MMRA: A Benchmark for Evaluating Multi-Granularity and Multi-Image Relational Association Capabilities in Large Visual Language Models

Anonymous ACL submission

Abstract

Current multi-modal benchmarks primarily focus on facts within individual images. However, they overlook the associative relations among multiple images, which necessitate conduct **commonsense reasoning** grounded in the associated knowledge at different granularities (i.e., "**image**" and "**entity**") and the ability to perceive **image order**. Therefore, we propose the multi-image relation association task and a meticulously curated **M**ulti-granularity **M**ulti-image **R**elational **A**ssociation (**MMRA**) benchmark, comprising 1,024 samples. In order to systematically evaluate current LVLMs, we establish an associational relation system among images that contain **11 subtasks** (e.g, UsageSimilarity, SubEvent, etc.) at two granularity levels (i.e., "**image**" and "**entity**") according to the relations in ConceptNet. Our experiments reveal that entity-level multi-image perception tasks pose a greater challenge for LVLMs compared to image-level tasks. Moreover, LVLMs perform poorly on spatial-related tasks, indicating that LVLMs have limited spatial awareness. Furthermore, we find that the LVLMs' **image order perception** capability is relatively poor and design a method to significantly improve the ability of LVLMs, which demonstrates that the majority of current LVLMs do not adequately consider image order perception during the pre-training process.

1 Introduction

Due to the development of Large Visual Language Models (LVLMs) (Li et al., 2023; Liu et al., 2024b,a; Bai et al., 2023; AI et al., 2024), there is growing interest in systematically and comprehensively defining benchmarks to assess the performance of LVLMs and guide future development in this field. However, current multi-modal benchmarks (Singh et al., 2019; Yuan Liu et al., 2023; Yue et al., 2024) focus on asking questions of a single image, and evaluation of LVLMs' multi-image association ability (e.g., "those images all depict outdoor scenes" as shown in Fig 1) is overlooked.

Current benchmarks overlook association relationships among multiple images. (1) The multi-image benchmarks, such as MuirBench (Wang et al., 2024) and MIRB (Zhao et al., 2024), merely focus on factual questions about visual elements in the images (e.g., ***How many gloves are there in the two pictures?***). However, they overlook the commonsense reasoning that is needed to mine the commonsense knowledge within two images (e.g., ***The truck in Image 1 is used for transporting goods + In Image 2, items are placed on the skateboard and glided along –>They share the same function: carry items.***). (2) Mining relations among multiple images across different granularities (e.g., entity vs. image level) and properties (e.g., spatial vs. temporal) poses varying challenges. Categorizing tasks by these dimensions helps diagnose LVLM performance gaps and guide targeted improvements. However, most tasks in existing benchmarks mainly focus on entity or text in images. (3) Current multi-image benchmark overlooks the model's ability to perceive the order of images. However, this capability is crucial for complex multi-image tasks, such as Image temporal order recognition.

To explore the multi-image association capabilities of LVLMs, we propose a multi-image relation association task, which requires LVLMs to discern the potential relations between two images (for instance, recognizing that the car and the knife, each present in different images, are both made of iron). We manually curated a high-quality **M**ulti-granularity **M**ulti-image **R**elational **A**ssociation (**MMRA**) benchmark, consisting of 1.024 samples, for evaluating the multi-image perception capabilities of LVLMs. Based on the relations in ConceptNet (Speer et al., 2017) and observations of potential connections between images, we define an associational relation system, which consists

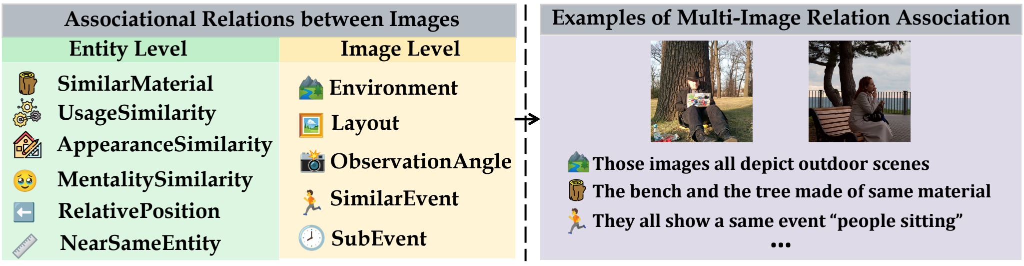

Figure 1: Overview of the MMRA benchmark. **Left**: image Associational Relations extended from the ConceptNet; **Right**: the examples of Multi-Image Relation Association task.

of 6 subtasks at the entity-level granularity (i.e., RelativePosition, NearSameEntity, etc.) and 5 subtasks at the image-level granularity (i.e., Layout, Environments, etc.) across different perspectives of mining relations between images (see Fig 1).

We employ an LVLM to generate detailed descriptions of the images and evaluate both LVLMs and LLMs using our MMRA benchmark across four distinct input configurations: Image+Question (IQ), Description+Question (DQ), Image+Description+Question (IDQ), and Question Only (QO). Furthermore, we reverse the image orders of MMRA to investigate the LVLMs' image order perception ability and annotate a training dataset, containing 1,500 samples, to improve the image order perception ability of LVLMs.

