CODIS: Benchmarking Context-Dependent Visual Comprehension for Multimodal Large Language Models

Fuwen Luo[1*], Chi Chen[1*], Zihao Wan[1], Zhaolu Kang[6], Qidong Yan[5], Yingjie Li[5], Xiaolong Wang[1], Siyu Wang[2], Ziyue Wang[1], Xiaoyue Mi[7], Peng Li[2,3†], Ning Ma[5], Maosong Sun[1†], Yang Liu[1,2,3,4]

[1]Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
[2]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China,
[4] Jiangsu Collaborative Innovation Center for Language Competence, Jiangsu, China
[5]Key Laboratory of Linguistic and Cultural Computing Ministry of Education, Northwest Minzu University, China
[6]College of Software, Jilin University, China
[7]Institute of Computing Technology, Chinese Academy of Sciences
{lfw23,chenchi19}@mails.tsinghua.edu.cn
lipeng@air.tsinghua.edu.cn, sms@tsinghua.edu.cn

Abstract

Multimodal large language models (MLLMs) have demonstrated promising results in a variety of tasks that combine vision and language. As these models become more integral to research and applications, conducting comprehensive evaluations of their capabilities has grown increasingly important. However, most existing benchmarks fail to consider that, in certain situations, images need to be interpreted within a broader context. In this work, we introduce a new benchmark, named as CODIS, designed to assess the ability of models to use context provided in free-form text to enhance visual comprehension. Our findings indicate that MLLMs consistently fall short of human performance on this benchmark. Further analysis confirms that these models struggle to effectively extract and utilize contextual information to improve their understanding of images. This underscores the pressing need to enhance the ability of MLLMs to comprehend visuals in a context-dependent manner. View our project website at https://thunlp-mt.github.io/CODIS.

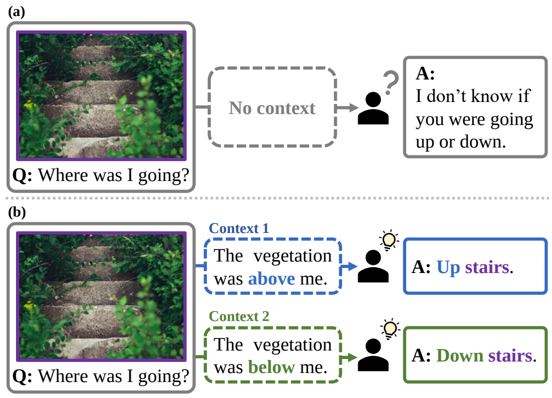


Figure 1: Interpretation of images can be significantly influenced by contextual information. In this instance, the determination of whether the photographer was ascending or descending a staircase remains ambiguous without supplementary context (a). However, when additional information is provided, indicating the position of the greenery relative to the observer, the direction of movement of the observer becomes clear (b). In the responses, words originating from the image and the two pieces of context are highlighted in purple, blue, and green, respectively.

1 Introduction

Recent years have witnessed a rapid advancement in multimodal large language models (MLLMs, OpenAI 2023; Gemini Team et al. 2023; Liu et al. 2023a; Dai et al. 2023). They have achieved remarkable results on various downstream tasks, such as image captioning (Luo et al., 2023; Wang et al., 2023; Chen et al., 2015), visual question answering (Shao et al., 2023; Liu et al., 2023c; Antol et al., 2015) and visual reasoning (Gupta and Kembhavi, 2023; Chen et al., 2023c; Zellers et al., 2019). As the performance of MLLMs continues to improve, the comprehensive assessment of their capabilities becomes increasingly important.

Previous research highlights a fascinating and vital human ability: understanding visual elements within broader contexts (Baldassano et al., 2016). Consider the example in Figure 1, where simply looking at an image does not reveal whether the stairs go up or down. However, introducing the context of whether the greenery is above or below

*Equal Contribution.
†Corresponding Authors.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10639–10659
August 11-16, 2024

Benchmark	Purpose	Ambiguity	Context	Answer Type	Evaluator
MME (Fu et al., 2023a)	Comprehensive Evaluation	✗	✗	Yes/No	Metrics
MMBench (Liu et al., 2023d)	Comprehensive Evaluation	✗	✗	Multi-choice	GPT
SEED-Bench (Li et al., 2023b)	Comprehensive Evaluation	✗	✗	Multi-choice	Metrics
VisDial (Das et al., 2017)	Visual Dialog	✗	✓	Open-ended	Metrics
MMDialog (Feng et al., 2022)	Visual Dialog	✗	✓	Open-ended	Metrics
HallusionBench (Guan et al., 2023)	Visual Hallucination	✓	✗	Yes/No	Metrics
Bingo (Cui et al., 2023)	Visual Hallucination	✓	✗	Open-ended	Human
SQUID-E (Sanders et al., 2022)	Visual Ambiguity	✓	✗	Multi-choice	Metrics
DCIC (Schmarje et al., 2022)	Visual Ambiguity	✓	✗	Multi-choice	Metrics
CODIS (Ours)	Image-context Comprehension	✓	✓	Close & Open-ended	Human / GPT

Table 1: Comparison of our proposed CODIS with recent vision-language benchmarks.

the viewpoint of the photographer enables humans to clarify the ambiguity and accurately interpret the image. In this study, we argue that MLLMs should also be evaluated from the perspective of context-dependent visual comprehension. From a scientific perspective, it is important to understand the extent to which MLLMs can leverage context to enhance their visual comprehension. Practically speaking, as MLLMs are increasingly applied in real-world scenarios, assessing their reliability in interpreting visual data within context is essential for ensuring they provide accurate responses.

However, the capability for context-dependent visual understanding has not been fully assessed by current benchmarks for MLLMs. Table 1 summarizes recent benchmarks for MLLMs. Most of these benchmarks (Fu et al., 2023a; Liu et al., 2023d; Li et al., 2023b; Guan et al., 2023; Cui et al., 2023; Sanders et al., 2022; Schmarje et al., 2022) do not pair images with additional context. Only two benchmarks, namely VisDial (Das et al., 2017) and MMDialog (Feng et al., 2022), include extra context to help in conversation with humans rather than to clarify the meaning of images as shown in Figure 1. This limitation means these benchmarks are not fully capable of testing the ability of MLLMs to understand images in a context-dependent manner.

To address this challenge, we introduce a new benchmark, named as CODIS, designed to evaluate the capability of MLLMs in **CO**ntext-**D**ependent **I**mage di**S**ambiguation. CODIS utilizes the visual question answering (VQA, Antol et al. 2015) format for this purpose, as shown in Figure 1(b). It stands out from existing benchmarks in three main aspects: first, each image in CODIS contains inherent ambiguity that can only be resolved with additional context; second, the questions are deliberately designed to highlight these ambiguities, requiring external context for accurate interpretation; third, for every image-question pair, we provide two contexts in a free-form text format. These contexts are subtly different yet lead to different interpretations of the image and, consequently, different answers. We have carefully curated all images, questions, and contexts by hand to maintain a high standard of quality and diversity. Our assessment of 14 widely-used MLLMs with CODIS indicates that the performance of these models in understanding context-dependent visuals significantly falls short of human capabilities. Further analysis demonstrates that the models struggle with identifying crucial contextual cues and extracting relevant visual features.

To summarize, our contributions are three-fold:

- We highlight the significance of context-dependent visual comprehension abilities for MLLMs.
- We introduce the CODIS benchmark to assess capabilities of MLLMs on context-dependent visual comprehension.
- Through our analysis, we uncover the deficiencies in MLLMs regarding context information extraction and visual information extraction, underscoring the immense potential for enhancement in the realm of context-dependent visual comprehension.

2 Related Work

2.1 Context in Visual Tasks

Previous works emphasize the importance of context in visual tasks (Wang and Zhu, 2023; Vo, 2021; Goh et al., 2004; Bar and Aminoff, 2003), exploring utilizing context in scene graph generation (Cong et al., 2023; Zheng et al., 2023; Zhu et al., 2022; Yang et al., 2018), object detection (Zou et al., 2023; Diwan et al., 2023; Xiao et al., 2023; Chen et al., 2023b), image inpainting (Zhang et al., 2023; Sargsyan et al., 2023; Xiang et al., 2023), etc. However, they mainly focus on visual context within images, e.g., relationships between objects, environment and background. External contexts, such as free-form texts, have also been studied. Previous works have shown their effectiveness in tasks such as floorplan reconstruction (Purushwalkam et al., 2021), and visual dialog (Das et al., 2017; Feng et al., 2022). The ability of multimodal in-context learning (Li et al., 2023a; Zhao et al., 2023; Sun et al., 2023; Li et al., 2023c) further demonstrates the significance of external contexts. However, in these works, context do not affect the interpretation of images, which is the main focus of our work. We strengthen the impact of context on visual comprehension.

2.2 Visual Ambiguities

Visual ambiguities arises from absence of visual information or other noises (Denison et al., 2018). It is an important phenomenon in human visual cognition. Previous works investigate visual ambiguities caused by optical illusions (Guan et al., 2023; Cui et al., 2023; Fu et al., 2023b). Images which may cause illusions are leveraged to evaluate visual hallucinations in MLLMs. Other works explore ambiguities in visual event classification (Sanders et al., 2022; Schmarje et al., 2022; Rajeswar et al., 2022), focusing on model robustness against ambiguous images. Different from these works, we assess capability of MLLMs to disambiguate images instead of just recognizing ambiguities.

3 CODIS

CODIS is proposed for evaluating the capability of MLLMs in context-dependent image disambiguation. Figure 2 presents several examples from our benchmark, highlighting the diversity of contexts covered. In this section, we first elaborate our taxonomy of context. Then, we delve into the instruction design and evaluation method of CODIS. Finally, we introduce procedures of data collection.

3.1 Taxonomy of Context

Given the extensive and varied nature of context information, it is challenging to catalog all forms of context comprehensively. Nevertheless, in our effort to establish a pioneering benchmark for context-dependent image disambiguation, we aim to cover a wide range of diverse and illustrative scenarios. Drawing inspiration from the types of information people require to understand the visual contents of an image, we have identified five representative types of context. These include three types of global context that pertain to the overall scene—namely, the global background, which encompasses location and orientation, temporal information, and cultural background. Additionally, we focus on two types of local context related to objects within the scene, specifically the attributes of objects and the relationships between people. We present examples in Figure 2 and provide an explanation of our classification as follows.

