SEED-Bench: Benchmarking Multimodal Large Language Models

Supplementary Material

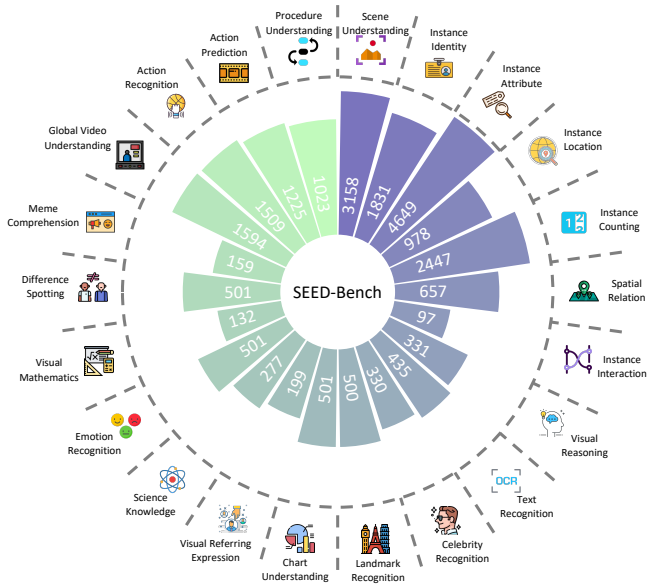

Figure 1. Overview of 22 evaluation dimensions in SEED-Bench capability L_1. The number in the bar denotes the number of multiple-choice questions in each dimension.

1. Evaluation Dimension

To thoroughly evaluate the diverse capabilities of MLLMs, SEED-Bench incorporates 27 assessment dimensions, encompassing Single-Image & Text Comprehension, Multiple-Images & Text Comprehension, Video & Text Comprehension, Interleaved Image & Text Comprehension, Image Generation, and Image & Text Generation. The dimensions within Single-Image & Text Comprehension, Multiple-Images & Text Comprehension, and Video & Text Comprehension are visually represented in Fig. 1.

Single-Image & Text Comprehension. The evaluation of single-image comprehension encompasses 16 dimensions, addressing global/object-level, recognition/reasoning, and various specialized domains.

- Scene Understanding: This dimension emphasizes global information in an image and necessitates 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, and 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.
- Instance Counting: This dimension necessitates that the model counts the number of specific objects in the image, understanding all objects and successfully counting the referred object's instances.
- 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.