We present our key insights as follows:

1. Based on the results of the IQ and QO setting, we found that closed-source models like GPT-4o, GPT-4v, and Gemini-Flash outperformed all open-source models. In particular, GPT-4o achieved SOTA overall performance. Additionally, different models exhibit significant performance variations across different subtasks. Some open-source models even surpassed GPT-4 in certain subtasks.
2. Compared to entity-level tasks, models generally perform better on image-level tasks, and their performance tends to be relatively poor in tasks related to spatial awareness. It indicates that current LVLMs have weak fine-grained multi-image association capabilities and are not proficient in handling spatial perception tasks.
3. We examine the image order perception capabilities of LVLMs by altering the order of input image pairs. With the exception of Idefics2, most open-source LVLMs scored relatively low. Moreover, to enhance the image order perception ability of LVLMs, we manually annotate a high-quality dataset for fine-tuning. As a result, the order perception ability of LVLMs is significantly improved through supervised fine-tuning (SFT). This suggests that current LVLMs are inadequate in modeling images' order during the pre-training phase.

2 Related Work

Large Visual Language Models. With the emergence of LLMs, researchers have applied it to the multimodal perception field. More and more LVLMs have achieved excellent success on single-image tasks, such as BLIP2 (Li et al., 2023), LLaVA (Liu et al., 2024b), LLaVA-Next (Liu et al., 2024a), QwenVL (Bai et al., 2023), CogVLM (Wang et al., 2023), and Yi-VL(AI et al., 2024). Those LVLMs all demonstrate exceptional ability on single image tasks, such as TextVQA (Singh et al., 2019), VQAV2 (Goyal et al., 2017), MMBench(Yuan Liu et al., 2023), GQA(Hudson and Manning, 2019). Although Fuyu-8B[1], Kosmos2 (Peng et al., 2023), and Flamingo (Alayrac et al., 2022) support interleaved input, they do not optimize in multi-image task.

Multi-Image Perception Model and Task. Currently, some researchers have realized the importance of the multi-image ability of LVLMs. Excepting Kosmos2, Fuyu and Flamingo, there are some models which support multi images input, such as Mantis, Idefic2, Phi3v and Mantis-Idefic2 (Sun et al., 2023; Laurençon et al., 2024; Rasheed et al., 2024; Jiang et al., 2024). Besides, the Emu2(Sun et al., 2023) is a generative multimodal model that supports the interleaved text-image inputs. And the video understanding models (Zhang et al., 2023; Ren et al., 2023) also have the multi-image perception ability, but it is relatively worse than LVLMs.

[1] https://www.adept.ai/blog/fuyu-8b

Meanwhile, there is also a lack of comprehensive and systematic evaluation of multi-image LVLMs. The earliest task is the description of the differences in the multi images, and researchers have developed many datasets, such as Spot-the-Diff and Birds-to-Words (Jhamtani and Berg-Kirkpatrick, 2018), etc. However, they are all generative tasks. Recently, the MuirBench (Wang et al., 2024) and the multi-image understanding benchmark (Zhao et al., 2024) focus on evaluating the LVLMs' ability, but they do not systematically define relations among images in real-life scenario.

Commonsense Reasoning. During the previous research in NLP, there are numerous works for commonsense reasoning (Du et al., 2022; Zhao et al., 2023; Gao et al., 2022; Jiang et al., 2021; Emelin et al., 2021) and would use many predefined commonsense knowledge (i.e., Knowledge Graph (Sap et al., 2019; Speer et al., 2017; Shen et al., 2023)). The Commonsense Knowledge Graph (CSKG), such as ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019), is comprehensively used in the commonsense reasoning tasks because they define numerous relations between event node and entity node. The current multi-image benchmarks (Wang et al., 2024; Zhao et al., 2024) do not define the relation system among images. Although VCD (Shen et al., 2024) uses the knowledge system in ConceptNet to mine the potential knowledge in a single image, it cannot be directly applied to the multi-image setting. In this work, we will define a relation system among different images and curate a benchmark.

3 Dataset Curation

3.1 Image Pair Selection

Given that most tasks in the MMRA benchmark require a specific relation between paired images, we use the semantic similarity of image captions to identify and select image pairs with relatively higher relevance. This aims to reduce the complexity of annotation. To be specific, we randomly chose the images in the LLaVA-665k-multi dataset and crawl some images from the internet to form an image pair. We then utilize the Sentence-BERT (Reimers and Gurevych, 2019) to calculate the semantic similarity and filter the image pair with a score below 0.5. Finally, we obtained 3,403 image pairs for annotation.

3.2 Subtask Definition

As shown in the Fig 6 in Appendix E, based on the perspective of humans observing images, we divide our tasks into two granularity levels (i.e., entity and the whole image). Because the ConceptNet comprehensively defines the relations among different textual events and entities, most of our subtasks are extended from it. Besides, we design some subtasks from a visual perspective (i.e., Layout and ObservationAngle).

Entity level. We primarily consider the mental state, appearance, and location information of different objects in the images, as well as the psychological characteristics of individual creatures.

- **RelativePosition (RP)**: The 'AtLocation' is an important relation in ConceptNet to express A is the inherent location of B. As for the entity in two images, we extend this relation into the subtask which judges the relative position of entities in the image. For example, we ask LVLMs to judge which two entities, respectively in different images, have the same relative position (e.g., all at the upper left of images).
- **NearSameEntity (NSE)**: The relation 'LocatedNear' in ConceptNet expresses "A and B are typically found near each other". Based on it, we design a subtask, 'NearSameEntity', which requires LVLMs to determine whether there are entities, respectively in different images, near the same object.
- **MentalitySimilarity (MS)**: 'HasProperty' in ConceptNet is a relation that describes the characteristics of an entity. We think the emotional property expressed by the images could directly affect humans. Thus, we extend this relation to a subtask that requires LVLMs to determine whether the creatures in two images have similar emotions, attitudes, or feelings (e.g., happy, excited, serious, surprised, etc.).
- **AppearanceSimilarity (AS)**: The physical characteristics of the entity is also an important factor. So we design a subtask that is also relevant to 'HasProperty' and that requires LVLMs to determine whether two images have entities that are physically similar in appearance (e.g., the shape and color of objects, the body and hairstyle of humans).
- **SimilarMaterial (SM)**: The relation 'MadeOf' in ConceptNet expresses 'A is

made of B'. Therefore, we design the subtask 'SimilarMaterial' which requires LVLMs to judge whether there are entities, respectively in different images, with the same production materials.