Location and orientation. Interpretations of scenes is closely related to where they happen. As Figure 2(a) shows, the scene of falling leaves indicates different times in a year when they happen in different countries. Orientation also plays an important role in image understanding. In Figure 2(b), direction which the photographer is facing can help determine direction in which the herd is moving.

Temporal information. Understanding the timing and sequence of events is crucial when we understand a scene. However, an image can only provide us with static information. Thus, context of temporal information is often crucial, and can help better restore the original appearance of events. Figure 2(c) shows an example of context which indicates dynamic changes of things and events, like opening or closing the door, while context of Figure 2(d) tells when pictures were taken. [1]

Cultural background. Cultural background has a profound impact on the way people think and act. Meaning of an action may have significant differences across different nations and cultures sometimes. Awareness of cultural background is crucial for understanding images. We employ cases covering various aspects of differences between cultures, such as laws, regulations, customs and traditions. See Figure 2(e) and Figure 2(f) for examples.

Attributes. To comprehensively understand objects in a scene, it is not enough to solely rely on visual appearances we see. Non-visible attributes

[1] One might argue that introducing video is the most direct method to resolve temporal ambiguities. However, videos are not always available in practice. By analogy to humans, a sufficiently capable MLLM should also be able to resolve such ambiguities without relying on videos. Therefore, we believe it is necessary to include this category of contexts in our benchmark.

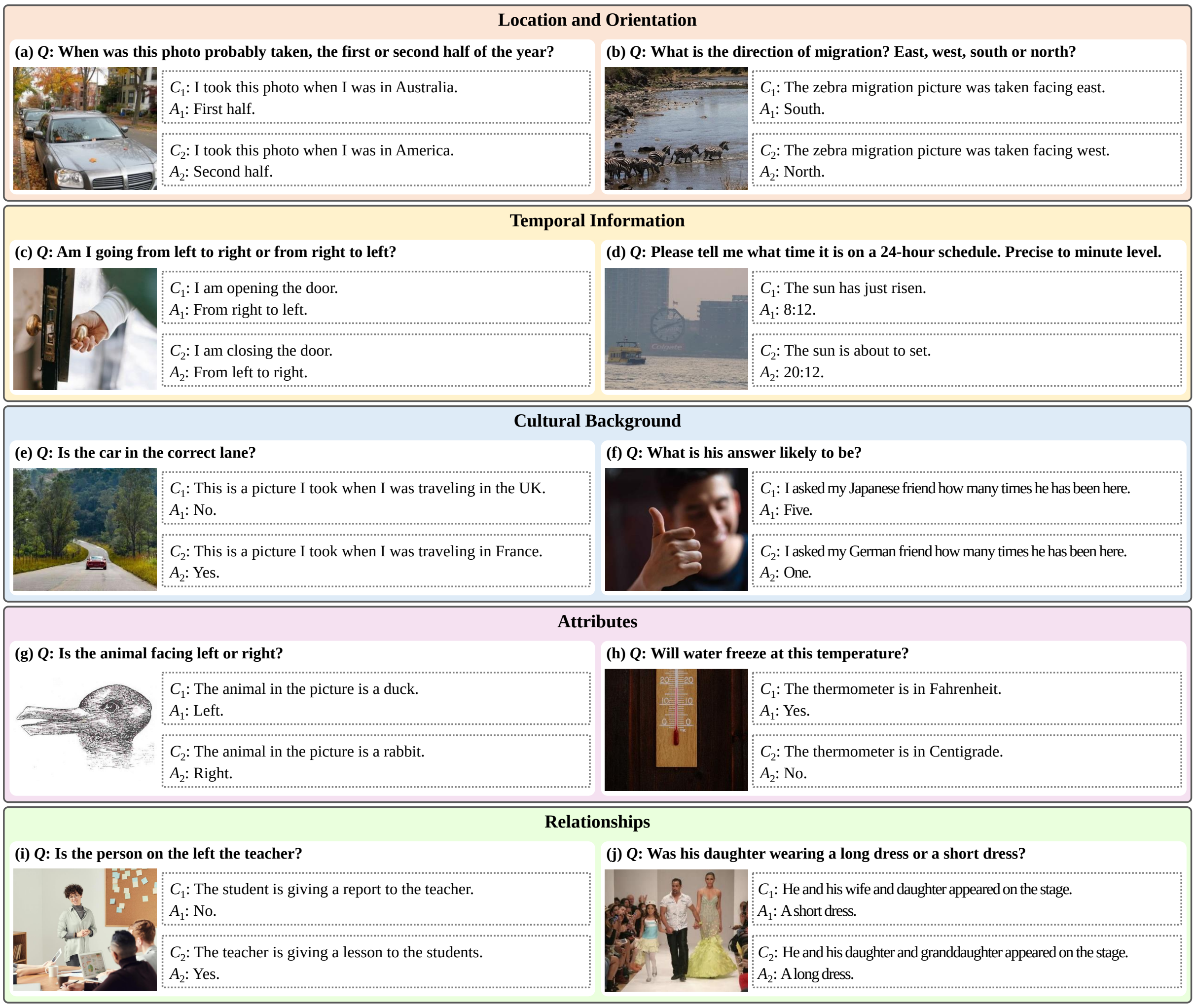


Figure 2: Taxonomy of our benchmark. We show two cases for each category. In each case, we show an image and a question, along with two pieces of different context and their corresponding answers. We use Q to denote questions, C for contexts and A for answers.

are also important. We consider two types of attributes, including *physical attributes*, such as structure, size, temperature, smell and touch, as Figure 2(g) shows, and *abstract attributes*, such as function and mechanism, as Figure 2(h) shows.

Relationships. We communicate with people in different ways based on our relationships with them. When the same event happens on people with different relationships, its meaning varies. It is important to understand relationships between people in visual comprehension. As Figure 2(i) and Figure 2(j) shows, relationships between people affect our judgement on their roles and identities.

3.2 Instruction Design

To prevent models from guessing the correct answers without fully understanding context, we organize our dataset in pairs. Each pair contains two queries (I, Q, C_1) and (I, Q, C_2). The queries have identical image I and question Q, but have two pieces of different context C_1 and C_2. We give MLLMs two queries separately, and get model outputs O_1 and O_2. Take Figure 2(a) as an example. In the first query, we provide a picture of falling leaves and indicate that the photo was taken in Australia. In the second query, we give different context that the photo was taken in America. We expect to get different outputs from models that the photo was taken in the first or the second half of the year.

3.3 Evaluation Method

We use O_1 and O_2 to denote model outputs of a pair of queries, and A_1 and A_2 to denote the groundtruth answers. The evaluation of model outputs can be represented as follows:

$$\text{Eval}(O_i) = \begin{cases} 1 & \text{if } O_i \text{ matches } A_i \\ 0 & \text{otherwise} \end{cases}, i = 1, 2.$$

Model	Parameters	Image Encoder	LLM	Alignment Module
GPT-4V (OpenAI, 2023)	-	-	-	-
Gemini (Gemini Team et al., 2023)		-	-	-
LLaVA-1.5-13B (Liu et al., 2023a)	~13B	$\text{CLIP}_{\text{ViT-L}}$-336px	LLaMA-2-13B	MLP
BLIP-2-11B (Li et al., 2023d)		$\text{EVA-CLIP}_{\text{ViT-G}}$	Flan-T5-xxl	Q-Former
InstructBLIP-13B (Dai et al., 2023)		$\text{EVA-CLIP}_{\text{ViT-G}}$	Vicuna-13B	Q-Former
mPLUG-Owl-2 (Ye et al., 2023)	~7B	$\text{CLIP}_{\text{ViT-L}}$	LLaMA-2-7B	Attention
MiniGPT4-7B (Zhu et al., 2023)		$\text{EVA-CLIP}_{\text{ViT-G}}$	Vicuna-7B	Linear
LLaVA-1.5-7B (Liu et al., 2023a)		$\text{CLIP}_{\text{ViT-L}}$-336px	LLaMA-2-7B	MLP
InstructBLIP-7B (Dai et al., 2023)		$\text{EVA-CLIP}_{\text{ViT-G}}$	Vicuna-7B	Q-Former
Otter-7B (Li et al., 2023a)		$\text{CLIP}_{\text{ViT-L}}$	MPT-7B	Attention
LLaVA-7B (Liu et al., 2023b)		$\text{CLIP}_{\text{ViT-L}}$	Vicuna-7B	Linear
Qwen-VL-Chat (Bai et al., 2023)		$\text{OpenCLIP}_{\text{ViT-bigG}}$	Qwen-7B	Attention
OpenFlamingo-7B (Awadalla et al., 2023)		$\text{CLIP}_{\text{ViT-L}}$	MPT-7B	Attention
BLIP-2-6.7B (Li et al., 2023d)		$\text{EVA-CLIP}_{\text{ViT-G}}$	OPT-6.7B	Q-Former

Table 2: API-based and open-source MLLMs selected for evaluation.

Following Fu et al. (2023a), we leverage two metrics, pair-wise accuracy Acc_p and query-wise accuracy Acc_q for our evaluation metrics, which are calculated as follows:

$$\text{Acc}_p = \frac{1}{n_p}\sum_{i=1}^{n_p}(\text{Eval}(O_{i1}) \times \text{Eval}(O_{i2})),$$

$$\text{Acc}_q = \frac{1}{2n_p}\sum_{i=1}^{n_p}(\text{Eval}(O_{i1}) + \text{Eval}(O_{i2})),$$

where n_p is number of data pairs. For Acc_p, models score only if their answers to a pair of queries are both correct. For Acc_q, models score for each single query they answer correctly.

3.4 Data Collection

In this section, we introduce the process of constructing our benchmark. Our data collection process includes three steps.

Image collection. We manually collect images that contain ambiguities which can only be resolved with external contexts. The majority of these images are real-scene images from the publicly available dataset ShareGPT4V (Chen et al., 2023a) and the Internet, while the remainder are created manually. The primary challenge in collecting images was that many appeared ambiguous at first glance but were actually unambiguous upon closer examination. Therefore, each image had to be carefully inspected and excluded from the collection if it was found to be unambiguous. In total, we collected 377 images.

