MUCAR: Benchmarking Multilingual Cross-Modal Ambiguity Resolution for Multimodal Large Language Models

Xiaolong Wang[*,1,3], **Zhaolu Kang**[*,4], **Wangyuxuan Zhai**[*,6], **Xinyue Lou**[6], **Yunghwei Lai**[1], **Ziyue Wang**[1], **Yawen Wang**[1], **Kaiyu Huang**[6], **Yile Wang**[✉,5], **Peng Li**[✉,2], **Yang Liu**[1,2]
[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]Jiuquan Satellite Launch Center (JSLC), Gansu, China
[4]School of Software & Microelectronics, Peking University, Beijing, China
[5]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[6] Beijing Jiaotong University, Beijing, China
wangxl22@mails.tsinghua.edu.cn, kangzl9966@gmail.com, zhaiwangyuxuan@bjtu.edu.cn
wangyile@szu.edu.cn, lipeng@air.tsinghua.edu.cn, liuyang2011@tsinghua.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated significant advances across numerous vision-language tasks. MLLMs have shown promising capability in aligning visual and textual modalities, allowing them to process image–text pairs with clear and explicit meanings. However, resolving the inherent ambiguities present in real-world language and visual contexts remains a challenge. Existing multimodal benchmarks typically overlook linguistic and visual ambiguities, relying mainly on unimodal context for disambiguation and thus failing to exploit the mutual clarification potential between modalities. To bridge this gap, we introduce MUCAR, a novel and challenging benchmark designed explicitly for evaluating multimodal ambiguity resolution across multilingual and cross-modal scenarios. MUCAR includes: (1) a multilingual dataset where ambiguous textual expressions are uniquely resolved by corresponding visual contexts, and (2) a dual-ambiguity dataset that systematically pairs ambiguous images with ambiguous textual contexts, with each combination carefully constructed to yield a single, clear interpretation through mutual disambiguation. Extensive evaluations involving 19 state-of-the-art multimodal models–encompassing both open-source and proprietary architectures–reveal substantial gaps compared to human-level performance, highlighting the need for future research into more sophisticated cross-modal ambiguity comprehension methods, further pushing the boundaries of multimodal reasoning.

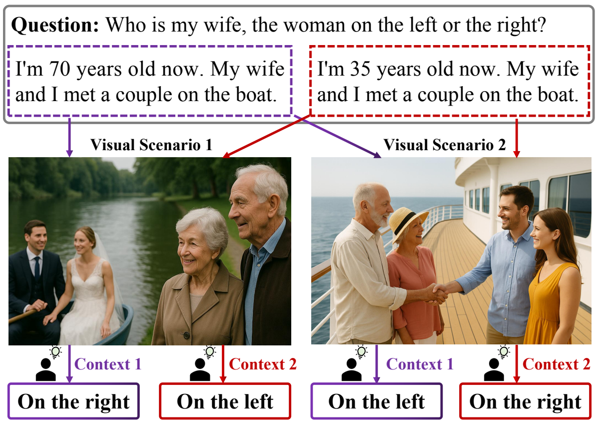


Figure 1: The interpretation of ambiguous text and visuals can be shaped by context and scenario. For instance, in "My wife and I met a couple on the boat," it is unclear whether "on the boat" modifies "met" or "a couple," while the image leaves the identity of the wife uncertain. Varying the visual setting (e.g., riverbank vs. cruise deck) and textual cues (e.g., age of the speaker) yields different resolutions. Each of the four context-image combinations leads to a distinct answer, with reasoning color-coded: purple (Context 1), red (Context 2).

1 Introduction

Multimodal Large Language Models (MLLMs; OpenAI 2023, 2024; Liu et al. 2023a; Dai et al. 2023) have advanced significantly in handling diverse inputs like text and images, with contextual understanding being key to their success in tasks such as question answering (Shao et al., 2023; Liu et al., 2023b; Antol et al., 2015), image captioning (Luo et al., 2023; Wang et al., 2023; Chen et al., 2015), and multimodal reasoning (Gupta and Kembhavi, 2023; Chen et al., 2023; Zellers et al., 2019). As MLLMs evolve, the ability to integrate multimodal context becomes crucial for accurate responses, underscoring the need for thorough evaluation of their contextual comprehension in real-world settings.

Prior studies have largely emphasized tasks with clear and unambiguous inputs (Fu et al., 2023a;

[*]Equal contribution.
[✉]Corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15038–15060
November 4-9, 2025 ©2025 Association for Computational Linguistics

Benchmark	Visual Ambiguity	Context Ambiguity	Multi-Languages	Evaluator
MME (Fu et al., 2023a)	✗	✗	✗	Metrics
MMBench (Liu et al., 2023c)	✗	✗	✗	GPT
MMT-Bench (Ying et al., 2024)	✗	✗	✗	GPT
MMStar (Chen et al., 2024a)	✗	✗	✗	Metrics
HallusionBench (Guan et al., 2023)	✓	✗	✗	Metrics
CODIS (Luo et al., 2024)	✓	✗	✗	Human / GPT
Illusory VQA (Rostamkhani et al., 2024)	✓	✗	✗	Human
MHaluBench (Chen et al., 2024b)	✓	✗	✗	GPT
MMA (Wang et al., 2024a)	✗	✓	✗	Metrics
VAGUE (Nam et al., 2024)	✗	✓	✗	Metrics
3AM (Ma et al., 2024)	✗	✓	✓	Metrics
UNPIE (Chung et al., 2024)	✗	✓	✓	GPT
MUCAR (Ours)	✓	✓	✓	Human / GPT

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

Ying et al., 2024; Li et al., 2023), frequently neglecting the ambiguity that naturally arises in both visual and textual modalities. Consider the example in Figure 2, neither the context nor the image alone can resolve the question "Who is my wife, the woman on the left or the right?", where a single sentence or image can often support multiple plausible interpretations depending on the specific scenario or provided context. Figure 2 illustrates this challenge clearly. Consider the sentence "My wife and I met a couple on the boat." This sentence contains structural ambiguity: it is unclear whether "on the boat" modifies the verb "met" (indicating the location of the meeting) or the noun phrase "a couple" (specifying the location of the couple). Simultaneously, the accompanying image introduces visual ambiguity concerning the referent of "the wife" among the depicted women. Notably, the ambiguity cannot be resolved independently within either modality; instead, mutual disambiguation arises when different textual scenarios (e.g., differing speaker ages) combine with different visual contexts (riverbank vs. cruise deck scenarios). Each unique combination yields a single, unambiguous interpretation, showing that textual and visual ambiguities can mutually clarify each other.