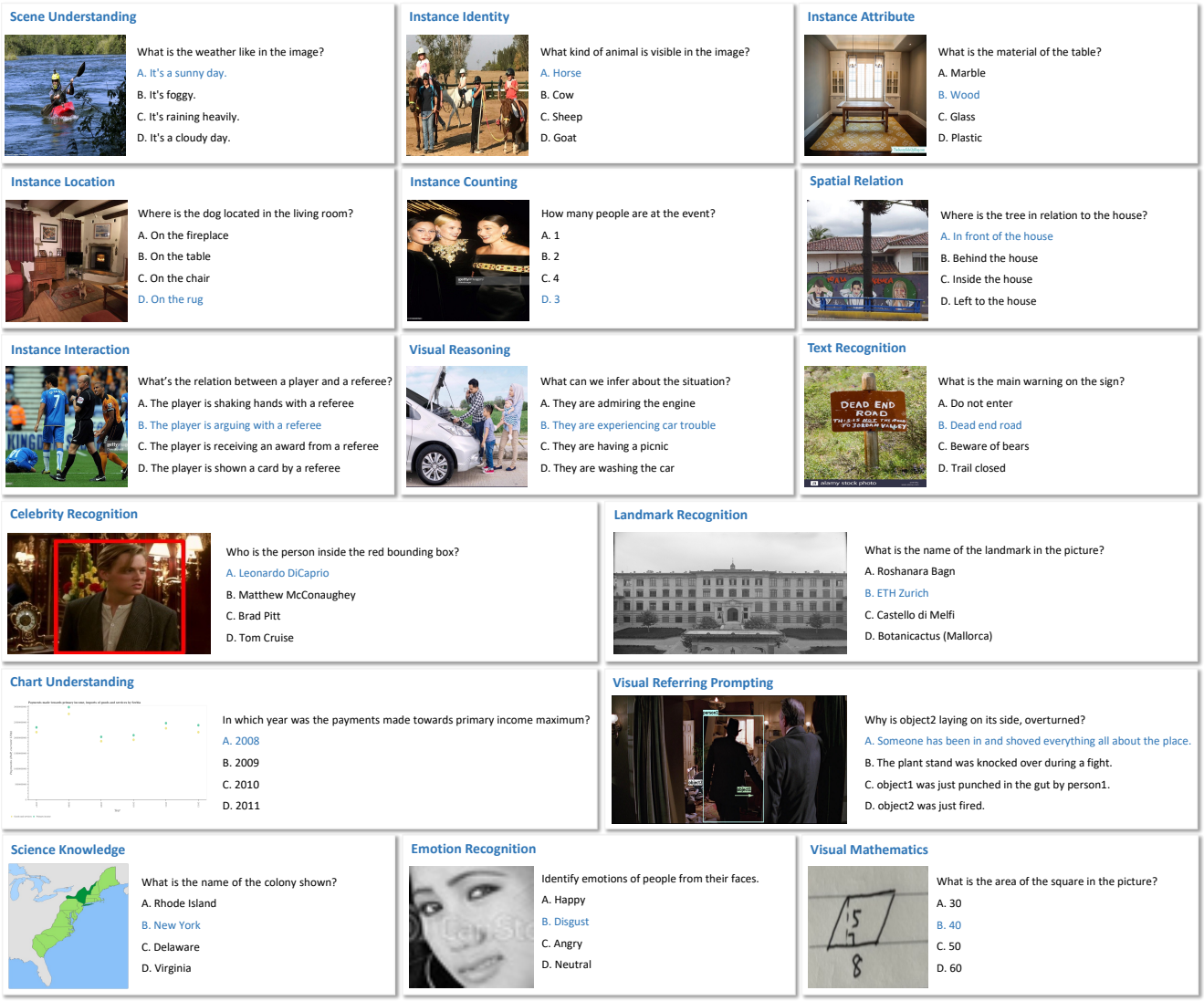

Figure 2. Data samples from a subset of evaluation dimensions in part-1 with single image as input, which encompasses capability L_1 in SEED-Bench.

- 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.

Multiple-Images & Text Comprehension. The evaluation of multiple-images comprehension comprises 2 dimensions: difference spotting and meme comprehension. These dimensions assess an MLLM's ability to extract information and discern differences from 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.

Video & Text Comprehension. For the evaluation of video comprehension, we propose 4 dimensions to assess an MLLM's ability to extract fine-grained information, temporal relationships, and reasoning through video content.

- 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

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 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 Location 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 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 3. Prompts of generating multiple-choice questions for different evaluation dimensions.

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	74.8
2	LLaVA-1.5	63.7
3	Kosmos-2	63.4
4	Emu	59.0
5	InstructBLIP	58.9

(1) Scene Understanding

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	70.5
2	LLaVA-1.5	62.4
3	Kosmos-2	57.1
4	Emu	50.0
5	InstructBLIP	49.7

(2) Instance Identity

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	67.6
2	LLaVA-1.5	66.7
3	InstructBLIP	61.7
4	Kosmos-2	58.5
5	Qwen-VL-Chat	54.8

(3) Instance Attribute

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	60.5
2	LLaVA-1.5	51.3
3	Qwen-VL-Chat	46.9
4	Kosmos-2	44.0
5	BLIP-2	39.1

(4) Instance Location

Rank	Model	Accuracy(%)
1	LLaVA-1.5	60.2
2	InstructBLIP	58.1
3	InstructBLIP Vicuna	56.5
4	InternLM-Xcomposer-VL	55.3
5	Qwen-VL-Chat	54.2

(5) Instance Counting

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	53.4
2	Qwen-VL-Chat	40.3
3	LLaVA-1.5	38.5
4	Kosmos-2	37.9
5	BLIP-2	36.2