Design of questions, context and answers. For each collected image, we manually write questions, context and answers for it. The data are compiled with following rules:

(1) Questions are targeted at the ambiguous parts of the images. They can not be answered if no additional context is provided due to ambiguities in the images.

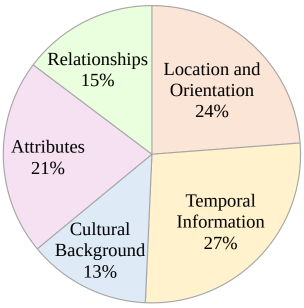
(2) For each image-question pair, two unique contexts with minor differences that lead to different interpretations of the image need to be created. As a result, these interpretations yield different answers to the question. Furthermore, we require that an answer can only be derived when both the image and its corresponding context are available. This means that neither the image nor the context alone is sufficient to provide an answer. For instance, in Figure 2(a), the photograph of falling leaves, when combined with the knowledge that the photograph was taken in Australia, allows us to conclude that it depicts autumn in the southern hemisphere, occurring in the first half of the year. In isolation, neither the image nor the context is adequate to reach this conclusion.

(3) To enhance the evaluation process, MLLMs are permitted to respond in free-form text. However, to maintain a balance between the ease of evaluation and ensuring objectivity, we require the answers to be relatively objective. For example, despite the lack of a strict format for the responses to the scenarios shown in Figure 2, we provide clear options for models to choose in most of the cases. This approach guarantees that the responses stay objective and are easy to evaluate.

Data verification. Five annotators participated in the data collection process. To ensure the quality of our dataset, each submission by an annotator was cross-checked by the other four. Data was retained only if it (1) was correct, (2) differed from existing data to a significant extent, thereby main-

Model	Loc & Ori Acc$_p$	Loc & Ori Acc$_q$	Temporal Acc$_p$	Temporal Acc$_q$	Cultural Acc$_p$	Cultural Acc$_q$	Attributes Acc$_p$	Attributes Acc$_q$	Relationships Acc$_p$	Relationships Acc$_q$	Overall Acc$_p$	Overall Acc$_q$	Context Awareness
Human													
Human	85.2	86.1	90.9	92.8	72.8	76.4	87.2	88.4	89.6	90.0	86.2	87.7	97.3
API-based Models													
GPT-4V	33.3	54.2	28.4	52.1	25.5	60.6	26.7	54.7	51.9	70.2	32.3	56.9	54.7
Gemini	21.4	49.4	29.5	51.1	21.3	56.4	24.0	52.0	34.6	58.7	26.1	52.7	43.6
Open-source ~13B Models													
LLaVA-1.5-13B	**6.0**	**41.1**	4.2	44.7	**10.6**	**50.0**	**14.7**	**51.3**	**13.5**	**54.8**	**9.1**	**47.5**	19.3
BLIP-2-11B	**6.0**	32.7	**8.4**	**45.8**	4.3	35.1	6.7	42.0	11.5	51.9	7.4	41.4	31.4
InstructBLIP-13B	**6.0**	39.3	2.1	41.6	4.3	**50.0**	4.0	44.7	7.7	51.0	4.5	44.2	14.1
Open-source ~7B Models													
mPLUG-Owl-2	**13.1**	42.3	**9.5**	41.6	**6.4**	42.6	**12.0**	44.7	**19.2**	51.9	**11.9**	44.1	31.7
MiniGPT4-7B	10.7	36.3	3.2	34.2	0.0	27.7	**12.0**	35.3	13.5	47.1	7.9	36.0	36.3
LLaVA-1.5-7B	11.9	**42.9**	5.3	44.7	4.3	43.6	9.3	39.3	7.7	47.1	7.9	43.3	21.5
InstructBLIP-7B	1.2	33.3	7.4	**45.8**	0.0	46.8	4.0	43.3	11.5	48.1	4.8	42.8	16.7
Otter-7B	2.4	32.7	5.3	41.1	4.3	28.7	0.0	26.0	5.8	40.4	3.4	34.1	19.3
LLaVA-7B	2.4	30.4	6.3	34.2	0.0	25.5	1.3	34.0	5.8	41.3	3.4	33.1	17.2
Qwen-VL-Chat	3.6	23.8	3.2	24.7	0.0	24.5	1.3	32.0	9.6	34.6	3.4	27.5	26.3
OpenFlamingo-7B	2.4	40.5	2.1	38.9	0.0	27.7	5.3	36.0	5.8	47.1	3.1	38.4	15.6
BLIP-2-6.7B	0.0	41.1	1.1	44.7	2.1	**48.9**	2.7	**46.0**	7.7	**53.8**	2.3	**46.0**	6.5

Table 3: Results of MLLMs on CODIS. All the model outputs are assessed by human. Consistency of human evaluation is investigated in Appendix A.



Scenario	Percentage
Actions	20%
Landscape	12%
Streetscape	11%
Indoors	11%
Human	10%
Transportation	9%
Sports	8%
Artificial	8%
Animals	7%
Others	4%

Figure 3: Distribution of five categories (left) and scenarios (right) of our CODIS benchmark.

taining high diversity, and (3) conformed to all the above criteria. If these conditions were not met, the annotator had to revise the data.

Ultimately, we retained 216 images and successfully collected a total of 706 queries, spanning five categories and encompassing a wide range of scenarios. We had to exclude 161 images because we were unable to collect eligibility questions or contexts for them. The distribution of categories and scenarios is illustrated in Figure 3.

4 Evaluation

4.1 Models

As shown in Table 2, we perform evaluation on 14 popular MLLMs, which are divided into three groups: (1) **API-based models:** GPT-4V (OpenAI, 2023) and Gemini (Gemini Team et al., 2023); (2) **Open-source ~7B models:** mPLUG-Owl-2 (Ye et al., 2023), MiniGPT-4-7B (Zhu et al., 2023), LLaVA-1.5-7B (Liu et al., 2023a), InstructBLIP-7B (Dai et al., 2023), Otter-7B (Li et al., 2023a), LLaVA-7B (Liu et al., 2023b), Qwen-VL (Bai et al., 2023), OpenFlamingo (Awadalla et al., 2023) and BLIP-2-6.7B (Li et al., 2023d); (3) **Open-source ~13B models:** LLaVA-1.5-13B, BLIP-2-11B and InstructBLIP-13B.

4.2 Results

We manually assess the outputs of MLLMs on CODIS benchmark and report the results in Table 3. "Human" refers to the average performance of five independent people, and the detailed results and analysis are shown in Appendix B.

Overall results. From an overall perspective, human is far ahead of all the MLLMs in all the categories, indicating that MLLMs have tremendous potential of improvement on context-dependent visual comprehension. API-based models significantly outperform open-source models, and GPT-4V gets the highest score among all the MLLMs. In terms of open-source models, mPLUG-Owl-2 has the best performance on Acc_p and LLaVA-1.5-13B has the best performance on Acc_q.

Disparity between Acc_p and Acc_q. We observe a large disparity between Acc_p and Acc_q of MLLMs, and open-source models have a larger disparity than API-based models. To better understand the reason of this phenomenon, we further investigate the metric of context awareness, which reflects model capability of recognizing different context and provide different responses accordingly. It is

Model	Output Variability
Human	81.9
GPT-4V	58.6
Gemini	53.5
mPLUG-Owl-2	36.8
LLaVA-1.5-13B	25.6

Table 4: Variations in model outputs when context is removed. See the calculation of output variability in Section 4.2.

Model	w/ Images w/o Captions	w/ Captions w/o Images
GPT-4V	35.1	51.0
Gemini	31.1	35.8
mPLUG-Owl-2	10.6	16.6
LLaVA-1.5-13B	11.9	20.5

Table 5: Variations in Acc_p when we replace the images with their captions.

calculated as follows:

$$\mathrm{ContextAwareness} = \frac{1}{n_p} \sum_{i=1}^{n_p} (O_{i1} \neq O_{i2}),$$

where n_p is number of data pairs. **We manually assess inequality of model outputs on semantic level.** The results are shown in the last column of Table 3. The scores of all the MLLMs are lower than 60, while human score is 97.3, indicating that for many data pairs, MLLMs fail to recognize the difference in context and they provide semantically equal response to the two queries. Thus, for a significant number of cases, models only correctly answer one of the two queries, leading to high Acc_q and low Acc_p scores. Additionally, context awareness is positively correlated with Acc_p, and has no clear correlation with Acc_q. Therefore, Acc_p is more possible to reflect model capability of context comprehension, and is more suitable as the primary metric for CODIS benchmark. We further investigate the capability of MLLMs on extracting information from context. By excluding context and providing only images and questions as inputs, we observe the metric of output variability, which reflects the variations in model outputs. It is calculated as follows:

$$\mathrm{OutputVariability} = \frac{1}{n_q} \sum_{i=1}^{n_q} (O_{nc}^i \neq O_c^i),$$

where n_q is the total number of queries, O_{nc} and O_c are outputs when context is given and is not given. **Inequality of model outputs is assessed manually on semantic level.** We conduct our experiment on API-based models GPT-4V and Gemini, and mPLUG-Owl-2 and LLaVA-1.5-13B, which have the highest Acc_p and Acc_q among open-source models, respectively. Additionally, we report human results for reference. As shown in Table 4, MLLMs score significantly lower than human, which means they are incapable to change their answers according to context they received. Their capability of context information extraction is needed to be improved. Examples of model responses when removing context are shown in Appendix C.

Comparison between ~7B and ~13B models. We observe from Table 3 that scaling up from ~7B to ~13B does not necessarily lead to an improvement in model performance. For LLaVA-1.5, its ~13B model significantly outperforms ~7B model on both Acc_p and Acc_q. However, for BLIP series models, Acc_p of InstructBLIP and Acc_q of BLIP-2 slightly drop when they scale up. Regrettably, significant differences exist between the two models, covering various aspects including data, training paradigms, and additional details. Consequently, we were unable to determine the primary causes of the differing behaviors observed in the two models. While intensive efforts are warranted to explore these causes, we believe that such an investigation exceeds the scope of this work.

Model performance on different categories. Among the five categories, most MLLMs perform best on relationships. For open-source MLLMs, it is possibly because their training data include dense image annotations (Liu et al., 2023b,a; Ye et al., 2023) such as Visual Genome (Krishna et al., 2017), or detailed captions (Li et al., 2023d; Dai et al., 2023; Zhu et al., 2023) which contain lots of data of relationships. We also observe that the score of MLLMs on cultural background is relatively low, indicating they are not sufficiently trained on culture-related data.