To systematically evaluate the capabilities of MLLMs to resolve such complex multimodal ambiguities, we introduce **MUCAR**, a novel benchmark specifically designed for **MU**ltilingual **C**ross-modal **A**mbiguity **R**esolution. Table 1 summarizes recent benchmarks designed to evaluate MLLMs in terms of *visual ambiguity*, *contextual ambiguity*, *multilinguality*, and the type of *evaluator* used (e.g., metrics, GPT, or human annotations). While early benchmarks such as MMT-Bench (Ying et al., 2024), MMStar (Liu et al., 2023c), and MME-RealWorld (Li et al., 2023) focus on general multimodal tasks, they lack coverage of ambiguity-related phenomena. More recent benchmarks like HallusionBench (Guan et al., 2023), Illusory VQA (Rostamkhani et al., 2024), and CODIS (Luo et al., 2024) begin to explore visual ambiguity, but often overlook contextual disambiguation or multilingual diversity. Notably, only a few benchmarks incorporate human evaluation, which is essential for assessing ambiguity understanding. To the best of our knowledge, MUCAR is the first benchmark to comprehensively address visual ambiguity, contextual ambiguity, and multilinguality, while integrating both human and GPT-based evaluation. This design enables a more rigorous and realistic assessment of ambiguity resolution capabilities in multimodal large language models. Our dataset and code are available at `https://github.com/THUNLP-MT/MUCAR`.

To summarize, our main contributions are:

- We construct **MUCAR**, the first *multilingual cross-modal ambiguity resolution benchmark*, featuring 1278 manually curated samples in Chinese, English, and Malay, including uniquely designed dual-ambiguity cases.
- We systemically evaluate **19 sota MLLMs** (both open-source and closed-source), revealing significant limitations in resolving multilingual multimodal ambiguities.

- We propose a **simple yet effective agent-based framework** for multimodal disambiguation, which improves performance through explicit cross-modal reasoning.

2 Related Work

Context Ambiguity Ambiguity is an inherent characteristic of linguistic text, emerging naturally due to the potential for multiple interpretations, especially in open-domain question answering tasks (Min et al., 2020; Sun et al., 2023). Existing research on ambiguity resolution in language primarily follows two directions. One line leverages contextual cues from surrounding text to resolve ambiguity, as seen in Gao et al. (2024); Lee et al. (2025), which uses in-context learning to disambiguate textual inputs. Another line exploits visual information to disambiguate language, particularly for polysemous word translation, such as in 3AM (Ma et al., 2024), which aligns ambiguous words with visual semantics.

Aligned with the multimodal ambiguity setting in MMA (Wang et al., 2024a), our work focuses on using visual input to disambiguate multilingual expressions. Beyond standard NLP ambiguities, we also address those arising from domain-specific and cultural differences. In particular, we construct a benchmark featuring both textual and visual ambiguity across multiple languages, aiming to evaluate the ability of MLLMs to resolve complex multimodal ambiguities in realistic scenarios.

Visual Ambiguity Visual ambiguity often stems from incomplete visual cues or interfering noise in the scene (Denison et al., 2018). Most previous vision-language benchmarks assume unambiguous (Liu et al., 2023c; Fu et al., 2023a; Liu et al., 2023c; Li et al., 2023) input or highlight the visual ambiguities caused by optical illusions (Guan et al., 2023; Rostamkhani et al., 2024; Cui et al., 2023; Fu et al., 2023b). Early Multimodal datasets like MS-COCO (Chen et al., 2015) focus on literal descriptions, while later works, e.g., CODIS (Luo et al., 2024) highlight the need for diverse context to reflect multiple valid interpretations. Inspired by CODIS (Luo et al., 2024), we assess the capability of MLLMs to disambiguate visual ambiguity through textual modalities instead of just recognizing ambiguities. Different from CODIS, we construct challenging dual-ambiguity instances, combining ambiguous visuals and texts that jointly resolve into a single interpretation, further testing the limits of multimodal reasoning.

3 MUCAR

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

3.1 Taxonomy

Given the extensive and varied nature of context information, comprehensive cataloging of all forms of context is challenging. With the aim of establishing an outstanding benchmark for disambiguation, we identified seven representative types. The first six types are inspired by the information people require to understand context. When collecting the data, we surprisingly found that in some cases, the combination of ambiguous text and ambiguous images led to mutual disambiguation. In other words, neither modality alone provided sufficient clarity, but together they resolved the ambiguity inherent in both. This observation inspired us to define the seventh type and collect the corresponding data. *To the best of our knowledge, we are the first to construct data for this type.* Figure 2 illustrates examples with corresponding classification explanations.

Polysemy. Navigli (2009) provides one formalization of polysemy, by referring to a word with two or more related meanings. These related meanings often share a conceptual core, with one meaning typically being an extension or variation of the other. Some contexts can be interpreted both literally and metaphorically. The “Polysemy” part in Figure 2 gives a good example.

Homonymy. Navigli (2009) also provides one formalization of homonymy. Opposite to polysemy, homonymy refers to a word having two or more unrelated meanings that stem from different historical origins, and the meanings of homonyms have no inherent connection. Disambiguating homonyms relies on the other elements. The “Homonymy” part in Figure 2 serves as a good example.

Grammar. This ambiguity occurs when sentence structures allow for multiple interpretations, often due to the placement of words or phrases. Such structural issues can make it unclear which part of the sentence a modifier applies to or the rela-

Figure 2: Taxonomy of our benchmark. We present one example for each category. Each example includes a context (C), a question (Q), and two different images with their corresponding answers and explanations.

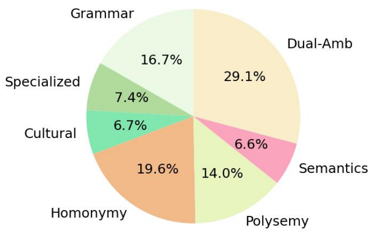
tionship between different clauses. The "Grammar" part in Figure 2 provides a clear illustration. As the "Grammar" part in the figure shows, in Chinese, "我的门没有锁" can be interpreted as "My door does not have a lock" or "My door has not been locked", here "锁" can be understood as a noun or verb, which leads to ambiguity.

Semantics. Understanding the timing and sequence of events is crucial when we understand a context. However, an isolated context can only provide us with static information, which is insufficient for dynamic events. The disambiguation can only be achieved when unambiguous image gives us more information. The "Semantics" case in Figure 2 provides a representative example.

Specialized. We define Specialized taxonomy to encompass terms or concepts that have distinct meanings across academic domains or personal situations. These terms often lead to ambiguity when encountered by individuals with varying background knowledge. The "Specialized" part in Figure 2 presents a good example.

Cultural. Some context can be interpreted differently depending on the cultural background of the interpreter, as words, symbols, and actions with specific meanings and connotations. Cultural norms, values, and historical experiences shape how individuals understand and react to information. This can lead to significant textual ambiguities where meaning is lost or distorted. The "Cultural" part in Figure 2 offers a good example.