- **UsageSimilarity (US)**: Apart from the aforementioned aspects, we have also devised a subtask that requires LVLMs to discern whether the entities, respectively in two images, have the same usage according to the ConceptNet's relation 'UsedFor' which express "the purpose of A is B".

Image level. We primarily consider the correlation between the events expressed by the whole image as well as the overall spatial structural similarities of different images.

- **Layout (LO)**: At the image granularity, we regard the layout of the image as a representation of the relation "AtLocation". We design a subtask that requires the LVLMs to determine whether there are similarities in layout between images according to the relation 'NearBy'.
- **Environment (Env)**: From the visual perspective, the environment of the image is also an important content that humans tend to notice (e.g., both images depict the streets of a European country with a Gothic architectural style). So, we design a subtask that lets LVLMs judge if the environments in those images are similar according to the relation 'AtLocation'.
- **SubEvent (SubE)**: The temporary relation is an important connection between two images. Therefore, we extend the relation 'SubEvent' to a subtask that requires LVLMs to determine whether the two images describe events that occurred at the same scene in two consecutive moments.
- **SimilarEvent (SimE)**: Excepting the 'SubEvent', the similar event is also a crucial factor when associating multi images. So we devise a subtask to evaluate the LVLMs' capability to find the same event that happened in the given two images.
- **ObservationAngle (OA)**: In addition to the 'Layout', we create a subtask for the model to determine whether one of the images is a close-up, inside shot, or different parallel angle shot of another image for the sake of exploring the view perception ability of LVLMs according to the relation 'LocatedNear' in ConceptNet.

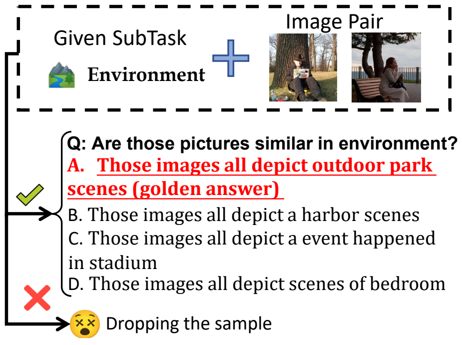

Figure 2: The process of annotation.

3.3 Data Annotation

We hire four annotators specializing in multimodal research to annotate data. Each annotator was assigned 2-3 tasks.

Annotation Process. As shown in Fig 2, each annotator is provided with two images and a certain subtask (i.e., Environment). Their responsibility is to determine whether they could design a question based on the given task for the image pair. If the image pair meets the task requirements, they proceed to annotate a question and options (either multiple-choice or true/false) for that pair. The annotator terminates annotating a task once they reach a predetermined number of labelled samples (i.e., 90) or once all the image pairs for that task have been annotated.

Quality Control. We conduct cross-validation on the annotated data. Specifically, each annotator reviews 2-3 tasks labeled by their peers. If any annotated samples do not meet the task requirements or if the answers derived from the images and options do not match the correct answer, those samples are removed. Quality control is concluded once all annotators agree that their verified portion satisfies the specified requirements.

3.4 Elimination of Answer Leakage from Questions and Options

When designing multiple-choice options at the entity level, we need to identify potential entities that could be regarded as the correct answer to the question and provide justifications. For example, as illustrated in Fig 1, 'both tree and bench are made of wood' can be the answer to the SimilarMaterial subtask. However, language models can sometimes

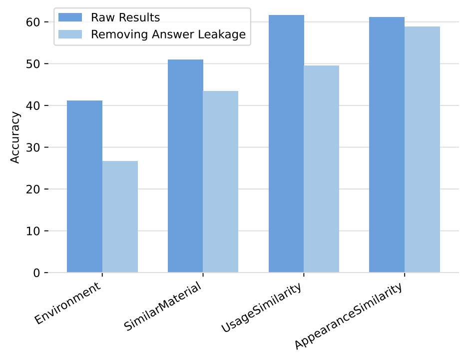

Figure 3: Comparing results before and after textual answer leakage elimination.

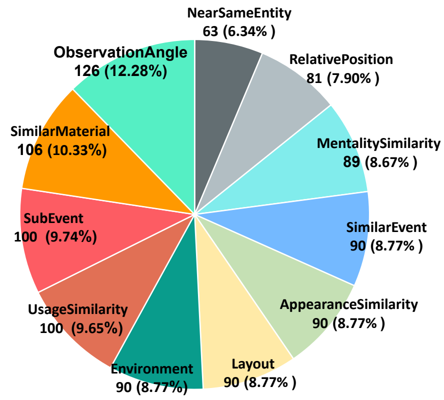

Figure 4: The number and ratio of each subtask in MMRA. The integers in the graph represent the number of samples in each task, while the percentages in parentheses indicate the proportion of each task.

deduce the correct answer simply by analyzing the textual content in the options. Additionally, annotators often unconsciously label the correct answer with greater detail and specificity, and the language model tends to choose these more detailed options. To eliminate these biases, we optimize the questions and options for subtasks where the language model scores higher than the expected accuracy by randomly answering the question. For instance, the expected accuracy for true/false questions is 50%, and for multiple-choice questions with four options, it is 25%.