5 Analyses

5.1 Visual Information Extraction

Extracting information from images according to context is the foundation of context-dependent visual comprehension task. We explore capability of MLLMs on extracting visual information. We replace images with detailed captions and observe the changes in Acc_p. For some of the pictures

Model	up: down	morning: afternoon	appropriate: not appropriate
Groundtruth	15:15	16:16	13:13
GPT-4V	14:12	7:12	16:9
Gemini	18:11	5:21	22:2
mPLUG-Owl-2	23:6	9:15	23:1
LLaVA-1.5-13B	21:8	19:7	22:4

Table 6: Bias in model outputs. We report frequencies of three pairs of specific words. Some outputs do not contain either of the two words, which makes total word frequencies in model outputs less than groundtruth.

Model	Acc_p Human Eval	GPT-4 Eval	Agreement
GPT-4V	32.3	31.2	97.6
Gemini	26.1	24.4	97.7
LLaVA-1.5-13B	9.1	8.5	97.6
BLIP-2-11B	7.4	7.9	94.8
InstructBLIP-13B	4.5	4.5	98.6
mPLUG-Owl-2	11.9	9.9	96.7
MiniGPT4-7B	7.9	6.8	91.8
LLaVA-1.5-7B	7.9	5.7	94.3
InstructBLIP-7B	4.8	4.2	98.7
Otter-7B	3.4	2.5	94.3
LLaVA-7B	3.4	2.5	92.8
Qwen-VL-Chat	3.4	3.4	97.7
OpenFlamingo-7B	3.1	3.1	100.0
BLIP-2-6.7B	2.3	2.5	98.1

Table 7: Comparison of human and GPT-4 evaluators. We report Acc_p based on human and GPT-4 evaluation, and agreement between human and GPT-4 evaluation.

are not easy to be described in natural language, we discard these images and only perform experiments on a subset of 96 images, 302 queries. We ensure that captions and context are sufficient for answering the questions. We conduct our experiment on GPT-4V, Gemini, mPLUG-Owl-2 and LLaVA-1.5-13B. The model outputs are assessed manually and results are shown in Table 5. For all the MLLMs, Acc_p increases when the images are replaced with captions, indicating that compared to images, MLLMs are more skilled at extracting information from text with the help of context. Model capability to extract visual information according to context is needed to be improved. Examples of model responses when replacing images with captions are shown in Appendix D.

5.2 Bias in Model Outputs

Bias significantly influences model performance. We explore biases in model outputs on three dimensions: location & orientation, timing and culture. We evaluate bias by calculating frequencies of three pairs of specific words, *up* and *down*, *morning* and *afternoon*, and *appropriate* and *not appropriate*, which are associated with the three dimensions, respectively. We evaluated the outputs of GPT-4V, Gemini, mPLUG-Owl-2, and LLaVA-1.5-13B, on a set of 88 queries whose groundtruth answers include these words. As shown in Table 6, although frequencies of the two words in each pair are equal in groundtruth answers, all the MLLMs demonstrate biases in their responses. This indicates that models tend to rely on their biases rather than the provided context in a certain number of cases, contributing to their poor performance on CODIS benchmark, especially Acc_p scores. Reducing bias is necessary for improving model capability on context-dependent visual comprehension. It also suggests that our design of data pairs and evaluation metric of Acc_p effectively prevent models from scoring high based on their biases rather than a true understanding of contexts.

5.3 Comparing Human and GPT-4 Evaluators

We explore the possibility of utilizing GPT-4 as automatic evaluator. We compare the results of human and GPT-4 evaluation, as Table 7 shows. The first two columns show Acc_p of models based on human and GPT-4 evaluation, and the third column shows agreement between human and GPT-4, which is calculated as follows:

$$\text{Agreement} = \frac{1}{n_q}\sum_{i=1}^{n_q}(\text{Eval}_\text{h}(O_i) = \text{Eval}_\text{g}(O_i)),$$

where n_q is the total number of queries, $\text{Eval}_\text{h}(\text{O}_\text{i})$ and $\text{Eval}_\text{g}(\text{O}_\text{i})$ are evaluation results of human and GPT-4 on model output O_i. From the table we can see that GPT-4 is consistent with human on evaluation of more than 90% of queries. Overall, there is a high degree of consistency between human and GPT-4 evaluators in assessing model performance. GPT-4 is capable of conducting evaluation on CODIS benchmark instead of human. We report detailed results of GPT-4 evaluation in Appendix E.

6 Conclusion

In this work, we introduce a new benchmark CODIS to evaluate capability of models to use context to enhance visual comprehension. Results suggest that MLLMs significantly fall behind human in context-dependent visual comprehension. Further analysis suggests that shortcomings in visual information extraction and bias in model outputs also account for low scores of MLLMs on CODIS. In summary, we pave a path for assessing capability of MLLMs on utilizing contextual information, which is a vital ability in real-world scenarios.

Limitations

There are still some limitations that need to be addressed. First, contexts in this work are brief due to the limited ability of current MLLMs in effectively utilizing contextual information. We believe it is essential to assess capability of extracting information from lengthy and intricate context to enhance visual comprehension in the future, as MLLMs evolve and become more powerful. Second, despite covering a wide range of scenarios and five representative types of context, there are still other forms of contexts that we intend to explore in future endeavors. Moreover, delving into context from other modalities such as audio and depth presents an interesting avenue for exploration. Moving forward, we aim to establish an automated pipeline for generating queries with context information spanning a broader range of scenarios and modalities.

Acknowledgements

This work is supported by the National Key R&D Program of China (2022ZD0160502) and the National Natural Science Foundation of China (No. 61925601, 62276152). We thank Sijie Cheng, Zhicheng Guo for their advice in paper writing. We thank Zonghan Yang, Yile Wang and Yuanchi Zhang for their valuable discussions.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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*.

Christopher Baldassano, Andre Esteva, Li Fei-Fei, and Diane M Beck. 2016. Two distinct scene-processing networks connecting vision and memory. *Eneuro*, 3(5).

Moshe Bar and Elissa Aminoff. 2003. Cortical analysis of visual context. *Neuron*, 38(2):347–358.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. ShareGPT4V: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. 2023b. DiffusionDet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO Captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. 2023c. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*.

Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. 2023. RelTR: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in GPT-4V(ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language odels with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Rachel N Denison, William T Adler, Marisa Carrasco, and Wei Ji Ma. 2018. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, 115(43):11090–11095.

Tausif Diwan, G Anirudh, and Jitendra V Tembhurne. 2023. Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications*, 82(6):9243–9275.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023a. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. 2023b. A challenger to GPT-4V? early explorations of Gemini in visual expertise. *arXiv preprint arXiv:2312.12436*.

G Gemini Team, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Joshua OS Goh, Soon Chun Siong, Denise Park, Angela Gutchess, Andy Hebrank, and Michael WL Chee. 2004. Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. *Journal of Neuroscience*, 24(45):10223–10228.

Tianrui Guan, Fuxiao Liu, Xiyang Wu Ruiqi Xian Zongxia Li, Xiaoyu Liu Xijun Wang, Lichang Chen Furong Huang Yaser Yacoob, and Dinesh Manocha Tianyi Zhou. 2023. HallusionBench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv e-prints*, pages arXiv–2310.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual Programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. MIMIC-IT: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. SEED-Bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. 2023c. Fine-tuning multimodal llms to follow zeroshot demonstrative instructions. *arXiv preprint arXiv:2308.04152*, 3.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

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

Yang Liu, Guanbin Li, and Liang Lin. 2023c. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023d. MM-Bench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. 2023. Semantic-conditional diffusion networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23359–23368.

OpenAI. 2023. GPT-4V(ision) System Card.

Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. 2021. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1183–1192.

Sai Rajeswar, Pau Rodriguez, Soumye Singhal, David Vazquez, and Aaron Courville. 2022. Multi-label iterated learning for image classification with label ambiguity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4783–4793.

Kate Sanders, Reno Kriz, Anqi Liu, and Benjamin Van Durme. 2022. Ambiguous images with human judgments for robust visual event classification. *Advances in Neural Information Processing Systems*, 35:2637–2650.

Andranik Sargsyan, Shant Navasardyan, Xingqian Xu, and Humphrey Shi. 2023. MI-GAN: A simple baseline for image inpainting on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7335–7345.

Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, et al. 2022. Is one annotation enough?-a data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215–33232.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. 2023. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*.

Melissa Le-Hoa Vo. 2021. The meaning and structure of scenes. *Vision Research*, 181:10–20.

Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, and Linlin Li. 2023. Controllable image captioning via prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2617–2625.

Xuan Wang and Zhigang Zhu. 2023. Context understanding in computer vision: A survey. *Computer Vision and Image Understanding*, 229:103646.

Hanyu Xiang, Qin Zou, Muhammad Ali Nawaz, Xianfeng Huang, Fan Zhang, and Hongkai Yu. 2023. Deep learning for image inpainting: A survey. *Pattern Recognition*, 134:109046.

Jinsheng Xiao, Haowen Guo, Jian Zhou, Tao Zhao, Qiuze Yu, Yunhua Chen, and Zhongyuan Wang. 2023. Tiny object detection with context enhancement and feature purification. *Expert Systems with Applications*, 211:118665.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

Xiaobo Zhang, Donghai Zhai, Tianrui Li, Yuxin Zhou, and Yang Lin. 2023. Image inpainting based on deep learning: A review. *Information Fusion*, 90:74–94.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. MMICL: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. 2023. Prototype-based embedding network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22783–22792.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, et al. 2022. Scene graph generation: A comprehensive survey. *arXiv preprint arXiv:2201.00443*.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object detection in 20 years: A survey. *Proceedings of the IEEE*.