Dual-ambiguity. This data type highlights our unique contribution. In this type, the context and the image are both ambiguous, but their combined information allows for clear disambiguation. Figure 2 gives a good example. The context ambiguity



Scenario	Percentage
People	24%
Landscape	13%
Cartoons	10%
Food	10%
Scene	7%
Indoors	7%
Actions	7%
Objects	7%
Streetscape	4%
Others	4%
Animals	3%
Sports	2%
Plants	2%

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

covers the six former types. Regarding image ambiguity, following CODIS (Luo et al., 2024), image ambiguity can be further categorized into distinct types, such as location and orientation, temporal information, cultural background, attributes, and relationships. However, for the purpose of this paper, we group these various types under the general term "image ambiguity". Thus, dual-ambiguity specifically denotes the situation where both the context and the image exhibit ambiguity.

3.2 Instruction Design

In order to ensure that model fully understands the context and image instead of making choices randomly, we organize our dataset in pairs. For the first six data types, the query can be represented as $(\mathcal{C}, \mathcal{Q}, \mathcal{I}_i)$. Each pair consists of an identical ambiguous context $\mathcal{C}$ and a question $\mathcal{Q}$, which are presented alongside i different unambiguous images $(\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_i)$.

For the dual-ambiguity data type, we manually group queries that look similar. Even within a pair, the context $\mathcal{C}_i$, the question $\mathcal{Q}_i$, and the image $\mathcal{I}_i$ may differ for each query instance. More formally, each pair can be represented as $(\mathcal{C}_i, \mathcal{Q}_i, \mathcal{I}_i)$.

Each $(\mathcal{C}, \mathcal{Q}, \mathcal{I})$ is independently input into MLLMs as a query without being influenced by other queries in the same pair. MLLMs receive i queries in the same pair independently and produce their outputs $(\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_i)$. These outputs are evaluated by comparing them with ground truth answers $(\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_i)$.

3.3 Data Collection

In this section, we detail how we construct this benchmark by using a three-step collection process.

Context Collection. We manually collected ambiguous contexts which can only be resolved with external images. Most of the contexts come from the Internet, while others are created manually. These contexts span three languages: Chinese, English, and Malay. We meticulously reviewed every context to exclude any that were unambiguous. In total, we collected 1278 contexts. The first six types comprise 906 contexts, and the remaining 372 contexts belong to the dual-ambiguity type.

Design of Questions, Images and Answers. For each context, we manually wrote questions and answers. As for images, we also manually collected them from the Internet or designed them ourselves. Specifically, for the dual-ambiguity type, the majority of images are sourced from CODIS (Luo et al., 2024), a benchmark constructed using ambiguous images. The data are compiled with following rules:

(1) Questions are designed to target ambiguous aspects within the contexts. Disambiguation of these contexts is not possible without the inclusion of external images.

(2) For each context and question, every unique image associated with them should lead to a distinct interpretation of the context, resulting in a unique answer per image. Crucially, the answer cannot be determined from the image or the context in isolation. The answer can be determined only when we give MLLMs the query in the format $(\mathcal{C}, \mathcal{Q}, \mathcal{I})$.

(3) To balance the performance and evaluation efficiency, MLLMs were required to generate outputs following a specific template with clear, predefined options. This method ensures the objectivity of responses and facilitates their efficient evaluation.

Data Verification. Five annotators participated in this process. To ensure dataset quality, each submission was cross-checked by the remaining four annotators. Data were retained only if they satisfied the following conditions: (1) correctness, (2) distinctiveness from existing data, and (3) compliance with all predefined criteria. Submissions that did not meet these conditions were returned to the annotator for revision.

Finally, our benchmark comprises 1278 queries and 501 $(\mathcal{C}, \mathcal{Q}, \mathcal{I})$ pairs, categorized into seven types. Figure 3 visually represents how the categories and scenarios are distributed.

3.4 Evaluation Metrics

For the k-th pair of queries, we decide to use $(\mathcal{O}_{k1}, \mathcal{O}_{k2}, \ldots, \mathcal{O}_{ki})$ to represent the model's outputs of a pair, and $(\mathcal{A}_{k1}, \mathcal{A}_{k2}, \ldots, \mathcal{A}_{ki})$ for the corresponding ground truth answers. We express the

(a) Without confusing options

Model	Polysemy Acc$_p$	Polysemy Acc$_q$	Homonymy Acc$_p$	Homonymy Acc$_q$	Grammar Acc$_p$	Grammar Acc$_q$	Semantics Acc$_p$	Semantics Acc$_q$	Specialized Acc$_p$	Specialized Acc$_q$	Cultural Acc$_p$	Cultural Acc$_q$	Dual-ambiguity Acc$_p$	Dual-ambiguity Acc$_q$	Overall Acc$_p$	Overall Acc$_q$
API-based Models																
GPT-4V	35.23	55.87	32.48	53.20	**43.93**	**65.89**	37.21	60.71	29.55	60.00	19.05	38.82	20.62	42.74	30.39	52.70
GPT-4o	**36.36**	**61.45**	**44.44**	**64.00**	41.12	65.42	**44.19**	**64.29**	**47.73**	**70.53**	**28.57**	**47.06**	21.65	**53.23**	**34.96**	**60.13**
Gemini-2.0-flash	19.48	52.87	38.18	63.14	17.78	50.56	19.51	50.00	30.77	63.53	25.00	46.91	12.36	36.16	23.46	52.11
Claude-3.5-Sonnet	3.41	16.76	2.56	16.00	5.61	21.03	0.00	14.29	6.82	17.89	0.00	5.88	**23.28**	51.10	9.37	26.36
Open-source Models > 7B																
Kimi-VL	3.37	36.31	5.13	46.40	5.56	43.46	2.38	40.48	2.27	36.84	0.00	47.06	14.08	52.27	4.74	45.15
Llama-3.2-Vision-11B	33.71	63.69	37.61	66.80	25.93	59.35	28.57	57.14	27.27	63.16	19.51	58.82	14.08	52.27	28.09	59.45
MiniCPM-o 2.6	38.20	66.48	38.46	66.40	**33.33**	63.55	28.57	60.71	**40.91**	**69.47**	**55.29**	29.27	4.23	20.53	34.00	53.10
Idefics3-8B-Llama3	39.33	67.04	37.31	62.80	31.48	62.62	35.71	65.48	31.32	56.84	24.09	56.47	19.72	49.87	32.38	58.91
InternVL2-8B	20.24	49.56	23.76	53.67	10.29	40.44	23.81	47.62	20.00	47.13	21.95	48.24	20.97	47.58	20.08	48.16
InternVL2.5-8B-MPO	42.26	69.32	43.56	68.81	25.00	60.29	**42.86**	**69.05**	37.50	58.62	19.51	57.65	24.19	51.88	35.73	61.68
InternVL2.5-8B-MPO-AWQ	**42.86**	**69.91**	**46.53**	**70.18**	30.88	**64.71**	33.33	61.90	37.50	60.92	21.95	**60.00**	**25.81**	**54.30**	**37.32**	**63.32**
Open-source Models ≤ 7B																
Deepseek-VL-Tiny	0.00	44.69	5.98	51.20	0.00	48.60	2.33	48.81	6.82	47.37	0.00	49.41	16.49	**54.03**	6.77	50.12
LLaVA-v1.6-vicuna-7b	15.73	54.19	21.37	54.80	14.81	54.21	21.43	55.95	11.36	42.11	12.20	51.76	16.90	50.40	16.46	52.26
LLaVA-v1.6-mistral-7b	19.11	55.87	23.08	58.00	12.96	54.67	16.67	58.33	15.91	49.47	2.44	47.06	16.90	50.40	16.26	53.60
Qwen2.5-VL-3B-Instruct	33.93	64.60	27.72	56.88	17.65	53.68	19.05	50.00	22.50	56.32	12.20	55.29	20.97	40.86	25.55	53.56
Qwen2.5-VL-7B-Instruct	**34.52**	60.47	32.67	60.09	29.41	58.09	**33.33**	50.00	**35.00**	**63.22**	14.63	54.12	19.35	49.19	**29.94**	56.29
mPLUG-Owl3-7B-240728	32.74	63.42	35.64	63.30	22.06	58.82	28.57	61.90	22.50	55.17	**48.24**	14.63	**27.42**	52.96	28.74	58.25
mPLUG-Owl3-2B-241014	33.73	**65.68**	36.08	64.71	27.47	**62.09**	30.95	**62.20**	30.00	60.92	10.00	55.74	16.49	51.61	25.65	**59.29**
LLaVA-OneVision	30.33	61.45	**38.46**	**64.80**	**30.56**	61.22	28.57	58.33	27.27	55.79	21.95	**60.00**	9.86	47.20	28.37	57.20