We refine the options and questions for four subtasks (i.e., UsageSimilarity, Environment, MadeOf, and AppearanceSimilarity), because language models exhibit relatively higher performance on them. As shown in Fig 3, we presented the accuracy changes of the Yi-1.5-9B model before and after answer leakage removal. We have significantly reduced the leakage of answers in the question and option texts. After refining our benchmark, the performances on these subtasks are close to the expected random accuracy rates for their respective task types.

For the UsageSimilarity subtask, the performance of language models remains significantly higher than random expectations. We hypothesize that this is because mining the similarity in usage between two entities, a type of general commonsense knowledge, relies heavily on the language model's inference capabilities. Additionally, the commonsense reasoning capabilities of language models make them adept at identifying subtle differences among the options.

Data Statistics As shown in Fig 4, we obtain a total of 1,024 annotated samples. To maintain the balance of samples of the subtasks, we endeavored to maintain that the number of samples for all tasks is around 90. The ObservationAngle task has the highest proportion in the entire benchmark, with a total of 126 samples (12.28%). Due to the difficulty of labeling in the NearSameEntity task, we removed some samples with inconsistent opinions from different annotators during the quality control process and this subtask only has 65 samples.

4 Experiment

4.1 Experiments Setting

To explore the impact of LVLM's image-captioning ability on its multi-image perception, we design four input settings: **(1) Image + Question (IQ).** In this setting, we just include the image pair and question in the prompt. **(2) Description + Question (DQ).** To investigate the impact of the image caption capability of LVLMs on the perception of multiple images, we include a detailed description of the image pair and question in the prompt. **(3) Image + Description + Question (IDQ).** Besides, we also include the image pair, its description, and question in the prompt to compensate for the content of the image that cannot be described in the text. **(4) Question Only (QO).** For the sake of inspecting whether the answer to the questions in our benchmark is leaked in the textual information of options and questions, we only input the question to let LVLMs answer.

4.2 Baselines

As shown in Tab 6 in Appendix, we evaluated our benchmark on both mainstream closed-source and open-source large models. Regarding close-source LVLMs, we choose OpenAI's **GPT4o** and **GPT4v**, as well as Google's **Gemini-Flash** and **Gemini-Pro**. As for the open-source LVLMs, we mainly evaluate those supporting multi-image inputs (i.e., **Idefics2**, **Qwen-VL-Chat**, **Phi3v**, **Mantis-Idefics2**). Besides, we also assess the open-source LLMs (i.e., **LLaMA**, **Qwen**, and **Yi**) under the text-only input setting. In addition to the above LVLMs, we further evaluate some small visual encoder models, such as **CLIP** (Radford et al., 2021) and **MetaCLIP** (Xu et al., 2023, 2024).

4.3 Evaluation Protocol

Prompt. As for each task, we all design a prompt to make LVLMs directly generate textual format answers to the question. Except for including the content of different input settings, we let LVLMs generate the 'A', 'B', 'C' or 'D' for the choice questions, and 'Yes' or 'No' for the T/F questions. Besides, we also add the options to the prompt. As for further details about our prompt design, please refer to the Tab 5 in Appendix A.

Retrieval Method. For MetaCLIP and CLIP, we directly calculate the similarity between the query (image+question) and the answer options, and choose the option with the highest similarity as the model-predicted answer. The details of the retrieval method are provided in Sec. B.

Answer Matching and Metric. Because the golden answer in our benchmark is in the format of option id (i.e., 'A', 'B', 'C' and 'D') or judgment (i.e., 'Yes' or 'No'), we design a rule to match the response of LVLMs with the golden answer. Finally, we use accuracy of the matching results as the score of those models. Please refer to Appendix E for details of our designed matching rule.

5 Result Analysis

5.1 Overall Analysis

As shown in Table 1, when inputting question and image pairs (Image+Question setting), the close-source model (i.e., GPT-4v, GPT-4o, Gemini-Pro, and Gemini-Flash) achieves the best performance on our MMRA benchmark, with overall accuracy surpassing 60%. In contrast, the overall performance of other open-source multi-image LVLMs ranges from 50% to 60%, with the exception of Qwen-VL-Chat whose score is only 47.45%. The Visual Encoder models, such as CLIP and Meta-CLIP, exhibit performance comparable to Qwen-VL-Chat and InternVL2-2B.

Although LVLMs demonstrate varying performances across different subtasks, their average performance at the entity level is generally lower than at the image level. The LVLMs' performance is notably high for the Environment (Env) and SubEvent (SubE) subtasks, with most of the LVLMs scoring over 80%. This may be because these subtasks primarily require abstract image-caption information, which LVLMs have learned during the pre-training phase. It is worth mentioning that spatial perception subtasks, {i.e., RelativePosition (RP), NearSameEntity (NSE), Layout (LO), and ObservationAngle (OA)}, remain challenging for LVLMs, as most models' accuracy is below 50% for these subtasks.

At the Question-Only (QO) setting, the performance of LLMs on the UsageSimilarity (US) task consistently exceeds 60%, which is comparable to the performance of multi-image LVLMs under IO setting. This suggests that the reasoning required by the UsageSimilarity (US) subtasks relies on commonsense knowledge inherent in the language model component of LVLMs. Under the QO setting, all models achieve significantly lower overall scores compared to the IQ setting, indicating that MMRA has been well-cleaned to prevent answer leakage in the textual content.

5.2 Impact of Image Input

As shown in Table 1, when providing both image pairs and questions (i.e., the Image + Question setting), multi-image LVLMs demonstrate significantly better performance compared to LLMs under the QO setting (i.e., Question Only). To highlight the performance improvement of LVLMs due to image input across various tasks, we calculate the average performance of all LLMs on each task as a standard. By comparing LVLMs' performance with this standard, we can quantify the actual enhancement brought about by incorporating images.

