstructBLIP outperforms other models on 8 evaluation dimensions and the possible explanations for this superior performance are as follows. (a) The instruction-tuning data of InstructBLIP contains totally 16M samples (larger than other instruction-tuning datasets), and covers a wide range of multi-modal tasks, even including QA data of OCR and temporal visual reasoning. (b) The weights of LLMs are frozen when performing instruction-tuning of InstructBLIP, which may alleviate catastrophic forgetting. However, InstructBLIP series models still perform poorly on action recognition and

procedure understanding that differ significantly from the instruction-tuning data. For instance, on action recognition that requires the understanding of fine-grained actions in Something-Something-v2, InstructBLIP series models can not achieve significant performance gain compared to LLMs (*i.e.*, lower than 2%). This indicates that InstructBLIP series models may fail to generalize well on the out-of-distribution data.

Despite excelling in instance localization, top MLLMs show weaker abilities in understanding spatial relationships between objects. While achieving 63.37% accuracy on the evaluation of instance location, InstructBLIP fails to reach 40% accuracy on the evaluation of spatial relations. This is because that recognizing relative spatial relationships between instances is generally more challenging than instance localization. Specifically, for each instance, while localization typically involves finding one position, there can be many more possible arrangements and combinations of spatial relationships between instances. Therefore, the recognition of spatial relationship requires handling increased variability. Additionally, spatial relationships between objects may cause ambiguity in some cases, making it difficult to determine their relationship. This requires a higher level of spatial reasoning.

Most MLLMs show poor performance for text recognition. Apart from InstructBLIP, all other models achieve an accuracy lower than 40% for text recognition due to the lack of textual elements in multimodal pre-training datasets. Since the ability to accurately identify and extract text from images is important, future work should develop models that are better equipped to handle text recognition by pre-training on datasets with rich textual elements in visual data.

VideoLLMs achieve promising results on spatial understanding. For example, VideoChat achieves 39.98% accuracy (ranking 4-th on the recognition of instance attribute, surpassing LLaVa by 11.55% and performing only 3.58% lower than the top-1 model. It shows that VideoChat's ability of spatial understanding does not degrade by jointly training on both image and video data during the pre-training and instruction-tuning stages.

Most MLLMs exhibit unsatisfactory performance on fine-grained temporal understanding. It is notable that on the evaluation of procedure understanding, the top-ranked model, VPGTrans, achieves an accuracy that is only 5% higher than that of LLaMA. The performance improvement of the following 4 MLLMs is even less than 1.2% compared with LLaMA. This demonstrates that it is extremely difficult for both the ImageLLMs and VideoLLMs to perform fine-grained temporal reasoning so that they can recognize and sort the key actions in a video.

VideoLLMs fail to achieve competitive performance on temporal understanding. Although VideoLLMs are instruction-tuned on video data, they do not exhibit a significant advantage on evaluation dimensions for temporal understanding. Surprisingly, two VideoLLMS (Video-ChatGPT and Valley) even perform worse than most ImageLLMs on action recognition, action prediction and procedure understanding. It indicates that the capabilities of existing VideoLLMs for fine-grained action recognition, temporal relationship understanding and temporal reasoning are still limited. Similar concerns about existing VideoLLMs are also presented in recent works [15, 16].

5 Conclusion

In this work, we propose a large-scale benchmark SEED-Bench to provide a comprehensive and objective evaluation of Multimodal Large Language models (MLLMs) on generative comprehension. SEED-Bench consists of 19K multiple-choice questions with accurate human annotations, which covers 12 evaluation dimensions for both the spatial and temporal understanding. We conduct a thorough evaluation of 18 models, analyzing and comparing their performances to provide insights for future research. We plan to launch and consistently maintain a leaderboard, offering a platform for the community to assess model performance. We will continue to further broadening the evaluation dimensions of SEED-Bench with more data.

Acknowledgements

We sincerely acknowledge Junting Pan (CUHK MMLab) for the insightful suggestions, Zhan Tong (Nanjing University) for the data processing, and Yi Chen (Tencent AI Lab) for the engaging discussions.

References

[1] 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.

[2] OpenAI. Gpt-4 technical report, 2023.

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

[4] FastChat. Vicuna. https://github.com/lm-sys/FastChat, 2023.

[5] 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.

[6] 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.

[7] 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.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.

[22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023.

[28] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[30] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021.

[31] https://github.com/PaddlePaddle/PaddleOCR. Paddleocr.

[32] 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.

[33] 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.

[34] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[35] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[36] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020.

[37] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014.

[38] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

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

[40] 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.

[41] ml_foundations. Openflamingo. https://github.com/mlfoundations/open_flamingo, 2023.

[42] 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.