A Consistency of Human Evaluation

A total of four individuals participated in human evaluation of model outputs. We examine the consistency of their assessments. A random sample of 100 queries out of 706 was selected for evaluation. Each of the four individuals made judgments on the model outputs of these 100 queries individually. A groundtruth value of 1 was assigned when the majority agreed the answer was correct, and a value of 0 was assigned when the majority agreed the answer was incorrect. In cases of a tie, a fifth individual was introduced. The average consistency between human judgments and groundtruths was then calculated as follows:

$$\text{Consistency} = \frac{1}{4n_q}\sum_{i=1}^{n_q}\sum_{j=1}^{4}(\text{Eval}_j(O_i) = \text{GT}_i),$$

where n_q represents the total number of queries, $\text{Eval}_j(O_i)$ denotes the evaluation of individual j on model output of query i and GT_i indicates the groundtruth assessment of model output of query i. The results are shown in Table 8.

Model	Human Consistency
GPT-4V	98.3
Gemini	99.3
mPLUG-Owl-2	97.3
MiniGPT4-7B	95.3
LLaVA-1.5-7B	97.5
InstructBLIP-7B	99.5
Otter-7B	97.0
LLaVA-7B	96.8
Qwen-VL-Chat	96.5
OpenFlamingo-7B	99.0
BLIP-2-6.7B	99.5
LLaVA-1.5-13B	98.8
BLIP-2-11B	97.0
InstructBLIP-13B	98.8

Table 8: Consistency of human evaluation on CODIS.

The results indicate that the consistency of human evaluation is high. As a result, to enhance efficiency in the remaining manual evaluation tasks in this study, we distributed tasks evenly among the four individuals instead of employing the majority vote method. Moreover, agreement between human and GPT-4 evaluators , as portrayed in Table 7, is close to human consistency, indicating proficiency of GPT-4 on evaluating model outputs is comparable to that of humans.

B Human Scores on CODIS

We enlisted participation of five volunteers to respond to the questions in CODIS. Their average age was 31 years old and they possess proficient daily English communication skills. As they are volunteers, they are not paid for the test. Before the test, all the volunteers were completely unaware of the questions and answers.

Testee	Acc_p	Acc_q
Human 1	91.8	92.6
Human 2	82.7	84.4
Human 3	89.0	89.4
Human 4	83.6	84.6
Human 5	83.9	87.4
Average	86.2	87.7

Table 9: Human performance on CODIS.

As shown in Table 9, humans perform well in CODIS benchmark, and achieve significantly higher scores than MLLMs. However, they do not reach 100% scores. Upon examining the results, we identify two primary reasons for this. Firstly, some questions require a lengthy chain of reasoning, which poses challenges for humans as well. Secondly, humans sometimes lack specific background knowledge, particularly in questions related to cultural backgrounds.

C Cases of Model Outputs When Removing Context

We provide three cases in Figure 4 to show alterations in model outputs resulting from the removal of context information. We use color green and red to emphasize model outputs in agreement and disagreement with the groundtruth answers, and use yellow to highlight the words associated with context information in model outputs.

D Cases of Model Outputs When Replacing Images with Captions

We provide three cases in Figure 5 to show alterations in model outputs when replacing images with captions. We employ the identical highlighting markers as those utilized in Appendix C.

E Detailed Results of GPT-4 Evaluation

We report detailed results based on GPT-4 evaluation in Table 10. We report Acc_p and Acc_q. The results exhibit a strong alignment with the results in Table 3 based on human evaluation.

Model	Loc & Ori Acc$_p$	Loc & Ori Acc$_q$	Temporal Acc$_p$	Temporal Acc$_q$	Cultural Acc$_p$	Cultural Acc$_q$	Attributes Acc$_p$	Attributes Acc$_q$	Relationships Acc$_p$	Relationships Acc$_q$	Overall Acc$_p$	Overall Acc$_q$
API-based Models												
GPT-4V	33.3	53.6	28.4	50.5	21.3	53.2	25.3	54.0	50.0	69.2	31.2	55.1
Gemini	20.2	48.8	27.4	50.0	21.3	54.3	22.7	51.3	30.8	54.8	24.4	51.3
Open-source ~13B Models												
LLaVA-1.5-13B	6.0	41.1	3.2	43.2	12.8	46.8	13.3	50.0	11.5	53.8	8.5	46.2
BLIP-2-11B	6.0	34.5	10.5	44.2	4.3	30.9	6.7	40.7	11.5	47.1	8.0	39.8
InstructBLIP-13B	6.0	39.9	2.1	41.1	6.4	46.8	4.0	44.7	5.8	48.1	4.5	43.3
Open-source ~7B Models												
mPLUG-Owl-2	13.1	39.9	9.5	40.0	4.3	41.5	9.3	42.7	11.5	48.1	9.9	41.9
MiniGPT4-7B	10.7	34.5	4.2	32.1	0.0	27.7	8.0	32.7	9.6	43.3	6.8	33.9
LLaVA-1.5-7B	8.3	37.5	1.1	36.3	2.1	40.4	9.3	37.3	7.7	48.1	5.7	39.1
InstructBLIP-7B	1.2	34.5	5.3	43.7	0.0	45.7	4.0	44.0	11.5	47.1	4.2	42.4
Otter-7B	2.4	31.5	3.2	35.3	0.0	23.4	1.3	27.3	5.8	34.6	2.5	31.0
LLaVA-7B	2.4	29.8	4.2	33.7	0.0	17.0	2.7	33.3	1.9	37.5	2.5	31.0
Qwen-VL-Chat	4.8	23.8	3.2	23.7	0.0	23.4	1.3	32.0	7.7	33.7	3.4	26.9
OpenFlamingo-7B	2.4	40.5	2.1	38.9	0.0	27.7	5.3	36.0	5.8	47.1	3.1	38.4
BLIP-2-6.7B	0.0	42.3	1.1	43.2	4.3	48.9	4.0	46.7	5.8	51.0	2.5	45.6

Table 10: Results of MLLMs on CODIS. All the model outputs are assessed by GPT-4.

F Prompt for Model Inference

As shown in Table 11, we design four prompts for model inference, and investigate impacts of detailed instructions and chain of thought prompting on model performance.

- **Detailed Instructions (DI)**: We stated that contexts contain helpful information that is not included in the image and encourage models to refer to contexts.
- **Chain of Thought Prompting (CoT)**: We asked models to think of the questions step by step, and then summarize the reasoning processes and answer the question in single words or phrases.

We randomly select 200 pairs, 400 queries from our benchmark and report Acc_p scores of GPT-4V, Gemini, mPLUG-Owl-2 and LLaVA-1.5-13B with the prompts above. The results are shown in Table 13.

The results indicated that DI and CoT help improve model performances. There were no significant differences in performance observed for LLaVA-1.5-13B when using either of the prompts, while the other models exhibit their best performances with DI and CoT. Therefore, based on the results, we decided to use the prompt with both DI and CoT for all the models in our evaluation process.

Additionally, it is worth noting that there are subtle variations in input formats and settings of different models, including placement of images, use of special tokens, etc. Throughout our evaluation, we adhere to the default configurations of each model. Details regarding input formats and settings of our evaluated MLLMs can be found in their documentations or repositories.

G Prompt for GPT-4 Evaluation

We report prompt for GPT-4 evaluation in Table 12. The results of GPT-4 evaluation in Table 7 and results in Table 10 are both based on the prompt.

H More Cases of CODIS

We present additional cases of CODIS, showcasing outputs of MLLMs on these cases in Figure 6, Figure 7, Figure 8, Figure 9 and Figure 10. We select two cases from each of the five categories. The incorrect model responses are highlighted in red. To aid readers in better comprehension, we provide explanations regarding the ambiguities in each image. It is important to note that these explanations are not visible to MLLMs or human volunteers who answered the questions in CODIS.

Prompt for Model Inference	
w/ DI w/ CoT	I'll give you an image and some additional context, which provides information closely related to the scene of the picture. Please answer my question based on the image and the context. Be sure to refer to the context and extract necessary information from it to help you answer the question because it contains helpful information that is not included in the image. Your answer should contain two parts. Two parts should be separated by a newline. In the first part, please think of the question step by step based on the image and context and output your reasoning process. In the second part, please summarize your reasoning process and directly answer the question in a single word or phrase. Context: [CONTEXT HERE] Question: [QUESTION HERE]
w/ DI w/o CoT	I'll give you an image and some additional context, which provides information closely related to the scene of the picture. Please answer my question based on the image and the context. Be sure to refer to the context and extract necessary information from it to help you answer the question because it contains helpful information that is not included in the image. Please answer in a single word or phrase. Context: [CONTEXT HERE] Question: [QUESTION HERE]
w/o DI w/ CoT	I'll give you an image and some additional context, which provides information closely related to the scene of the picture. Please answer my question based on the image and the context. Your answer should contain two parts. Two parts should be seperated by a newline. In the first part, please think of the question step by step based on the image and context and output your reasoning process. In the second part, please summarize your reasoning process and directly answer the question in a single word or phrase. Context: [CONTEXT HERE] Question: [QUESTION HERE]
w/o DI w/o CoT	I'll give you an image and some additional context, which provides information closely related to the scene of the picture. Please answer my question based on the image and the context. Please answer in a single word or phrase. Context: [CONTEXT HERE] Question: [QUESTION HERE]

Table 11: Prompt for model inference.

Prompt for Evaluation
Please evaluate the output of models based on the given question and groundtruth and tell me whether the output is right. Please pay attention to the following rules: 1. The output contains rationale of the reasoning process and answer which is summarized from the reasoning process. Please extract the answer from the output and make your judgement only based on answer, NOT rationale. 2. The answer is right if it follows the question in meaning and is consistent with the groundtruth. 3. Do not be too strict about the answer. Format different from the groundtruth and minor grammar issues are allowed. If you think the answer is correct according to the groundtruth, please output "right", otherwise output "wrong". You can only print "right" or "wrong" and nothing else. Here is the question: [QUESTION HERE] Here is the groundtruth: [GROUNDTRUTH HERE] Here is the output: [OUTPUT HERE]

Table 12: Prompt for GPT-4 evaluation.

Model	w/ DI w/ CoT	w/ DI w/o CoT	w/o DI w/ CoT	w/o DI w/o CoT
GPT-4V	40.0	22.0	34.0	25.0
Gemini	32.0	16.0	27.0	11.0
mPLUG-Owl-2	13.0	14.0	11.0	13.0
LLaVA-1.5-13B	15.0	4.0	11.0	6.0

Table 13: Model performances with different prompts. We report Acc_p.