As shown in Fig 5 in Appendix F, compared to the entity level, the relative improvement at the image level is better, which also indirectly confirms that the entity-level multi-image relation association task requires the model to be able to perceive more image details (the relative improvement at the entity level is around 0.1, while that of the image level is around 0.3). At the entity level, while the overall performance on the MentalitySimilarity (MS) is comparable to other subtasks, the improve-

Setting	Model	Overall	Entity Level						Image Level				
			RP	US	MS	SM	AS	NSE	Env	LO	SimE	SubE	OA
IQ	GPT4o	**67.29**	45.68	66.67	65.17	44.34	68.89	**63.49**	88.89	47.78	77.78	**97.00**	70.75
	GPT4v	66.63	38.75	**70.71**	60.67	44.76	**71.11**	51.61	87.77	**64.44**	**78.89**	92.00	66.04
	Gemini-Pro	65.01	**48.15**	67.68	69.66	**47.17**	67.78	56.92	82.22	54.44	60.00	82.00	**73.02**
	Gemini-Flash	60.33	34.56	66.66	70.78	25.47	68.88	53.84	83.33	60.00	48.88	93.00	57.14
	Idefics2	56.93	37.04	65.66	69.66	28.30	44.44	53.97	87.78	36.67	72.22	88.00	45.24
	Mantis-Idefics2	57.59	35.80	62.63	68.54	41.51	52.22	41.27	82.22	20.00	74.44	91.00	56.35
	Phi3v	51.75	48.15	64.65	62.92	47.17	61.11	46.03	86.67	34.44	56.67	51.00	20.63
	Qwen-VL-Chat	47.45	37.04	58.59	68.54	34.91	48.89	41.27	73.33	33.33	61.11	50.00	23.02
	InternVL2-26B	58.78	48.15	64.65	**76.40**	37.73	63.33	57.14	**93.33**	42.22	63.33	52.00	53.17
	InternVL2-2B	47.97	11.90	61.11	67.42	44.44	58.73	46.67	50.00	31.11	59.05	46.67	40.57
	InternVL2-1B	43.71	16.67	62.22	64.04	34.57	42.86	47.78	32.00	30.00	52.38	53.33	34.91
	CLIP	45.05	50.00	50.00	44.94	43.21	30.16	57.78	51.00	45.56	32.32	50.00	40.57
	MetaCLIP	48.37	51.59	68.89	65.17	33.33	31.75	42.22	61.00	28.89	64.65	47.78	36.79
QO	LLaMA-3-8B-Instruct	31.76	34.57	62.63	24.72	34.91	32.22	42.86	28.89	31.11	31.11	6.00	25.40
	LLaMA-3-70B-Instruct	23.66	38.27	60.61	12.36	26.42	6.67	34.92	35.56	31.11	6.67	0.00	14.29
	Qwen1.5-32B-Chat	32.36	39.51	64.65	11.24	40.57	36.67	49.21	33.33	31.11	42.22	0.00	17.46
	Qwen1.5-72B-Chat	37.11	33.33	63.64	51.69	33.96	41.11	34.92	28.89	31.11	50.00	50.00	0.00
	Qwen2-7B-Chat	40.43	43.21	65.66	50.56	30.19	42.22	42.86	35.56	31.11	52.22	50.00	11.91
	Qwen2-72B-Chat	38.97	35.80	64.65	46.07	45.28	46.67	39.68	27.78	31.11	48.89	44.00	7.14
	Yi-1.5-9B-Chat	41.68	44.44	60.61	46.07	43.40	58.89	30.16	26.67	31.11	40.00	50.00	26.98
	Yi-34B-Chat	41.57	34.57	51.52	47.19	37.74	55.56	26.98	25.56	45.56	48.89	49.00	32.54
	Yi-1.5-34B-Chat	26.78	25.93	63.64	39.33	43.40	11.11	36.51	26.67	20.00	5.56	7.00	17.46
	Mantis-Idefics2	32.68	27.16	18.18	50.56	20.75	54.44	23.81	21.11	33.33	48.89	50.00	21.43
	Qwen-VL-chat	40.04	28.40	53.54	55.06	38.68	53.33	26.98	37.78	33.33	54.44	50.00	11.11
	Phi3	42.17	41.98	65.66	44.94	41.51	46.67	38.10	30.00	31.11	48.89	50.00	25.40
	Idefics2	37.44	22.22	61.62	51.69	29.25	42.22	28.57	34.44	31.11	51.11	50.00	13.49
	InternVL2-8B	31.27	25.93	58.59	15.73	35.85	41.10	39.68	31.11	31.11	1.11	50.00	17.46
	InternVL2-26B	35.64	35.80	62.63	19.10	38.68	42.22	38.10	40.00	35.56	6.67	50.00	25.40

Table 1: The main results of current LVMLs and LLMs on our MMRA benchmark. The IQ and QO represent the Image+Question input and Question Only input, respectively.