(6) Spatial Relation

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	76.3
2	Kosmos-2	55.7
3	Qwen-VL-Chat	55.7
4	Emu	49.5
5	BLIP-2	48.5

(7) Instance Interaction

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	76.1
2	Kosmos-2	60.7
3	LLaVA1.5	59.8
4	Emu	58.3
5	MiniGPT-4	57.1

(8) Visual Reasoning

Rank	Model	Accuracy(%)
1	LLaVA-1.5	69.0
2	Kosmos-2	68.1
3	Emu	61.4
4	InstructBLIP	61.4
5	InternLM-Xcomposer-VL	61.4

(9) Text Recognition

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	86.1
2	Kosmos-2	82.1
3	mPLUG-Owl	70.9
4	Emu	68.8
5	Qwen-VL-Chat	62.4

(10) Celebrity Recognition

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	78.0
2	Emu	61.6
3	Qwen-VL-Chat	55.6
4	Otter	53.0
5	IDEFICS-9B-Instruct	52.8

(11) Landmark Recognition

Rank	Model	Accuracy(%)
1	LLaVA	30.3
2	VPGTrans	30.1
3	InstructBLIP Vicuna	27.9
4	InternLM-Xcomposer-VL	27.2
5	InstructBLIP	26.4

(12) Chart Understanding

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	60.3
2	Kosmos-2	48.2
3	LLaVA-1.5	45.7
3	Emu	45.7
5	mPLUG-Owl	44.2

(13) Visual Referring Expression

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	84.8
2	LLaVA-1.5	56.7
3	BLIP-2	52.4
4	InstructBLIP	47.7
5	mPLUG-Owl	44.0

(14) Science Knowledge

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	68.9
2	LLaMA-Adapter V2	39.7
3	Otter	37.3
4	InstructBLIP	34.5
5	VideoChat	34.33

(15) Emotion Recognition

Rank	Model	Accuracy(%)
1	Qwen-VL-Chat	28.8
2	Kosmos-2	28.0
3	MultiModal-GPT	27.3
3	OpenFlamingo	27.3
3	VPGTrans	27.3

(16) Visual Mathematics

Rank	Model	Accuracy(%)
1	IDEFICS-9B-Instruct	56.5
2	InternLM-Xcomposer-VL	47.7
3	Video-ChatGPT	46.1
4	GVT	41.5
5	MultiModal-GPT	40.1

(17) Difference Spotting

Rank	Model	Accuracy(%)
1	Video-ChatGPT	61.4
2	GVT	59.2
3	InternLM-Xcomposer-VL	56.6
4	MultiModal-GPT	56.5
5	InstructBLIP Vicuna	55.4

(18) Meme Comprehension

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	58.6
2	Kosmos-2	48.5
3	LLaVA1.5	46.7
4	LLaVA	46.1
5	IDEFICS-9B-Instruct	44.1

(19) Global Video Understanding

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	49.9
2	Qwen-VL-Chat	42.8
3	Emu	42.7
4	Kosmos-2	40.8
5	LLaVA-1.5	39.4

(20) Action Recognition

Rank	Model	Accuracy(%)
1	InstructBLIP	40.5
2	Kosmos-2	39.5
3	Emu	37.9
4	InternLM-Xcomposer-VL	37.6
5	BLIP-2	36.2

(21) Action Prediction

Rank	Model	Accuracy(%)
1	VPGTrans	33.5
2	Kosmos-2	30.0
3	MiniGPT-4	28.6
4	LLaVA1.5	28.1
5	VideoChat	27.4

(22) Procedure Understanding

Rank	Model	Accuracy(%)
1	Emu	51.7
2	NExt-GPT	46.7
3	MiniGPT-4	45.8
4	IDEFICS-9B-Instruct	45.8
5	Otter	42.5

(23) In-Context Captioning

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	36.7
2	IDEIFCS-9B-Instruct	34.7
3	GVT	34.7
4	InstructBLIP	34.7
5	OpenFlamingo	32.7

(24) Interleaved Image-Text Analysis

Rank	Model	Accuracy(%)
1	Emu	46.8
2	NExt-GPT	45.1

(25) Text-to-Image Generation

Rank	Model	Accuracy(%)
1	Emu	43.2
2	NExt-GPT	19.8

(26) Next Image Prediction

Rank	Model	Accuracy(%)
1	NExt-GPT	36.7
2	Emu	34.2

(27) Text-Image Generation

Figure 4. Each task leaderboard of SEED-Bench.