Question: What were they doing? Greeting or saying goodbye to each other?
Context: He was separated from his newborn daughter and met again when she was five years old.

w/ Context
Human: Greeting.
GPT-4V: Greeting. The context specifies that the man was separated from his newborn daughter and met her again when she was five years old. This reuniting scenario typically involves a greeting.
Gemini: Greeting. They were separated from each other for five years, so this is probably the first time they have seen each other in five years.
mPLUG-Owl-2-7B: Greeting. The context mentions that the man was separated from his newborn daughter and met again when she was five years old. This suggests that they were reunited after a long time.
LLaVA-1.5-13B: Greeting. The context suggests that they were separated for a long time and have just met again, so it is more likely that they are greeting each other.

w/o Context
Human: Indeterminate.
GPT-4V: Ambiguous. Without contextual information about what is happening before or after this moment, it is not possible to definitively determine whether they are greeting or saying goodbye.
Gemini: It's not possible to tell whether they were greeting or saying goodbye based on the image and the context.
mPLUG-Owl-2-7B: Saying goodbye. The girl is hugging a stuffed animal and the man is holding her. This suggests that they are saying goodbye to each other.
LLaVA-1.5-13B: Greeting.

Question: Which direction was the photographer facing? East, west, south or north?
Context: I went out with my friend in the early morning.

w/ Context
Human: West.
GPT-4V: West. Given that the context states that the photo was taken in the early morning, it is known that the sun rises in the east. Therefore, if the shadows are cast in front of the photographer, the photographer must be facing the opposite direction of the sunrise.
Gemini: West. The shadows are pointing to the west. In the early morning, the sun rises in the east, so the photographer was facing west.
mPLUG-Owl-2-7B: East. The sun rises in the east and sets in the west. The shadows are cast from the people in the east, which means the sun is rising in the east. The photographer was facing east.
LLaVA-1.5-13B: East. The shadows of the people are cast to the left, which indicates that the photographer was facing east.

w/o Context
Human: Indeterminate.
GPT-4V: Without additional context, the photo does not provide enough information to conclusively answer the question. If the photo was taken in the morning, the photographer would be facing west, as the sun would be rising in the east and casting shadows forward.
Gemini: East. The shadows are pointing towards the west. Therefore the photographer was facing east.
mPLUG-Owl-2-7B: East. Step 1: Identify the direction of the sun. Step 2: Determine the shadow direction based on the sun's direction. Step 3: Identify the direction that matches the shadow direction.
LLaVA-1.5-13B: East.

Question: Will he get a fine?
Context: He drove at 40 kph.

w/ Context
Human: No.
GPT-4V: No. The context provided informs us that he drove at 40 kph. Since he was driving within the legal speed limit set by the traffic sign, he did not exceed the maximum allowed speed.
Gemini: He drove at 40 kph. The speed limit is 50 kph. So he did not exceed the speed limit. He will not get a fine. No.
mPLUG-Owl-2-7B: Yes. The speed limit is 50 kph. The driver was driving at 40 kph. The driver is driving faster than the speed limit. Driving faster than the speed limit is against the law. The driver might get a fine for driving faster than the speed limit.
LLaVA-1.5-13B: No. The image shows a red and yellow sign with the number 50 on it. The context suggests that the person drove at 40 kph.

w/o Context
Human: Indeterminate.
GPT-4V: Without additional context such as the person's actual driving speed, their adherence to the speed limit, and the relevant traffic laws, it's impossible to accurately predict from this image alone whether someone will get a fine.
Gemini: No. The context does not provide any information about the driver's speed, so I cannot determine whether he will get a fine.
mPLUG-Owl-2-7B: Yes. The image shows a sign with the number 50 on it. The number 50 is often associated with speed limits. The sign suggest that there might be a restriction in place.
LLaVA-1.5-13B: Yes.

Figure 4: Three cases to show alterations in model outputs resulting from the removal of context information.

Question: Is the scene dangerous?

Context: The speed of the car is 60 km/h.

Caption: From the photographer's point of view, the car is on the left side of the road, facing the photographer, and more than half of the car is in the grass by the side of the road.

w/ Image w/o Caption

GPT-4V: No. The scene does not appear to be inherently dangerous based on the image and the provided context of traveling at 60 km/h.

Gemini: No. The context states that the speed of the car is 60 km/h. This is not particularly fast, but it is still a significant speed. Overall, the scene does not appear to be particularly dangerous.

mPLUG-Owl-2-7B: The speed of the car is 60 km/h. Since the speed limit is not mentioned, we cannot determine if the car is speeding or not, and we cannot determine if the scene is dangerous or not.

LLaVA-1.5-13B: No.

w/ Caption w/o Image

GPT-4V: Yes. A car traveling at 60 km/h that has veered off the road into the grass suggests a potential loss of control, making the scene dangerous.

Gemini: Yes. Some part of the car is on the road, and more specifically on the wrong side of the road. The context tells us that the car is travelling at a speed of 60 km/h. This is relatively fast. The combined information leads to the conclusion that the car is dangerously out of control and is therefore posing a danger.

mPLUG-Owl-2-7B: Yes. A two-lane highway typically has a speed limit, and the car is going 60 km/h, which is above the speed limit. The scene is dangerous due to the car being above the speed limit and more than half in the grass by the side of the road.

LLaVA-1.5-13B: Yes. The car is traveling at a speed of 60 km/h, which is relatively fast for a rural highway. Additionally, more than half of the car is in the grass by the side of the road, indicating that it may have veered off the road or is in the process of doing so. This situation could be dangerous for the car's occupants and other road users.

Question: What does the sign say?

Context: The guard noticed me that I could not exit the building through this door.

Caption: In the picture is a door with a sign on it. The words on the sign can be read in two ways, horizontally: "DO NOT ENTER, EXIT ONLY", and vertically: "DO NOT EXIT, ENTER ONLY".

w/ Image w/o Caption

GPT-4V: The sign says "DO NOT ENTER EXIT ONLY". The image clearly shows the sign on the door stating "DO NOT ENTER EXIT ONLY," which means that the door is meant only for exiting the building and not for entering it.

Gemini: The sign says "Do not enter, exit only". The guard told me that I could not exit the building through this door. So this is an exit. The sign says "Do not enter". So this is an exit only door.

mPLUG-Owl-2-7B: Do Not Enter Exit Only. The image shows a sign on a door that says "Do Not Enter Exit Only."

LLaVA-1.5-13B: Do not enter exit only.

w/ Caption w/o Image

GPT-4V: Enter Only. The caption mentions that the sign can be read in two ways, horizontally and vertically. Since the guard said you cannot exit, the sign must say "Exit Only" when read horizontally and "Enter Only" when read vertically.

Gemini: DO NOT EXIT, ENTER ONLY. From the caption, we know that the sign can be read in two ways. And the context tells us that the guard noticed the person that he could not exit the building through this door. This implies that the sign is actually saying "DO NOT EXIT, ENTER ONLY".

mPLUG-Owl-2-7B: The caption states that the sign can be read in two ways. The context mentions that the guard noticed the person could not exit the building through the door. Based on the context, the sign says "DO NOT EXIT, ENTER ONLY" because the person was unable to exit the building through the door.

LLaVA-1.5-13B: The sign says "DO NOT EXIT, ENTER ONLY". Since the context states that the guard noticed that the person could not exit the building through this door, it is more likely that the sign is meant to be read as "DO NOT EXIT, ENTER ONLY". This would make sense in the context of the image, as it would prevent people from leaving the building through this door.

Question: mPLUG-Owl-2-7B

Context: I am currently in my car, driving in the United Kingdom.

Caption: The picture shows a two-way lane, and the photographer is on the right side of the road.

w/ Image w/o Caption

GPT-4V: Based on the context of driving in the UK, where driving on the left side of the road is the correct practice, and without contradictory evidence in the image, it would be logical to assume that you are driving in the correct lane.

Gemini: Yes. I am driving in the United Kingdom. In the United Kingdom, cars drive on the left side of the road. Therefore, I am driving in the correct lane.

mPLUG-Owl-2-7B: In the image, there are no road markings or signs that can be seen. I will consider the context provided, which states that I am currently in my car, driving in the United Kingdom. The context does not provide any specific information about the lane I should be in. Based on the image and context, I cannot determine if I am driving in the correct lane.

LLaVA-1.5-13B: Yes.

w/ Caption w/o Image

GPT-4V: No. Since the context states that you are driving in the United Kingdom and the picture shows the right side of the road, you would only be in the correct lane if you were driving in the direction towards the photographer.

Gemini: No. I am in the United Kingdom, which is a country that drives on the left side of the road.

mPLUG-Owl-2-7B: The caption states that the photographer is on the right side of the road, which means that the correct lane to be in is the one on the right. Yes, you are driving in the correct lane.

LLaVA-1.5-13B: No. The image shows a two-way lane, which means there are two directions of traffic. The photographer is on the right side of the road, so if you are in the United Kingdom, you would be driving on the left side of the road. In this case, you would be in the incorrect lane.

Figure 5: Three cases to show alterations in model outputs when replacing images with captions.

Question: In which direction is the airplane in this picture flying? East, west, south or north?

Explanation: The sun rises in the east and sets in the west. Given that the picture is taken in the morning or afternoon, the orientation of the airplane can be determined.

Context 1: I took this picture in the early morning.
Answer: West.

GPT-4V: It would suggest that the aircraft is flying westward, towards the setting position of the sun from the morning perspective. \nWest
Gemini: The airplane is flying west.
LLaVA-1.5-13B: East.
BLIP-2-11B: The airplane is flying east because the sun is rising in the east.
InstructBLIP-13B: West.
mPLUG-Owl-2-7B: The airplane is flying east.
MiniGPT4-7B: The airplane is flying in the east direction.
LLaVA-1.5-7B: The airplane in this picture is flying west.
InstructBLIP-7B: The airplane is flying west.
Otter-7B: The airplane in the picture is flying to the south.
LLaVA-7B: The airplane in the picture is flying towards the sun, which that it is flying towards the east.
Qwen-VL-Chat: The airplane could be flying in any direction, including east, west, south, or north.
OpenFlamingo-7B: The airplane is flying north.
BLIP-2-6.7B: East.

Context 2: I took this picture in the late afternoon.
Answer: East.