Setting	Model	Overall	Entity Level						Image Level				
			RP	US	MS	SM	AS	NSE	Env	LO	SimE	SubE	OA
DQ	LLaMA-3-8B-Instruct	53.43	46.91	60.61	57.30	29.25	57.78	57.14	77.78	46.67	62.22	51.00	47.62
	LLaMA-3-70B-Instruct	60.31	40.74	67.68	62.92	37.74	61.11	41.27	88.89	58.89	70.00	73.00	57.14
	Qwen1.5-32B-Chat	58.46	40.74	67.68	59.62	37.74	67.42	53.97	86.67	**66.67**	73.33	52.00	43.65
	Qwen1.5-72B-Chat	60.06	45.68	**69.70**	75.28	41.51	48.89	60.32	84.44	51.11	74.44	56.00	56.35
	Qwen2-7B-Chat	51.98	39.51	64.65	57.99	32.08	61.80	60.32	85.56	32.22	48.89	68.89	30.16
	Qwen2-72B-Chat	**61.53**	**49.38**	66.67	69.66	**47.17**	50.00	**63.49**	92.22	64.44	72.22	51.00	55.56
IDQ	Idefics2	56.35	39.51	63.64	**75.28**	24.53	46.67	57.14	**88.89**	33.33	68.89	**82.00**	45.24
	Qwen-vl-chat	43.76	27.16	51.52	57.30	34.91	44.44	49.21	62.22	30.00	67.78	50.00	17.46
	Phi3v	53.72	43.21	62.63	73.03	41.51	**55.56**	55.56	87.78	40.00	62.22	54.00	26.98
	Mantis-Idefics2	55.93	35.80	62.63	71.91	29.25	48.89	42.86	85.56	21.11	**75.56**	**82.00**	55.56

Table 2: The results of DQ and IDQ setting on our MMRA benchmark.

ment attributed to the inclusion of images is the most significant. This suggests that current LVLMs have a robust capacity to perceive mental states during pre-training. As a result, multi-image LVLMs can effectively harness the information in images to analyze the relation between multiple images in the context of individuals' mental states.

5.3 Impact of Image Descriptions

We use LLaVA-NeXT-100B to obtain the image caption and input it as extra information, and the results are presented in Tab 2. Under the DQ setting, with the combination of descriptions of image pair, all LLMs' performance is highly improved, and the overall result of Qwen2-72B-Chat surpasses Gemini-Flash and is second only to GPT-4v, GPT-4o, and Gemini-Pro. This demonstrates that multi-image understanding capability of LVLMs mainly stems from content that they precept from images.

The key to improving LVLMs' multi-image association ability lies in enhancing the model's fine-grained perception capabilities. As for IDQ setting, after including image descriptions, the performance of LVLMs does not change significantly, proving image descriptions obtained by LLaVA-NeXT-100B overlap with the content perceived by LVLMs themselves. Although the LVLMs still surpass LLMs at the Image Level, they underperform LLMs at the Entity Level, indicating that LVMLs' fine-grained image perception ability is limited.

Different tasks have varying requirements for the visual module of the LVLMs. As for the image level task, the LVLMs' performance is not obviously improved at IDQ setting, while the LLMs' results are close to that of LVLMs with the input of images' descriptions. It demonstrates that the multi-image perception at the image level relies on the visual module of LVLMs. With regard to the tasks at the entity level, in the IDQ setting, the performance of LVLMs varied the most on the MentalitySimilarity (MS) task, even surpassing GPT-4v and GPT-4o. This indicates that entity-level fine-grained tasks require LVLMs to perceive more detailed textual descriptions.

6 Image Order Perception

6.1 Evaluating Image Order Perception

Understanding the sequential order of images is crucial for interpreting the relations between multiple images, which is essential for tackling complex multi-image tasks, such as sorting images. In certain subtasks of the MMRA benchmark, the order of input images can change the answer to the associated questions.

To examine the LVLMs' ability of perceiving images' order, we reverse the input images' order for four specific subtasks: RelativePosition (RP), SimilarMaterial (SM), NearSimilarEntity (NSE), and ObservationAngle (OA), and each subtask has options that are directly related to the images' order. Additionally, we introduce a new option, "All of the above options are incorrect" as the correct choice. Subsequently, we evaluate the performance of LVLMs on these subtasks under both normal and reverse settings, reporting the average performance across both configurations.

Current LVLMs do not have a strong ability to perceive the order of images. As illustrated in Table 3, we present the accuracy of various LVLMs. Idefics2 demonstrates commendable image order perception, achieving an overall score close to 60%. In contrast, most current LVLMs exhibit inadequate image order perception abilities, with overall scores below 35%. This discrepancy suggests that current open-source LVLMs have not adequately addressed image sequence tasks during their pre-training processes.

6.2 Improving LVLMs' Image Order Perception Ability

Training data curation. To improve the capability of LVLMs' order perception ability, we manually curate 1.5 thousand training data for the associated subtasks (i.e., RP, SM, NE, and OA). Specifically, we continually hire 5 postgraduate students to annotate the samplings under the selected subtasks following the criterion described in Sec. 6.1.

Model	Overall	RP	SM	NSE	OA
Idefics2	**54.12**	**65.55**	**53.30**	**68.26**	**29.37**
Mantis	25.22	31.32	20.76	20.64	28.18
Phi3v	36.85	45.07	47.17	38.89	16.27
Qwen-VL	17.35	18.52	17.93	21.43	11.51

Table 3: The results of the Sequence Perception task.

Model	Overall	RP	SM	NSE	OA
Idefics2	54.12	**65.55**	53.30	68.26	29.37
Qwen-VL	17.35	18.52	17.93	21.43	11.51
Ours	**61.01**	63.98	**60.31**	**69.80**	**49.97**

Table 4: Comparing the baseline and our model.

Training method. To enable the LVLMs learning the order of input images, we curate the reverse sample of the collected data. As each sample with two images in the correct order, we reverse the order of the images. Then we change the golden answer to "All of the above options are incorrect" as described in Sec. 6.1. After that, we combine the normal training data and the reverse training data to fine-tune QwenVL.