(b) With confusing options

Model	Polysemy Acc$_p$	Polysemy Acc$_q$	Homonymy Acc$_p$	Homonymy Acc$_q$	Grammar Acc$_p$	Grammar Acc$_q$	Semantics Acc$_p$	Semantics Acc$_q$	Specialized Acc$_p$	Specialized Acc$_q$	Cultural Acc$_p$	Cultural Acc$_q$	Dual-ambiguity Acc$_p$	Dual-ambiguity Acc$_q$	Overall Acc$_p$	Overall Acc$_q$
API-based Models																
GPT-4V	26.51	47.93	20.62	41.18	26.37	52.20	21.43	45.12	20.00	47.13	6.67	26.23	15.46	30.65	19.93	40.45
GPT-4o	**32.53**	**50.30**	**25.77**	**45.10**	**31.87**	**56.04**	**23.81**	**46.34**	**22.50**	**51.72**	**13.33**	29.51	**17.53**	**36.29**	**23.92**	**44.51**
Gemini-2.0-flash	8.06	36.80	8.82	32.41	14.06	44.53	8.70	40.91	4.76	25.58	6.67	**32.26**	11.02	31.23	10.00	34.85
Claude-3.5-Sonnet	3.61	22.49	4.12	22.55	4.40	23.08	2.38	20.73	2.50	27.59	3.33	13.11	15.59	35.64	7.56	26.50
Open-source Models > 7B																
Kimi-VL	1.13	31.84	1.71	39.60	1.85	40.19	0.00	36.90	0.00	32.63	0.00	41.18	5.71	18.82	1.50	31.98
Llama-3.2-Vision-11B	21.35	56.98	37.61	**63.60**	13.89	43.93	14.29	45.24	11.37	48.42	17.07	**54.12**	5.71	34.67	19.87	48.00
MiniCPM-o 2.6	20.22	53.07	34.18	60.40	28.70	58.41	30.95	59.52	18.18	60.00	14.63	50.58	2.86	12.63	23.50	44.40
Idefics3-8B-Llama3	29.21	58.10	**39.32**	62.80	24.07	55.14	**35.71**	**60.71**	**31.82**	**62.11**	17.07	51.76	12.86	**50.54**	28.14	56.37
InternVL2-8B	13.48	35.20	8.55	33.20	7.41	34.58	7.14	36.90	20.45	41.05	7.32	36.47	8.06	31.45	9.78	34.24
InternVL2.5-8B-MPO	26.97	58.10	27.35	54.80	27.78	56.54	28.57	58.33	25.00	52.63	**19.51**	51.76	20.97	46.51	25.95	53.01
InternVL2.5-8B-MPO-AWQ	**31.46**	**63.69**	33.33	59.20	**34.26**	**62.62**	30.95	59.52	31.81	57.89	14.63	52.94	**22.58**	50.27	**29.94**	**57.31**
Open-source Models ≤ 7B																
Deepseek-VL-Tiny	0.00	21.30	0.00	21.08	0.00	30.77	0.00	24.39	5.00	31.03	0.00	27.87	9.28	35.48	3.47	28.61
LLaVA-v1.6-vicuna-7b	13.48	51.95	19.66	58.00	14.81	56.07	19.05	54.76	11.36	49.47	14.63	**55.29**	11.43	46.77	15.19	52.54
LLaVA-v1.6-mistral-7b	19.10	57.54	23.93	59.20	12.96	54.21	26.19	**69.91**	18.18	51.58	2.44	49.41	14.29	47.58	17.14	53.71
Qwen2.5-VL-3B-Instruct	25.84	56.42	28.21	58.00	25.00	57.94	19.05	53.57	11.36	50.53	14.63	42.35	**24.19**	45.43	22.95	52.23
Qwen2.5-VL-7B-Instruct	13.48	42.46	19.66	50.80	27.78	55.61	26.19	51.19	27.27	60.00	9.76	44.71	20.97	**48.66**	20.76	50.12
mPLUG-Owl3-7B-240728	21.35	56.42	30.77	60.00	22.22	56.54	26.19	59.52	18.18	53.68	9.76	43.53	**24.19**	47.85	23.35	53.79
mPLUG-Owl3-2B-241014	24.10	53.85	30.93	60.78	25.27	59.34	**33.33**	58.54	**32.50**	**60.92**	**16.67**	49.18	13.40	46.77	22.70	54.28
LLaVA-OneVision	**40.44**	**68.16**	**39.32**	**65.60**	**34.26**	**63.08**	**33.33**	64.29	31.82	60.00	**17.07**	**55.29**	8.57	41.67	**31.40**	**57.39**
Human																
Human	75.00	86.59	73.50	86.00	70.09	85.05	78.57	89.29	79.55	89.36	68.29	83.53	81.43	87.12	74.66	86.42