GPT-4V: The airplane would be flying westward, heading towards the setting sun.\n\nWest
Gemini: The answer is: east.
LLaVA-1.5-13B: East.
BLIP-2-11B: The airplane is flying west because the sun is setting in the west.
InstructBLIP-13B: West.
mPLUG-Owl-2-7B: The airplane is flying in the opposite direction, so it is flying east.
MiniGPT4-7B: The airplane is flying in the east direction.
LLaVA-1.5-7B: The airplane in the picture is flying west.
InstructBLIP-7B: The airplane is flying west.
Otter-7B: The airplane in the picture is flying to the south.
LLaVA-7B: The airplane in the picture is flying towards the sun, which suggests that it is flying towards the west.
Qwen-VL-Chat: The airplane could be flying in any direction, including east, west, south, or north.
OpenFlamingo-7B: The airplane is flying north.
BLIP-2-6.7B: East.

Question: Is the train arriving or leaving the platform?

Explanation: The remote end of the train is outside the platform. Thus, if the train is coming towards me, it is arriving the platform, and if the train is going away from me, it is leaving the platform.

Context 1: The subway is coming towards me.
Answer: Arriving the platform.

GPT-4V: Since the context ,we can conclude that the train is arriving at the platform rather than leaving it. Arriving.
Gemini: The train is still moving and the station platform is visible in the background. Therefore, the train is arriving at the platform.
LLaVA-1.5-13B: Arriving.
BLIP-2-11B: The train is coming towards me so it is arriving.
InstructBLIP-13B: The train is arriving at the platform.
mPLUG-Owl-2-7B: Arriving.
MiniGPT4-7B: The train is arriving at the platform.
LLaVA-1.5-7B: Based on the context and the image, the train is approaching the platform.
InstructBLIP-7B: The train is arriving at the platform.
Otter-7B: The train is arriving at the platform.
LLaVA-7B: The train is arriving at the subway station platform.
Qwen-VL-Chat: The train is arriving at the platform.
OpenFlamingo-7B: Arriving.
BLIP-2-6.7B: Arriving.

Context 2: The subway is going away from me.
Answer: Leaving the platform.

GPT-4V: Based on this statement, it can be inferred that the train is in the process of departing from the station platform. The train is leaving the platform.
Gemini: Going away from me" means the train is leaving the platform. Answer: leaving
LLaVA-1.5-13B: Leaving.
BLIP-2-11B: The train is leaving the platform.
InstructBLIP-13B: The train is leaving the platform.
mPLUG-Owl-2-7B: Leaving.
MiniGPT4-7B: The train is leaving the platform because it is traveling away from the observer, who is standing on the platform.
LLaVA-1.5-7B: Leaving.
InstructBLIP-7B: The train is arriving at the platform.
Otter-7B: The train is arriving at the platform.
LLaVA-7B: The train is leaving the platform.
Qwen-VL-Chat: Based on the context provided, the train is going away from the person, so it is leaving the platform.
OpenFlamingo-7B: Arriving.
BLIP-2-6.7B: Arriving.

Figure 6: More cases of category of location and orientation.

Question: Am I opening or closing the door?

Explanation: The person in the picture may be opening the door to get in the car or closing the door after he got out of the car.

Context 1: My friend is saying hello to me across the street.
Answer: Closing the door.

GPT-4V: Based on the image provided, it would be more plausible that you are closing the car door after exiting the vehicle to engage with your friend. Closing the door.
Gemini: Opening.
LLaVA-1.5-13B: Opening.
BLIP-2-11B: Opening.
InstructBLIP-13B: You are closing the door.
mPLUG-Owl-2-7B: Opening.
MiniGPT4-7B: The man is getting out of the taxi.
LLaVA-1.5-7B: Opening.
InstructBLIP-7B: Closing.
Otter-7B: I am opening the door.
LLaVA-7B: he is opening the door to let your friend in.
Qwen-VL-Chat: Based on the context, it is not clear whether the person is opening or closing the door.
OpenFlamingo-7B: Opening.
BLIP-2-6.7B: Opening.

Context 2: My friend is saying goodbye to me across the street.
Answer: Opening the door.

GPT-4V: Given that saying goodbye usually happens when someone is leaving, and, the most likely interpretation of these two pieces of information is that you are opening the door .
Gemini: The man is pulling the door handle, which means he is opening the door.
LLaVA-1.5-13B: Opening.
BLIP-2-11B: Opening.
InstructBLIP-13B: Closing the door.
mPLUG-Owl-2-7B: Opening.
MiniGPT4-7B: It is not possible to determine whether the man is opening or closing the taxi door.
LLaVA-1.5-7B: Opening.
InstructBLIP-7B: Closing.
Otter-7B: I am opening the door.
LLaVA-7B: He is either opening or closing the door.
Qwen-VL-Chat: The person is opening the door of a taxi.
OpenFlamingo-7B: Opening.
BLIP-2-6.7B: I'm opening the door.

Question: What is the moving direction of my hand? Up or down?

Explanation: The person in the picture has his hand on the zipper of his jacket. However, whether he is pulling the zipper up or down is ambiguous.

Context 1: I am taking off my clothes.
Answer: Down.

GPT-4V: Based on the context that you are taking off your clothes and the image that I see, Taking the clothing off would move your hands downward after unzipping. Your hand is moving down.
Gemini: In order to take off his clothes, he needs to unzip his jacket. To unzip the jacket, he needs to pull the zipper down. Therefore, the moving direction of his hand is down.
LLaVA-1.5-13B: Up.
BLIP-2-11B: Down.
InstructBLIP-13B: The moving direction is up.
mPLUG-Owl-2-7B: Answer: Up.
MiniGPT4-7B: The direction of the hand is not specified in the image or context.
LLaVA-1.5-7B: The hand is holding the jacket, and it is moving up.
InstructBLIP-7B: The moving direction is down.
Otter-7B: Down.
LLaVA-7B: The moving direction of your hand is down, as you are taking off your clothes.
Qwen-VL-Chat: It is not possible to determine the moving direction of the hand based on the given context.
OpenFlamingo-7B: Up.
BLIP-2-6.7B: Up.

Context 2: I am putting on my clothes.
Answer: Up.

GPT-4V: Given the context and the typical motions involved in dressing, they could be moving their hand in either direction, depending on whether they are starting to zip or finishing Up or down.
Gemini: If he wants to put on his clothes, he needs to pull the zipper up. So the moving direction of his hand is up. Answer: Up.
LLaVA-1.5-13B: Up.
BLIP-2-11B: Up. My hand is moving up.
InstructBLIP-13B: Up.
mPLUG-Owl-2-7B: Answer: Up.
MiniGPT4-7B: It is not possible to determine whether it is moving up or down.
LLaVA-1.5-7B: My hand is moving up.
InstructBLIP-7B: The moving direction of the hand in the image is down.
Otter-7B: Down.
LLaVA-7B: The moving direction of your hand is up, as you are putting on your clothes.
Qwen-VL-Chat: The answer to the question "What is the moving direction of my hand? Up or down?" cannot be determined.
OpenFlamingo-7B: Down.
BLIP-2-6.7B: Up.

Figure 7: More cases of category of temporal information.

Question: Would this door number probably make me feel uncomfortable?

Explanation: The ambiguity of this picture lies in the cultural differences. For Chinese people, number 4 is not very auspicious, but many people in other countries do not avoid number 4.

Context 1: I am a tourist from China.
Answer: Yes.

GPT-4V: Since you are a tourist from China, the number 14 itself does not have as strong a cultural stigma as the number 4 alone, so it might not necessarily make you feel uncomfortable.
Gemini: In Chinese culture, the number 14 sounds similar to mean "will die for sure", So this door number might make Chinese people feel uncomfortable.
LLaVA-1.5-13B: Yes.
BLIP-2-11B: The door number is a number that is not in my native language.
InstructBLIP-13B: Yes.
mPLUG-Owl-2-7B: No.
MiniGPT4-7B: It is not possible to determine whether the door number would make the tourist feel uncomfortable based on the image and context..
LLaVA-1.5-7B: The building's location is not relevant to the tourist's feelings of comfort or discomfort.
InstructBLIP-7B: Yes.
Otter-7B: The correct answer should be: yes.
LLaVA-7B: Yes, in Chinese culture, the number 14 is considered unlucky.
Qwen-VL-Chat: No.
OpenFlamingo-7B: No.
BLIP-2-6.7B: No.

Context 2: I'm a tourist from the United States.
Answer: No.

GPT-4V: Being a tourist from the United States, as the number 14 does not have the same cultural connotations of bad luck or superstition associated with it in the United States. No.
Gemini: In the United States, this door number would probably not make me feel uncomfortable..
LLaVA-1.5-13B: Yes.
BLIP-2-11B: The door number is a number that is not used in the United States.
InstructBLIP-13B: Yes.
mPLUG-Owl-2-7B: No, the door number 14 would not make me feel uncomfortable.
MiniGPT4-7B: It is likely that the door number would make the tourist feel uncomfortable.
LLaVA-1.5-7B: No.
InstructBLIP-7B: Yes.
Otter-7B: The answer provided is not appropriate to the question.
LLaVA-7B: Yes, the door number 14 might make me feel uncomfortable.
Qwen-VL-Chat: No, the door number 14 would not make me feel uncomfortable.
OpenFlamingo-7B: Yes.
BLIP-2-6.7B: No.

Question: Is my brother's behavior legal?

Explanation: The ambiguity of this picture lies in the differences in laws and regulations of different countries. For most states in America, only those over 21 years old can enter the casinos. In the UK, those over 18 years old can enter the casinos.

Context 1: My brother in the UK is celebrating his twentieth birthday here.
Answer: Yes.

GPT-4V: The brother's behavior of gambling at the age of twenty in the UK is legal.
Gemini: The legal age for gambling in the UK is 18. So it is legal for him to be in a casino. Yes.
LLaVA-1.5-13B: It is not possible to determine if the behavior of the people in the image is legal or not.
BLIP-2-11B: The answer: yes.
InstructBLIP-13B: No, it is not legal.
mPLUG-Owl-2-7B: The brother's behavior is legal.
MiniGPT4-7B: The image and context do not provide enough information to answer the question.
LLaVA-1.5-7B: Since my brother is in the UK and of legal age, it is likely that his behavior is legal.
InstructBLIP-7B: Yes.
Otter-7B: Depends on whether the game being played is legal in the country where it is taking place.
LLaVA-7B: It is not clear if the UK is one of them.
Qwen-VL-Chat: It would be necessary to look up the UK's gambling laws and regulations.
OpenFlamingo-7B: My brother's behavior is legal.
BLIP-2-6.7B: Yes, it is legal.