Result analysis. As shown in Tab.4, our designed training data brings a significant improvement to the QwenVL, even surpassing the Idefics2. Specifically, our model achieves an overall score of 61.01%, with an improvement of 43.66%, surpassing Idefics2 by 6.89%. It demonstrates that the multi-image input method of current LVLMs has the capability to learn to perceive the images' order. However, **the pre-training and SFT phase of LVLMs do not consider the dimension of multiple image orders**.

7 Conclusion

The multi-image perception capabilities of LVLMs are often overlooked. To systematically assess these capabilities, we establish a relational system among images and manually annotate a sophisticated multi-granularity, multi-image relation association benchmark (MMRA). Our evaluation of multi-image LVLMs reveals that they perform poorly on fine-grained (entity-level) and spatial perception subtasks. Compared results of IDQ setting with those of IQ setting, we find that these models lack robust image detail perception abilities.

Limitations

In this work, due to resource constraints, our exploration of improving model performance in this work was conducted with a limited amount of training data (only 1.5k samples), which does not fully exploit the potential of current VLMs.

Ethics Statement

The dataset used in our research is constructed using publicly available data sources, ensuring that there are no privacy concerns or violations. We do not collect any personally identifiable information, and all data used in our research is obtained following legal and ethical standards. In the stage of data annotation, we employed three graduate students experienced in Multimodal Reasoning filed. We paid the graduate students approximately $13 per hour, well above the local average wage, and engaged in constructive discussions if they had concerns about the process.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 432–446. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 698–718. Association for Computational Linguistics.

Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. Comfact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1656–1675. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4024–4034. Association for Computational Linguistics.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "i'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4380–4397. Association for Computational Linguistics.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad S. Khan. 2024. Llava++: Extending visual capabilities with llama-3 and phi-3.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2023. Timechat: A time-sensitive multimodal large language model for long video understanding. *ArXiv*, abs/2312.02051.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Xiangqing Shen, Yurun Song, Siwei Wu, and Rui Xia. 2024. Vcd: Knowledge base guided visual commonsense discovery in images. *arXiv preprint arXiv:2402.17213*.

Xiangqing Shen, Siwei Wu, and Rui Xia. 2023. Dense-atomic: Towards densely-connected ATOMIC with high knowledge coverage and massive multi-hop paths. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13292–13305. Association for Computational Linguistics.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326f.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative multimodal models are in-context learners. *CoRR*, abs/2312.13286.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. Muirbench: A comprehensive benchmark for robust multi-image understanding. *CoRR*, abs/2406.09411.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Hu Xu, Po-Yao Huang, Xiaoqing Ellen Tan, Ching-Feng Yeh, Jacob Kahn, Christine Jou, Gargi Ghosh, Omer Levy, Luke Zettlemoyer, Wen tau Yih, Shang-Wen Li, Saining Xie, and Christoph Feichtenhofer. 2024. Altogether: Image captioning via re-aligning alt-text.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying clip data.

Yuanhan Zhang Yuan Liu, Haodong Duan, Bo Li, Yike Yuan Songyang Zhang, Wangbo Zhao, Jiaqi Wang, Conghui He, Ziwei Liu, and Dahua Lin Kai Chen. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 543–553. Association for Computational Linguistics.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*.

Wenting Zhao, Justin T. Chiu, Claire Cardie, and Alexander M. Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14883–14896. Association for Computational Linguistics.

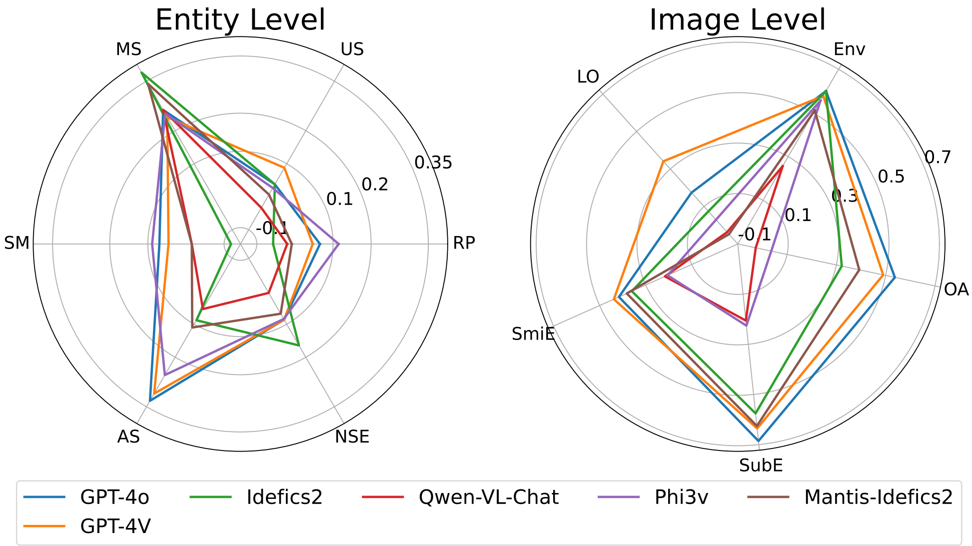

Figure 5: The relative improvement of LVLMs on MMRA benchmark.

Question Type	Prompt Template
T/F Question	You will be giving one question and two images. Please only answer the question with Yes or No. Questions: {question}. Please give me your answer.
Choice Question	You will be giving one question, two images, and four options, one of them is correct. Please choose one of the four options. The question is: {Question}. The options are: [A: {A}, B: {B}, C: {C}, D: {D}] Please tell me the answer in the format if [A], [B], [C] or [D].

Table 5: The designed prompt template for the task in our MMRA benchmark.