Table 2: Results of MLLMs on MUCAR benchmark under two settings: (top) without confusing options and (bottom) with confusing options. For humans, the two experimental settings show little difference, so we chose the more challenging **with confusing options** setting for human evaluation.

evaluation of these model outputs as follows:

$$\text{Eval}(\mathcal{O}_{ki}) = \begin{cases} 1 & \text{if } \mathcal{O}_{ki} \text{ matches } \mathcal{A}_{ki} \\ 0 & \text{otherwise} \end{cases}, i \in \mathbb{Z}^+$$

Following Fu et al. (2023a), our evaluation utilizes two metrics, pair-wise accuracy Acc_p and query-wise accuracy Acc_q, these metrics can be calculated as follows:

$$\text{Acc}_p = \frac{1}{n_p} \sum_{k=1}^{n_p} \prod_{i=1}^{n_k} \text{Eval}(\mathcal{O}_{ki}),$$

$$\text{Acc}_q = \frac{1}{n_q} \sum_{k=1}^{n_p} \sum_{i=1}^{n_k} \text{Eval}(\mathcal{O}_{ki}).$$

where n_k denotes the number of queries within each pair, n_p represents the total number of pairs, and n_q is the total number of individual queries. Acc_p denotes the accuracy of judging each individual query's correctness independently. Acc_q requires that a pair is considered correct only if MLLMs correctly judge all queries within the pair.

4 Experiments

4.1 Models

We evaluate a total of 19 models covering a range of scales and architectures. Our evaluated proprietary models include GPT-4V (OpenAI, 2023), GPT-4o (OpenAI, 2024), Gemini (Gemini Team et al., 2023), and Claude-3.5-Sonnet (Anthropic, 2024). For open-source models, we include Deepseek-VL-Tiny (Lu et al., 2024), Kimi-VL (Team et al., 2025), Llama-3.2-Vision-11B (Meta, 2024), MiniCPM-o 2.6 (Yao et al., 2024), InternVL2.5 series (Wang et al., 2024b; Chen et al., 2024c), LLaVA-v1.6-vicuna-7b (Liu et al., 2024), Qwen2.5-VL series (Bai et al., 2025). Details of these models are listed in Table 11 in Appendix D.

4.2 Main Results

Main experimental results on our benchmark of all 19 models are reported in Table 2.

Overall Performance. Across all evaluated models, InternVL2.5-8B-MPO-AWQ achieves the best overall accuracy ($Acc_q = 63.32\%$), followed closely by InternVL2.5-8B-MPO (61.68%) and MiniCPM-o 2.6 (59.45%). Among proprietary models, GPT-4o outperforms the others, obtaining an overall accuracy of 60.13%, slightly higher than GPT-4V (52.70%). In contrast, Claude-3.5-Sonnet and Kimi-VL underperform, showing limited ability in disambiguation tasks.

Results with Different Model Size. Open-source models with scales *larger than 7B* generally outperform smaller ones, with all top-performing models falling within this range of scale, which is likely to benefit from richer training data and more advanced architectures. In comparison, models *smaller than or equal to 7B* show a clear performance gap. Although certain models, such as LaVA-Onevision (57.20%) and mPLUG-Owl2-2B (59.29%), perform competitively, most smaller models struggle with complex ambiguities, particularly in semantic and cultural contexts.

Results on Different Categories. We further break down the results by disambiguation categories, and find that InternVL2.5-8B-MPO-AWQ consistently leads in most categories, especially in Homonymy (70.18%), Grammar (64.71%), and Semantics (69.05%). Notably, MiniCPM-o 2.6 excels in the Specialized category (69.47%), suggesting domain knowledge plays a key role. In the Cultural category, which requires understanding cross-cultural references, models like GPT-4o (47.06%) and InternVL2.5-8B-MPO-AWQ (60.00%) show relatively stronger performance. On the other hand, most models perform poorly in the Polysemy and Dual-Ambiguity categories, reflecting the inherent challenges in resolving subtle or cross-modal ambiguities.

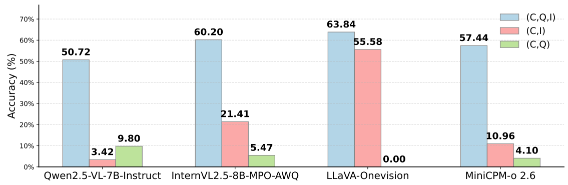


Figure 4: Accuracy under different input settings. $\mathcal{C}$ = Context, $\mathcal{Q}$ = Question, $\mathcal{I}$ = Image. For example, $(\mathcal{C}, \mathcal{Q}, \mathcal{I})$ includes all three; $(\mathcal{C}, \mathcal{Q})$ and $(\mathcal{C}, \mathcal{I})$ include only the specified components.

Model	Overall Acc_p	Overall Acc_q
Qwen2.5-VL-7B-Instruct (w/)	20.73	50.72
Qwen2.5-VL-7B-Instruct (w/o)	7.06 (↓13.67)	12.57 (↓38.15)
InternVL2.5-8B-MPO-AWQ (w/)	30.98	60.20
InternVL2.5-8B-MPO-AWQ (w/o)	4.56 (↓26.42)	8.38 (↓51.82)
LLaVA-OneVision (w/)	34.62	63.84
LLaVA-OneVision (w/o)	0.00 (↓34.62)	0.66 (↓63.18)
MiniCPM-o 2.6 (w/)	26.42	57.44
MiniCPM-o 2.6 (w/o)	1.14 (↓25.28)	4.63 (↓52.81)

Table 3: Ablation study: Only input $(\mathcal{Q}, \mathcal{I})$, with confusion options. Performance drop (↓) indicates the gap compared to full input.

In summary, model size and architecture significantly affect cross-modal disambiguation performance. Larger models and instruction-finetuned models usually demonstrate better generalization. The disparity across categories reveals the diverse challenges in context-dependent reasoning, particularly in categories involving semantic, cultural, or compound ambiguities.

5 Analysis and Discussion

5.1 Ablation Study

Figure 4 reports the accuracy of four models under three input settings: $(\mathcal{C}, \mathcal{Q}, \mathcal{I})$, $(\mathcal{C}, \mathcal{I})$, and $(\mathcal{C}, \mathcal{Q})$. All models achieve the highest accuracy with full input, with LLaVA-OneVision reaching 63.84%, followed by InternVL2.5 (60.20%), MiniCPM-o 2.6 (57.44%), and Qwen2.5 (50.72%). Removing the question while keeping context and image $(\mathcal{C}, \mathcal{I})$ results in moderate drops—for example, LLaVA drops to 55.58%, and InternVL2.5 to 21.41%. In contrast, removing the image $(\mathcal{C}, \mathcal{Q})$ leads to drastic degradation: LLaVA drops to 0.00%, and Qwen2.5 to 3.42%. This highlights the essential role of image information in resolving multimodal ambiguity.