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.

Interleaved Image & Text Comprehension. For the evaluation of interleaved image-text data comprehension, we introduce 2 dimensions: in-context captioning and interleaved image-text analysis. These dimensions assess an MLLM's ability to extract information from arbitrary image-text data.

- 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

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	64.2
2	LLaVA-1.5	50.8
3	Kosmos-2	49.5
4	Qwen-VL-Chat	46.0
5	InstructBLIP	45.5

(1) Single-Image & Text Comprehension

Rank	Model	Accuracy(%)
1	Video-ChatGPT	53.8
2	IDEFICS-9B-Instruct	52.5
3	InternLM-Xcomposer-VL	52.2
4	GVT	50.4
5	MultiModal-GPT	48.3

(2) Multi-Image & Text Comprehension

Rank	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	42.8
2	Kosmos-2	39.7
3	Emu	36.1
4	LLaVA-1.5	35.7
5	InstructBLIP	35.7

(3) Video & Text Comprehension

Rank	Model	Accuracy(%)
1	Emu	41.1
2	IDEFICS-9B-Instruct	40.3
3	GVT	38.6
4	Otter	36.6
5	InstructBLIP	35.7

(4) Interleaved Image & Text Comprehension

Rank	Model	Accuracy(%)
1	Emu	45.0
2	NExt-GPT	32.4

(5) Image Generation

Rank	Model	Accuracy(%)
1	NExt-GPT	36.7
2	Emu	34.2

(6) Image & Text Generation

Figure 5. Subgroup task leaderboard of SEED-Bench.

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.

Image Generation. 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.

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

- Text-Image Creation: This dimension focuses on a model's ability to generate images with text. It evaluates the model's capacity to produce accurate text and visual content.

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, enabling us to construct diverse and challenging multiple-choice questions.

For dimensions 1-9, we utilize the CC3M [36] 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 [13]. 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 [27], IC13 [14], IIIT5k [31], and SVT [41] datasets to enlarge this dimension.

For the celebrity recognition dimension, we use celebrity data from MME [8] and MMBench [25] to conduct this dimension. As celebrity recognition comprises 4-choice questions in MMBench and T/F questions in MME, we use GPT-4 to generate confusing options for MME data to construct 4-choice questions.

For the landmark recognition dimension, we use the Google landmark dataset v2 [42] train set as the data source and generate selections by randomly selecting other landmark names.

For the chart understanding dimension, we use the plotQA [30] test set and generate selections using GPT-4 by inputting corresponding image captions.

For the visual referring expression dimension, we use the VCR [46] valid dataset as the data source, and we use four methods to indicate the object in the picture: drawing a

bounding box, drawing a circle, drawing a mask, and drawing an arrow.

For science knowledge, we use the scienceQA [26] test set, which contains image data for each question as the data source.

For emotion recognition, we use the fer2013 [5] test dataset as the image source and use the 6 emotions in the dataset as selections.

For visual mathematics, we use the math part of the MME [8] dataset and generate some questions by humans.

For difference spotting, we use the SD part of the MIMICIT [18] dataset as the image source and generate selections using GPT-4.

For meme comprehension, we generate questions by humans.

For global video understanding, we select the Charades [37] test dataset as the video source, as the videos in the dataset contain rich information. For each video, we use tag2text [13] to generate each second caption and grit [43] 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 action prediction, we adopt Something-Something-v2 (SSV2)[11], and Epic-kitchen 100 [4] datasets to build questions and let human annotators filter the questions. SSV2 is an action recognition dataset that includes 174 fine-grained categories of basic actions with everyday objects, and we adopt 1509 videos from its validation set. We also select 138 long videos from the Epic-kitchen 100 dataset with temporally annotated action labels. Moreover, videos and fine-grained action segmentation annotations in the Breakfast dataset [16] are utilized for the procedure understanding task.