A Designed Template

In this part, we present our designed prompt template for both Choice Question and T/F Question in the Tab 5.

B The Details of Retrieval Method

Our approach leverages the strong alignment between text and image representations learned by multimodal retrieval models such as CLIP. Specifically, we compute the embedding of the query and add it to the embeddings of image1 and image2. The resulting representation is then compared with the embeddings of the answer options using a dot product to measure similarity. The option with the highest similarity score is selected as the model's final prediction. We will include a more detailed explanation in the final version of the paper, as one additional page is permitted.

C The Information of Our Baselines.

We present the pre-training information and supporting of our used baselines in Tab 6.

D Result Exact Matching Rule

Due to significant differences in the response styles of various LLMs and chat templates, the content format of model answers can vary greatly. To address this discrepancy and accurately reflect the responses of different models, we have developed a specialized Exact Matching Rule.
For Multiple-Choice questions: First, we use regular expressions to attempt to directly extract the matching content within parentheses, i.e., extracting Answer: "A" from "(A)". If this is unsuccessful, we then attempt to match option labels (A-D) from the entire response content and return the option with the highest match count. If the response does not contain any option label information, we try to match the option content directly within the response and return the corresponding option label.
For True/False questions: We use regular expressions to match "yes" or "no" within the response content. If there are multiple matches, we return the result that appears the most frequently.

E Sampled examples from MMRA benchmark

In order to comprehensively show our benchmark, we select a sample for each task and present then in the Figure 6. We design two kinds of tasks (i.e., Choice Question and T/F Question). For each

Model	Pre-training Data	Supporting Input	Parameters
GPT4o&GPT4v	/	Text, Multi Images, Audio	/
Gemini-Flash	/	Text, Multi Images, Audio, Video	/
Idefics2	Internet Crawled Data (Wikipedia and OBELICS), Public Multimodal Dataset, LAION-COCO, PDFA (en), IDL, Rendered-text, WebSight	Text, Multi Images	8B
Qwen-VL-Chat	LAION-en, LAION-zh, In-house Data, LAION-COCO, DataComp, Coyo, CC12M, CC3M, SBU, COCO Caption	Text, Multi Images	8B
Phi3v	/	Text, Multi Images	26B
InternVL2	/	Text, Multi Images, Video	8B
Mantis-Idefics2	Mantis-Instruction dataset	Text, Multi Images	8B
LLaMA-3	/	Text Only	8B, 70B
Qwen1.5&Qwen2	Internet Crawled Data	Text Only	7B, 32B, 72B
Yi-Chat&Yi-1.5-Chat	Web Documents from Common Crawl	Text Only	9B, 43B

Table 6: The pre-training information and supporting input of the baselines. "_" refers to non-public or not fully public data.

example, we show the image pair, question and options.

F Relative Improvement of LVLMs

We present the relative improvement of LVLMs between the IQ and QO settings.

G Error analysis

To better analyze the shortcomings of LVLMs, we examined instances where GPT-4o made errors on relatively challenging subtasks such as RelativePosition, MadeOf, NearSameEntity, and Layout.

As presented in Fig 7, LVLMs often select entities that do not appear in the image when answering fine-grained questions. For example, for subtasks like 'RelativePosition' and 'NearSameEntity', LVLMs sometimes choose options featuring entities that are not present in the image (e.g., beer and tray).

We believe this issue arises because LVLMs primarily depend on the reasoning capabilities of the language model. The textual relations in the options can significantly interfere with the LVLMs' judgments, leading them to overlook the visual input, particularly for fine-detailed questions.

In scenarios where neither image contains the correct answer for the subtask, we introduced an alternative option to express there is no association between the two images, such as 'there are no entities of the same material in fig1 and fig2'. When LVLMs cannot identify the correct answer, they tend to select this option, suggesting no connection between the two images.

Regarding the 'Layout' subtask, it appears that current LVLMs have a limited ability to grasp the key elements within images. They sometimes fail to determine whether both images prominently feature a main entity.

Figure 6: Sampled MMRA examples for each task. The bold and underlined options indicate they are the golden answers.

RelativePosition

Question: Which two entities in Fig1 and Fig2 are in the same relative position within the images?
QA_type: Choice QA

Options:
A. shutter in figure one and window in figure two
B. hinge in figure one and baby bird in figure two
C. doorframe in figure one and the marks left by a impact in figure two
D. doorframe in figure one and string in figure two

Golden answer: C **GPT4O's answer: D**

SimilarMaterial

Question: Which two entities, respectively in Fig1 and Fig2, are made of the same material?
QA_type: Choice QA

Options:
A. doorknob in fig1 and microwave door frame in fig2
B. the surf in fig1 and the bus in fig2
C. there are no entities of the same material in fig1 and fig2
D. the surf in fig1 and the road surface in fig2

Golden answer: C **GPT4O's answer: D**

NearSameEntity

Question: Which two entities, respectively in Fig1 and Fig2, near or adjacent to a same object?
QA_type: Choice QA

Options:
A. spoon in figure one and folk in figure two
B. wine in figure one and cup in figure two
C. beer cap in figure one and tray in figure two
D. beer in figure one and tray in figure two

Golden answer: C **GPT4O's answer: D**

Layout

Question: What are the similarities between these two pictures in terms of structure and layout?
QA_type: Choice QA

Options:
A. the distribution of entities in the pictures follows a similar pattern or arrangement
B. there is no obvious relation between the pictures in terms of layout.
C. each picture has a prominent entity

Golden answer: C **GPT4O's answer: A**

Figure 7: The error analysis of GPT4o on our MMRA benchmark.