Table 3 presents the ablation results using only question and image inputs, with confusing options included. All models exhibit significant performance degradation when context is removed. For example, LLaVA-OneVision drops from 63.84% to 0.66% in Acc_q (↓63.18), and from 34.62% to 0.00% in Acc_p (↓34.62). Similar trends are observed for InternVL2.5 and MiniCPM-o, which also suffer large drops in both metrics. These results underscore the importance of contextual information in resolving ambiguity, especially in the presence of visually or semantically confusing alternatives.

5.2 Discussion

As shown in Figure 5, this example illustrates how the interpretation of the phrase *"666"* is highly dependent on cultural and visual context, highlighting the necessity of cross-modal disambiguation. **Scenario 1:** The accompanying image shows the Forbidden City in Beijing, indicating a modern Chinese cultural context. In this setting, "666" is widely used as internet slang to express praise, meaning "awesome" or "skillful." *Answer: Positive*. **Scenario 2:** The image depicts a European Gothic cathedral—Notre-Dame de Paris—evoking a Western Christian context. Here, "666" is traditionally associated with the "number of the beast" from the Bible, conveying a negative connotation. *Answer: Negative*.

5.3 Further Exploration: An Agent-Based Framework for Ambiguity Resolution

To better address the cross-modal ambiguity resolution, we propose an agent-based framework, as shown in Figure 5. First, the model takes the task description, question, and ambiguous context/image as input to identify the ambiguity and its significance. Second, it generates contextual evidence to resolve the ambiguity. Finally, the model bridges the context and image through logical reasoning to produce the final answer.

Formally, given a task description $\mathcal{D}$, a question $\mathcal{Q}$, and a multimodal ambiguous context $\mathcal{X}$ (e.g., an image and text), the agent-based model $\mathcal{M}$ solves the task through a three-step process as illustrated in Figure 5:

Step 1: Ambiguity Detection. The model first detects the ambiguity and explains its significance:

$$\mathcal{A} = \mathcal{M}(\mathcal{D} \oplus \mathcal{Q} \oplus \mathcal{X} \oplus \mathcal{T}_1),$$

where $\mathcal{T}_1$ is a prompt guiding the model to identify potential ambiguity and why it matters, and $\mathcal{A}$ denotes the ambiguity explanation.

Step 2: Contextual Evidence Extraction. The model then extracts relevant contextual evidence to

Figure 5: Pipeline of our proposed framework. We first identify ambiguity and its significance from the task description, question, and context/image. Then we generate contextual evidence to resolve the ambiguity. Finally, we bridge the context and image through reasoning to produce the final answer.

resolve the ambiguity:

$$\mathcal{E} = \mathcal{M}(\mathcal{D} \oplus \mathcal{Q} \oplus \mathcal{X} \oplus \mathcal{A} \oplus \mathcal{T}_2),$$

where $\mathcal{T}_2$ instructs the model to generate explanatory evidence from the context, and $\mathcal{E}$ denotes the extracted evidence.

Step 3: Logical Reasoning and Final Answer. Finally, the model uses evidence to logically align context and image, and generate the final answer:

$$\mathcal{R} = \mathcal{M}(\mathcal{D} \oplus \mathcal{Q} \oplus \mathcal{X} \oplus \mathcal{E} \oplus \mathcal{T}_3),$$

where $\mathcal{T}_3$ prompts the model to conduct reasoning and provide the answer $\mathcal{R}$.

Unified Prompt Alternatively, the entire process can be completed with a unified prompt:

$$\mathcal{A}, \mathcal{E}, \mathcal{R} = \mathcal{M}(\mathcal{D} \oplus \mathcal{Q} \oplus \mathcal{X} \oplus \mathcal{T}),$$

where $\mathcal{T} = \mathcal{T}_1 \oplus \mathcal{T}_2 \oplus \mathcal{T}_3$.

To demonstrate the generality of our framework across different types of ambiguity, we also report performance of our framework on other wide discussed benchmarks, including CODIS (Luo et al., 2024) and MMA (Wang et al., 2024a). As shown in Table 4, our method consistently outperforms all baselines across benchmarks. Compared to CoT

Model	CODIS		MMA	MUCAR		Overall	
	Acc_p	Acc_q	Acc	Acc_p	Acc_q	Acc_p	Acc_q
Vanilla	36.26	59.49	72.0	32.35	53.91	34.31	54.61
CoT	36.81	60.76	68.0	28.57	56.41	32.69	52.44
CODIS	36.80	60.30	71.0	28.29	56.96	32.55	53.19
OURS	**42.49**	**63.46**	**84.0**	**44.87**	**66.78**	**43.68**	**64.11**

Table 4: Results of our method on CODIS, MMA and MUCAR based on `GPT-4o-2024-11-20`.

and CODIS-specific prompting, our approach is not only simple but also effective, demonstrating strong potential across different types of ambiguity.

6 Conclusion

We present MUCAR, a benchmark designed to evaluate MLLMs in resolving ambiguities across visual, textual, and multilingual contexts. Unlike prior benchmarks, MUCAR targets cross-modal disambiguation through multilingual text and image-text ambiguity cases. Evaluation of 19 state-of-the-art MLLMs reveals a clear gap from human-level performance, highlighting the need for more context-aware and cross-modally grounded models. We also introduce a simple agent-based framework that improves disambiguation through explicit reasoning. MUCAR aims to guide future research toward more robust and interpretable multimodal systems in this direction.

Limitations

While MUCAR offers a novel and rigorous benchmark for multilingual cross-modal ambiguity resolution, it has several limitations. It covers only three languages, limiting generalizability to low-resource or typologically diverse languages. The curated examples may not capture the complexity and noise of real-world multimodal data. Its partial reliance on GPT-based evaluation introduces potential biases, and the agent-based reasoning framework, though effective in structured tasks, may struggle with open-ended scenarios. Future work should address broader linguistic coverage, real-world settings, and more robust reasoning methods.

Ethics Statement

We ensured that all images in our dataset comply with ethical and legal standards. A small portion of images were manually synthesized to cover rare scenarios, and 0.5% were personally collected; both underwent strict quality checks to minimize bias. The remainder were obtained from platforms with free usage rights (e.g., Unsplash, Pexels, Pixabay). All images were carefully reviewed to ensure quality, fairness, and copyright compliance. Given their high quality and small proportion, synthetic images are unlikely to affect evaluation outcomes.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62306161, 62276152, 62236011).