For in-context captioning, we use the ground-truth captions generated by the instance attribute dimension and instance counting dimension to formulate questions and correct options. We then use captions from other dimensions to create distractors. For each caption in the instance attribute, we employ GPT-4 to classify and ensure that the same properties are described for different objects across various images.

For interleaved image-text analysis data, we generate questions by humans.

For text-to-image generation, we firstly use GPT-4 to modify the target categories or attributes in the prompt of CC-500 [7] dataset and ABC-6k [7] dataset and form a four-choice question. We then use Stable-Diffusion-XL [35] to generate each prompt and let human annotators to filter unqualified data.

For the next image prediction dimension, we use Epic-kitchen 100 [4] dataset and start-end frame in action prediction dimension to form this dimension.

For text-image creation, we generate questions by humans.

Part 1	Model	Accuracy(%)
1	InternLM-Xcomposer-VL	59.2
2	LLaVA-1.5	47.3
3	Kosmos-2	46.3
4	Qwen-VL	43.1
5	Emu	42.5

Part 2	Model	Accuracy(%)
1	Emu	41.1
2	IDEFICS-9B-Instruct	40.3
3	GVT	38.6
4	Otter	36.6
5	InstructBLIP	35.7

Part 3	Model	Accuracy(%)
1	Emu	41.4
2	NExt-GPT	33.9

Figure 6. Part leaderboard of SEED-Bench.

3. Automatic Pipeline

In this section, we provide a detailed discussion of the automatic pipeline for constructing multiple-choice questions for dimensions 1-9.

Visual Information Extraction. For constructing questions related to spatial understanding, we interpret the rich information in each image with texts using multiple pre-trained 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 [20] and Tag2Text [13] 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 [15] to segment each instance in the image and obtain their bounding boxes according to the segmentation results. The object labels are obtained using Tag2Text [13]. Besides,

Table 1. Evaluation results of various MLLMs in 'Single-Image & Text Comprehension' part of SEED-Bench. The best (second best) is in bold (underline). The corresponding brackets for each task indicate the number of associated questions.

Model	Language Model	Scene Understanding (3138)	Instance Identity (1831)	Instance Attribute (4649)	Instance Location (978)	Instance Counting (2447)	Spatial Relation (657)	Instance Interaction (97)	Visual Reasoning (331)	Text Recognition (435)	Celebrity Recognition (330)	Landmark Recognition (500)	Chart Understanding (501)	Visual Referring Expression (199)	Science Knowledge (277)	Emotion Recognition (501)	Visual Mathematics (132)
BLIP-2 [21]	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
InstructBLIP [3]	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
InstructBLIP Vicuna [3]	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
LLaVA [24]	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
MiniGPT-4 [50]	Vicuna-7B	56.3	49.2	45.8	37.9	45.3	32.6	47.4	57.1	41.8	55.2	45.2	20.2	41.2	43.3	24.2	25.0
VPGTrans [47]	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
MultiModal-GPT [10]	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
Otter [19]	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
OpenFlamingo [32]	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
LLaMA-Adapter V2 [9]	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
GVT [40]	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
mPLUG-Owl [45]	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
Kosmos-2 [34]	Decoder only 1.3B	63.4	57.1	58.5	44.0	41.4	37.9	55.7	60.7	68.1	82.1	51.4	21.2	48.2	43.7	30.7	28.0
Qwen-VL-Chat [1]	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**
LLaVA-1.5 [23]	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
IDEFICS-9B-Instruct [17]	LLaMA-7B	48.2	38.2	37.8	32.9	29.0	32.4	37.1	54.1	45.5	52.4	52.8	22.6	42.7	33.2	26.6	21.2
InternLM-Xcomposer-VL [49]	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**	**68.9**	25.8
VideoChat [22]	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
Video-ChatGPT [29]	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
Valley [28]	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
Emu [38]	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
NExt-GPT [44]	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

Table 2. Evaluation results of various MLLMs in 'Multi-Images & Text Comprehension' part, 'Video & Text Comprehension' part, 'Interleaved Image & Text Comprehension' part, 'Image Generation' part, 'Image & Text Generation' part of SEED-Bench. The best (second best) is in bold (underline). The corresponding brackets for each task indicate the number of associated questions.