Context 2: My brother in America is celebrating his twentieth birthday here.
Answer: No.

GPT-4V: Based on the context and the provided image, if your brother is gambling at a casino in the United States at the age of 20, his behavior is likely illegal.
Gemini: The legal age for gambling in America is 21, so my brother's behavior is not legal. No.
LLaVA-1.5-13B: No.
BLIP-2-11B: The final answer: yes.
InstructBLIP-13B: No.
mPLUG-Owl-2-7B: Answer: No.
MiniGPT4-7B: It is important to check the local laws to determine whether the behavior is legal.
LLaVA-1.5-7B: Answer: No.
InstructBLIP-7B: It is best to consult with local authorities or legal experts to determine the answer.
Otter-7B: To determine if the game is legal it's up to the area where it takes place.
LLaVA-7B: To determine if his behavior is legal, one would need more information about the specific location.
Qwen-VL-Chat: It is not possible to determine whether my brother's behavior is legal or not.
OpenFlamingo-7B: My brother's behavior is legal.
BLIP-2-6.7B: Yes, it is legal.

Figure 8: More cases of category of cultural background.

Question: Can I eat the box of meat?

Explanation: The box of meat can be eaten before May 8. Whether I can eat it depends on the date today.

Context 1: It is May 6th today.
Answer: Yes.

GPT-4V: Yes, based on the \"Use by\" date of May 8th and today being May 6th, you can eat the box of meat.
Gemini: It is already May 6th, so the meat is past its \"Use By\" date. Therefore, it is not safe to eat the meat.\n\nNo.
LLaVA-1.5-13B: Yes.
BLIP-2-11B: The box of meat is expired. The answer: no.
InstructBLIP-13B: No, the box of meat is expired.
mPLUG-Owl-2-7B: Answer: Yes.
MiniGPT4-7B: Yes, you can eat the box of meat.
LLaVA-1.5-7B: Yes.
InstructBLIP-7B: No.
Otter-7B: The answer provided is not related to an image. Please provide an image or a context related to a image.
LLaVA-7B: Yes, you can eat the box of meat. The image shows a box of meat with a price tag on it, and the context indicates that it is May 6th. This suggests that the meat is fresh and ready for consumption.
Qwen-VL-Chat: Therefore, it is recommended to check the quality of the meat before eating it.
OpenFlamingo-7B: No.
BLIP-2-6.7B: Yes, you can eat the box of meat..

Context 2: It is May 9th today.
Answer: No.

GPT-4V: The box of meat is past its use-by date; therefore, it is not recommended to eat it.
Gemini: The \"Use By\" date on the package is May 8th, which has already passed. Therefore, it is not safe to eat the meat.\n\nNo.
LLaVA-1.5-13B: Yes.
BLIP-2-11B: The box of meat is expired. The answer: no.
InstructBLIP-13B: No, the box of meat is expired.
mPLUG-Owl-2-7B: No.
MiniGPT4-7B: Therefore, it is not clear whether the box of meat in the image is safe to eat.
LLaVA-1.5-7B: No.
InstructBLIP-7B: No, it is a packaging box.
Otter-7B: The answer provided is not appropriate to the question.
LLaVA-7B: Yes, you can eat the box of meat. The image shows a box of meat with a price tag, and the context indicates that it is May 9th today. This suggests that the meat is fresh and ready to be consumed.
Qwen-VL-Chat: Therefore, it is not safe to assume that the box of meat is still good to eat without further information.
OpenFlamingo-7B: Yes.
BLIP-2-6.7B: Yes, you can eat the box of meat.

Question: Is the person in danger?

Explanation: The baby may get scalded if the cup was filled with hot water. However, it would be harmless if the cup was filled with cool water.

Context 1: The cup is filled with hot water.
Answer: Yes.

GPT-4V: Since the child is actively pouring the hot water over themselves, the risk of injury from the heat is imminent.
Gemini: The cup is filled with hot water. The baby is taking a bath. The baby is not in danger.
LLaVA-1.5-13B: No.
BLIP-2-11B: The cup is filled with hot water. The person is a baby. The baby is playing with the water. The baby is not in danger. The answer: no.
InstructBLIP-13B: Yes, the person is in danger.
mPLUG-Owl-2-7B: Since the cup is filled with hot water, it is likely that the baby is in danger of being burned or scalded.\n\nAnswer: Yes, the person is in danger.
MiniGPT4-7B: No, the person is not in danger.
LLaVA-1.5-7B: Answer: No
InstructBLIP-7B: No.
Otter-7B: The image does not indicate that the person is in danger.
LLaVA-7B: Yes, the person is in danger.
Qwen-VL-Chat: No, the person is not in danger.
OpenFlamingo-7B: No.
BLIP-2-6.7B: No.

Context 2: The cup is filled with cold water.
Answer: No.

GPT-4V: The child does not appear to be in any danger that can be discerned from the image and the context provided.\n\nNo danger evident.
Gemini: The baby is taking a bath. The water is cold. The baby looks happy. So the baby is not in danger.
LLaVA-1.5-13B: No.
BLIP-2-11B: The person is playing with water. The water is cold. The person is not in danger. The answer: no.
InstructBLIP-13B: No, the person is not in danger.
mPLUG-Owl-2-7B: No, the person is not in danger.
MiniGPT4-7B: No, the person is not in danger.
LLaVA-1.5-7B: Answer: No.
InstructBLIP-7B: No.
Otter-7B: The image does not indicate that the person is in danger.
LLaVA-7B: Yes, the person is in danger. It is essential to ensure that the child is supervised and the water is placed in a safe location to prevent any potential hazards.
Qwen-VL-Chat: No, the person is not in danger.
OpenFlamingo-7B: The answer provided is not appropriate to the question.
BLIP-2-6.7B: No.

Figure 9: More cases of category of attributes.

Question: Is the principal standing on the left or on the right?

Explanation: We construct two scenarios. Models should determine whether the principal was standing on the left side or the right side accordingly.

Context 1: Our principal insisted on delivering his speech in the rain.
Answer: Left.

GPT-4V: Therefore, based on the given context and the observable actions in the image, it makes sense to conclude that the principal is the person giving the speech on the left.The principal is standing on the left.
Gemini: The person who is not holding an umbrella is the principal, and he is standing on the left.
LLaVA-1.5-13B: Left.
BLIP-2-11B: The principal is standing on the right.
InstructBLIP-13B: On the right.
mPLUG-Owl-2-7B: The man holding the umbrella is standing on the left side.
MiniGPT4-7B: The principal is standing on the left side of the image.
LLaVA-1.5-7B: Right.
InstructBLIP-7B: The principal is standing on the left.
Otter-7B: The principal is standing on both the left and the right.
LLaVA-7B: The principal is standing on the left side of the man holding the umbrella.
Qwen-VL-Chat: Based on the context provided, the principal is standing on the left.
OpenFlamingo-7B: The principal is standing on the left.
BLIP-2-6.7B: The principal is standing on the right.

Context 2: Our principal cared for his students. He sheltered the student who was giving his speech from the rain.
Answer: Right.

GPT-4V: The principal is standing on the right, holding the umbrella.
Gemini: So the principal is standing on the right.
LLaVA-1.5-13B: Left.
BLIP-2-11B: the left.
InstructBLIP-13B: The principal is standing on the right.
mPLUG-Owl-2-7B: The principal is standing on the left side of the image.
MiniGPT4-7B: The principal is standing on the left side of the image.
LLaVA-1.5-7B: The principal is standing on the right.
InstructBLIP-7B: The principal is standing on the left.
Otter-7B: The principal is standing on the right.
LLaVA-7B: The principal is standing on the left side of the student who is giving the speech.
Qwen-VL-Chat: Based on the context provided, the principal is standing on the right.
OpenFlamingo-7B: The principal is standing on the right.
BLIP-2-6.7B: The principal is standing on the right.

Question: Is my friend on the left or on the right in the perspective of the photographer?

Explanation: We construct two scenarios. Models should determine whether my friend was on the left side or the right side accordingly.

Context 1: My daughter was sitting on my lap. My friend said she was cute.
Answer: On the left.

GPT-4V: Therefore, the friend would be the remaining adult, who is on the left from the photographer's perspective.The friend is on the left.
Gemini: The friend is on the right.
LLaVA-1.5-13B: Right.
BLIP-2-11B: The friend is on the left.
InstructBLIP-13B: The friend is on the right in the perspective of the photographer.
mPLUG-Owl-2-7B: The friend is on the right side of the perspective of the photographer.
MiniGPT4-7B: The friend is on the right in the perspective of the photographer.
LLaVA-1.5-7B: My friend is on the left in the perspective of the photographer.
InstructBLIP-7B: The friend on the left.
Otter-7B: The woman standing next to her is holding a coffee cup, and the child is sitting on her lap.
LLaVA-7B: My friend is on the right.
Qwen-VL-Chat: Therefore, we cannot determine the friend's location based on the given context alone.
OpenFlamingo-7B: My friend is on the left.
BLIP-2-6.7B: The photographer is on the left.

Context 2: My friend said my daughter was cute. She let her sit on her lap and played with her.
Answer: On the right.

GPT-4V: The friend is on the right from the perspective of the photographer.
Gemini: The friend is on the right.
LLaVA-1.5-13B: Left.
BLIP-2-11B: The woman on the left is my daughter. The woman on the right is my friend.
InstructBLIP-13B: The friend is on the right in the perspective of the photographer.
mPLUG-Owl-2-7B: The friend is not visible in the image.
MiniGPT4-7B: The answer is that the woman holding the baby is on the left side of the image .
LLaVA-1.5-7B: My friend is on the left in the perspective of the photographer.
InstructBLIP-7B: The photographer is on the left of the image, as the woman on the left is sitting at the table with the baby.
Otter-7B: The friend is on the left.
LLaVA-7B: In the perspective of the photographer, the friend is on the right.
Qwen-VL-Chat: Therefore, the friend is on the left side.
OpenFlamingo-7B: My friend is on the left.
BLIP-2-6.7B: The photographer is on the left.

Figure 10: More cases of category of relationships.