References

Anthropic. 2024. Claude 3: A new generation of ai. Accessed on May 5, 2025.

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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are we on the right way for evaluating large vision-language models?

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Unified hallucination detection for multimodal large language models.

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*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024d. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

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

Jiwan Chung, Seungwon Lim, Jaehyun Jeon, Seungbeen Lee, and Youngjae Yu. 2024. Can visual language models resolve textual ambiguity with visual cues? let visual puns tell you! In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2452–2469, Miami, Florida, USA. Association for Computational Linguistics.

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*.

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.

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*.

Lingyu Gao, Aditi Chaudhary, Krishna Srinivasan, Kazuma Hashimoto, Karthik Raman, and Michael Bendersky. 2024. Ambiguity-aware in-context learning with large language models.

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*.

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.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions.

Youngwon Lee, Seung won Hwang, Ruofan Wu, Feng Yan, Danmei Xu, Moutasem Akkad, Zhewei Yao, and Yuxiong He. 2025. Agentic verification for ambiguous query disambiguation.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer.

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

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, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning.

Yang Liu, Guanbin Li, and Liang Lin. 2023b. 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. 2023c. MM-Bench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding.

Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, Peng Li, Ning Ma, Maosong Sun, and Yang Liu. 2024. CODIS: Benchmarking context-dependent visual comprehension for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10639–10659, Bangkok, Thailand. Association for Computational Linguistics.

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.

Xinyu Ma, Xuebo Liu, Derek F Wong, Jun Rao, Bei Li, Liang Ding, Lidia S Chao, Dacheng Tao, and Min Zhang. 2024. 3AM: An ambiguity-aware multi-modal machine translation dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1–13, Torino, Italia. ELRA and ICCL.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Heejeong Nam, Jinwoo Ahn, Keummin Ka, Jiwan Chung, and Youngjae Yu. 2024. VAGUE: Visual Contexts Clarify Ambiguous Expressions. *arXiv e-prints*, page arXiv:2411.14137.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

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

OpenAI. 2024. Hello GPT-4o. Accessed on May 5, 2025.

Mohammadmostafa Rostamkhani, Baktash Ansari, Hoorieh Sabzevari, Farzan Rahmani, and Sauleh Eetemadi. 2024. Illusory vqa: Benchmarking and enhancing multimodal models on visual illusions.

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.

Weiwei Sun, Hengyi Cai, Hongshen Chen, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2023. Answering ambiguous questions via iterative prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7669–7683, Toronto, Canada. Association for Computational Linguistics.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. 2025. Kimi-VL technical report.

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.

R. Wang, S. Song, L. Ding, S. S. Gu, M. Gong, Y. Iwasawa, and J. Guo. 2024a. Mma: Benchmarking multimodal large language model in ambiguity contexts. *preprint openreview*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

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.

A Dataset Details

A.1 Dataset Distribution

Category	Chinese	English	Malay	Total
Polysemy	63	64	52	179
Homonymy	42	176	32	250
Grammar	64	72	78	214
Semantics	24	18	42	84
Specialized	26	60	8	94
Cultural	24	66	0	90
Dual-Amb	126	126	120	372
Overall	**364**	**582**	**332**	**1278**

Table 5: Distribution of ambiguity categories across different languages.

MUCAR Table 5 summarizes the distribution of seven ambiguity types across **Chinese**, **English**, and **Malay**. **English** has the most annotated instances (582), followed by **Chinese** (364) and **Malay** (332). *Homonymy* is notably dominant in English, while *Dual-Ambiguity* remains consistently high across all languages. *Cultural* ambiguity appears in Chinese and English but is absent in Malay. *Grammar* and *Polysemy* are relatively balanced, whereas *Specialized* and *Semantics* vary more significantly. These trends reflect both shared and language-specific ambiguity patterns.

Prompt for GPT-4o evaluation.
I'll give you an image. Please answer my question based on the image. Directly select the correct option (A, B, C, D, or E). Use the following format to answer: Answer: [ONLY the option letter; not a complete sentence] Only give me the reply according to this format, don't give me any other words. Now, please answer this question. Question: [QUESTION HERE] Options: [OPTIONS HERE]

Table 6: Prompt for GPT-4o evaluation.

Model	Consistency Score
GPT-4o	96.25
Gemini-2.0-flash	96.25
Qwen2.5-VL-7B-Instruct	97.35
InternVL2.5-8B-MPO-AWQ	97.57
mPLUG-Owl3-7B-240728	95.26

Table 7: Consistency results of different models.

A.2 Information about Annotators

In our study, the annotators are the co-authors of the paper. We opted for this arrangement to ensure careful and consistent evaluation, given the nuanced nature of the tasks involving ambiguity and cross-linguistic interpretation. Below, we provide detailed information about the annotators, including their linguistic fluency, country/region of residence, professional background, and gender. Table 8 shows the information about annotators. It is important to note that all the images used in the dataset were carefully selected and manually reviewed. The selection process involved 8 data curators, who cross-checked the images to ensure quality and relevance. Only those images that received unanimous approval from all curators were included in the dataset. This approach ensured that the images met the necessary standards for consistency and quality.

A.3 Potential Risks Analysis

As for the details of data collection, we estimate that approximately 2.3% of the images are AI-generated and 0.5% of the images are taken personally, while the rest are crawled from online sources (e.g., Unsplash, Pexels, Pixabay).

Regarding the potential impact of synthetic images on the evaluation of multimodal ambiguity resolution, all synthetic images were carefully crafted to closely resemble real-world scenes and passed strict manual quality checks. They were designed to supplement rare or hard-to-collect cases, enhancing benchmark coverage without introducing bias. Given their high quality and small proportion in the dataset, any potential impact on model performance evaluation is expected to be negligible.