Model	Language Model	part 1						part 2		part 3		
		Multi-Images & Text Comprehension		Video & Text Comprehension				Interleaved Image & Text Comprehension		Image Generation		Image & Text Generation
		Difference Spotting (501)	Meme Comprehension (159)	Global Video Understanding (1594)	Action Recognition (1509)	Action Prediction (1225)	Procedure Understanding (1023)	In-Context Captioning (120)	Interleaved Image-Text Analysis (49)	Text-to-Image Generation (1008)	Next Image Prediction (81)	Text-Image Creation (79)
BLIP-2 [21]	Flan-T5-XL	17.8	38.6	42.5	37.7	36.2	22.9	40.0	30.6	-	-	-
InstructBLIP [3]	Flan-T5-XL	22.8	35.2	41.5	36.1	**40.5**	24.5	36.7	34.7	-	-	-
InstructBLIP Vicuna [3]	Vicuna-7B	36.5	55.4	40.4	38.6	31.2	15.6	26.7	32.7	-	-	-
LLaVA [24]	LLaMA-7B	27.0	50.0	44.1	36.2	25.1	18.6	40.0	20.4	-	-	-
MiniGPT-4 [50]	Vicuna-7B	19.0	46.7	39.0	38.7	27.4	28.6	**45.8**	22.5	-	-	-
VPGTrans [47]	LLaMA-7B	24.6	44.0	37.8	38.2	20.9	**33.5**	19.2	28.6	-	-	-
MultiModal-GPT [10]	Vicuna-7B	40.1	56.5	37.6	38.7	25.3	24.4	39.2	30.6	-	-	-
Otter [19]	LLaMA-7B	27.4	46.7	36.6	37.9	26.0	24.8	42.5	30.6	-	-	-
OpenFlamingo [32]	LLaMA-7B	39.9	54.9	37.6	38.4	25.2	24.1	38.3	32.7	-	-	-
LLaMA-Adapter V2 [9]	LLaMA-7B	29.1	52.2	41.9	38.2	18.8	20.3	-	-	-	-	-
GVT [40]	Vicuna-7B	41.5	59.2	40.4	29.7	26.3	24.1	42.5	34.7	-	-	-
mPLUG-Owl [45]	LLaMA-7B	33.5	54.9	42.0	37.8	18.3	19.3	29.2	28.6	-	-	-
Kosmos-2 [34]	Decoder only 1.3B	25.2	42.8	48.5	40.8	39.5	30.0	24.2	22.5	-	-	-
Qwen-VL-Chat [1]	Qwen-7B	34.3	47.2	39.7	42.8	29.6	19.1	42.5	28.6	-	-	-
LLaVA-1.5 [23]	Vicuna-7B	35.7	50.3	46.1	39.4	29.4	28.1	39.2	22.5	-	-	-
IDEFICS-9B-Instruct [17]	LLaMA-7B	**56.5**	48.4	42.7	38.6	23.6	20.5	45.8	34.7	-	-	-
InternLM-Xcomposer-VL [49]	InternLM-7B	47.7	56.6	**58.6**	**49.9**	37.6	24.9	27.5	**36.7**	-	-	-
VideoChat [22]	Vicuna-7B	30.3	51.6	41.5	34.0	30.6	27.4	40.0	30.6	-	-	-
Video-ChatGPT [29]	LLaMA-7B	46.1	**61.4**	42.6	32.2	27.0	19.0	37.5	24.5	-	-	-
Valley [28]	LLaMA-13B	37.1	52.2	31.5	32.1	21.9	26.5	35.8	28.6	-	-	-
Emu [38]	LLaMA-13B	29.3	37.1	41.9	42.7	37.9	21.8	51.7	30.6	**46.8**	**43.2**	34.2
NExt-GPT [44]	Vicuna-7B	24.2	39.0	35.5	33.8	25.6	24.5	46.7	24.5	45.1	19.8	**36.7**

we also utilize attribute detector [48] to obtain the attributes of each instance in the image. Finally, we employ GRiT [43] 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 [12] for detecting textual elements.

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 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 [33] 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 for generating multiple-choice questions for different evaluation dimensions are listed in Fig. 3.

Automatic Filtering. Our benchmark aims to evaluate 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 [6], Flan-T5-XXL [2] and LLaMA-7B [39] 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 dimensions. If one question can not be answered based on the visual input, does not have any correct choice, or has multiple correct choices, it will be discarded by human annotators.

4. Evaluation Results

Detailed evaluation result for 22 models in 27 tasks is presented in Tab. 1 and Tab. 2. In these tables, the best and second-best performances for each task are indicated in bold and underlined, respectively.

Additionally, the leaderboards for each task, sub-part, and part are displayed in Fig. 4, Fig. 5, and Fig. 6.

5. More Observation

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. Emu, IDEFICS-9B-Instruct and Otter 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 demonstrates 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.

Compared to 'Single-Image & Text Comprehension' and 'Multiple-Images & Text Comprehension', MLLMs still have significant potential for improvement in 'Video & Text Comprehension'. Although MLLMs achieve competitive performance in 'Single-Image & Text Comprehension' and 'Multiple-Images & Text Comprehension' sub-part tasks, most MLLMs struggle to accurately answer questions in 'Video & Text Comprehension' sub-part tasks, particularly in 'Action Prediction' and 'Procedure Understanding'.

As input data becomes more complex, MLLMs exhibit reduced performance. As illustrated in Fig. 5, the performance of top MLLMs decreases as the complexity of the input data increases. This demonstrates that MLLMs still have significant potential for improvement when processing complex data constructs.

InternLM-Xcomposer-VL achieves top performance in 14 out of 24 evaluation dimensions. InternLM-Xcomposer-VL [49] demonstrates exceptional performance across 24 dimensions, achieving the top performance in 14 tasks and ranking within the top 5 in 21 tasks, as shown in Fig. 4.

Emu exhibits outstanding performance across all parts of SEED-Bench. As depicted in Fig. 6, Emu [38] ranks within the top 5 for each part and achieves the top position in parts 2 and 3 of SEED-Bench, demonstrating its versatility and robustness.

References

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

[2] 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. 8

[3] 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. 7

[4] 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. 6

[5] Dumitru, Ian Goodfellow, Will Cukierski, and Yoshua Bengio. Challenges in representation learning: Facial expression recognition challenge, 2013. 6

[6] FastChat. Vicuna. https://github.com/lm-sys/FastChat, 2023. 8

[7] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 6

[8] 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. 5, 6

[9] 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. 7

[10] 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. 7

[11] 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. 6

[12] https://github.com/PaddlePaddle/PaddleOCR. Paddleocr. 7

[13] 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. 5, 6

[14] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013. 5

[15] 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. 6

[16] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014. 6

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

[18] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 6

[19] 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. 7

[20] 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. 6

[21] 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

[22] 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. 7

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 7

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 7

[25] 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. 5

[26] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 6

[27] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)*, 7:105–122, 2005. 5

[28] 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. 7

[29] 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. 7

[30] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020. 5

[31] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012. 5

[32] ml_foundations. Openflamingo. https://github.com/mlfoundations/open_flamingo, 2023. 7

[33] OpenAI. Gpt-4 technical report, 2023. 7

[34] 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. 7

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6

[36] 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. 5

[37] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 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*, pages 510–526. Springer, 2016. 6

[38] 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. 7, 8

[39] 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. 8

[40] 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. 7

[41] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International conference on computer vision*, pages 1457–1464. IEEE, 2011. 5

[42] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 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*, pages 2575–2584, 2020. 5

[43] 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. 6, 7

[44] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 7

[45] 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. 7

[46] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 5

[47] 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. 7

[48] 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. 7

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

[50] 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. 7