B Human Evaluation

B.1 Results of Human Evaluation

We have conducted a more comprehensive human evaluation. Specifically, we conducted evaluations with 3 human evaluators for Chinese and English, and 2 human evaluators for Malay. The final evaluation scores were averaged across the evaluators for each language. These new results provided a concrete basis for comparison against the current MLLMs performance. Table 9 is the full breakdown of human performance across various categories. For humans, the two experimental settings show little difference, so we chose the more challenging with confusing options setting for human

Name	Gender	Fluent Languages	Country/Region	Professional Background
author1	Male	Chinese, English	China	Computer Science, Telecommunications, Systems Science
author2	Female	English, Chinese	England	Computer Science, Telecommunications, Finance, Law
author3	Female	Malay, English, Chinese	Malaysia	Computer Science
author4	Male	Chinese, English	China	Computer Science
author5	Female	Chinese, English	China	Computer Science, Economics
author6	Female	Chinese, English	China	Computer Science, Economics, Statistics, Data Science, Humanities and Social Sciences
author7	Male	English, Chinese, Japanese	China	Computer Science, Data Science
author8	Male	Chinese, English	China	Computer Science, Statistics

Table 8: Information about Annotators

Model	Polysemy		Homonymy		Grammar		Semantics		Specialized		Cultural		Dual-ambiguity		Overall	
	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q
Human	75.00	86.59	73.50	86.00	70.09	85.05	78.57	89.29	79.55	89.36	68.29	83.53	81.43	87.12	74.66	86.42

Table 9: Human Evaluation Results. For humans, the two experimental settings show little difference, so we chose the more challenging **with confusing options** setting for human evaluation.

evaluation. As shown in the table, our benchmark features higher difficulty and a larger dataset. The evaluation scores of human evaluators are significantly higher than those of large models, which highlights a strong gap between human-level performance and current MLLMs. Specifically, the best overall $\mathrm{Acc_p}$ of current MLLMs is only 37.32, but human $\mathrm{Acc_p}$ reaches 74.66.

B.2 Consistency between GPT and Human Evaluation

Due to the diverse and often unpredictable nature of large model outputs, it is difficult to strictly evaluate them based on exact match criteria, as this approach would not always be accurate. Additionally, human evaluation is highly time-consuming and labor-intensive, especially when dealing with large datasets. So, we adopt GPT-based evaluation when assessing model performance.

To ensure the reliability of GPT-based evaluation, we conducted a consistency analysis between human and GPT judgments across multiple models by using prompt in Table 6. The results in Table 7 show high agreement rates, where the GPT-based binary judgments aligned with human evaluation in over 94% of the cases, supporting the feasibility of using GPT-based evaluation in place of manual evaluation.

C Prompt for Model Inference

Table 10 presents the detailed prompts used during model testing. In the main experiments, we employed English prompts; additionally, we conducted ablation studies using Chinese and Malay prompts to evaluate the impact of different evaluation languages on the experimental results. The table also lists the prompts used in the three ablation settings: $(\mathcal{Q}, \mathcal{I})$, $(\mathcal{C}, \mathcal{Q})$, and $(\mathcal{C})$, where $(\mathcal{Q}, \mathcal{I})$ uses the *Question and Image* as input, $(\mathcal{C}, \mathcal{Q})$ uses the *Context and Question*, and $(\mathcal{C})$ uses only the *Context*.

D Evaluated Models

We evaluate a total of 19 models covering a range of scales and architectures. Our evaluated proprietary models include GPT-4V (OpenAI, 2023), GPT-4o (OpenAI, 2024), Gemini (Gemini Team et al., 2023), and Claude-3.5-Sonnet (Anthropic, 2024). For open-source models, we include Deepseek-VL-Tiny (Lu et al., 2024), Kimi-VL (Team et al., 2025), Llama-3.2-Vision-11B (Meta, 2024), MiniCPM-o 2.6 (Yao et al., 2024), InternVL2-8B (Chen et al., 2024d), InternVL2.5-8B-MPO (Wang et al., 2024b), InternVL2.5-8B-MPO-AWQ (Chen et al., 2024c), LLaVA-v1.6-vicuna-7b (Liu et al., 2024), Qwen2.5-VL-3B-Instruct (Bai et al., 2025), Qwen2.5-VL-7B-Instruct (Bai et al., 2025). Details of these models are listed in Table 11. Table 11 presents a comprehensive overview of the Multimodal Large Language Models (MLLMs) evaluated in our benchmark. The models are categorized into two groups: API-based models and open-source models. For each model, we list its parameter size category (greater than or less than 7 billion), the vision encoder architecture, the underlying language model (LLM) backbone, and the employed vision-to-language (V2L) adapter. API-based models such as GPT-4V and Gemini do not publicly disclose architectural details, while open-source models span a variety of encoders (e.g., SigLIP, CLIP ViT, InternViT), LLM backbones (e.g., Llama, Qwen, InternLM), and adapter

Prompt for Model Inference	
Main experiment In English	I'll give you an image. Please answer my question based on the image. Directly select the correct option (A, B, C, D, or E). Use the following format to answer: Answer: [ONLY the option letter; not a complete sentence] Only give me the reply according to this format, don't give me any other words. Now, please answer this question. Question: [QUESTION HERE] Options: [OPTIONS HERE]
Ablation Study In Chinese	我会给你一张图片。请根据图片回答我的问题。直接选择正确选项(A, B, C, D, 或E)。请使用以下格式回答： Answer: [仅为选项字母；不是完整句子] 请只按照此格式回复我，不要给出任何其他文字。现在，请回答这个问题。 问题：[QUESTION HERE] 选项：[OPTIONS HERE]
Ablation Study In Malay	Saya akan berikan anda imej. Sila jawab soalan saya berdasarkan imej tersebut. Pilih terus pilihan yang betul (A, B, C, D, atau E). Gunakan format berikut untuk menjawab: Answer: [HANYA huruf pilihan; bukan ayat penuh] Berikan saya jawapan mengikut format ini sahaja, jangan berikan perkataan lain. Sekarang, sila jawab soalan ini. Soalan: [QUESTION HERE] Pilihan: [OPTIONS HERE]
Ablation Study Only input $(\mathcal{Q}, \mathcal{I})$	I'll give you an image. Please answer my question based on the image. Directly select the correct option (A, B, C, D, or E). Use the following format to answer: Answer: [ONLY the option letter; not a complete sentence] Only give me the reply according to this format, don't give me any other words. Now, please answer this question. Question: [QUESTION HERE] Options: [OPTIONS HERE]
Ablation Study Only input $(\mathcal{C}, \mathcal{Q})$	Please answer my question. Directly select the correct option (A, B, C, D, or E). Use the following format to answer: Answer: [ONLY the option letter; not a complete sentence] Only give me the reply according to this format, don't give me any other words. Now, please answer this question. Question: [QUESTION HERE] Options: [OPTIONS HERE]
Ablation Study Is $\mathcal{C}$ ambiguous?	Please determine whether this sentence is ambiguous. If the sentence is ambiguous, please answer 'Yes.'; otherwise, answer 'No.'. Please respond directly with 'yes' or 'no', without any additional content. Sentence: [SENTENCE HERE]

Table 10: Prompt for model inference.

types (e.g., MLP, Linear, XAttn LLM). This table highlights the diversity in architectural design choices across MLLMs.

Table 12 reports results for Chinese and Malay prompts under two settings: with and without confusing options. The four sections present detailed model performance for each language and setting combination.

As shown in Table 12, models achieve the highest performance without confusing options, with InternVL2.5-8B-MPO-AWQ reaching 65.82 for Chinese prompts and 65.27 for Malay prompts. When confusing options are introduced, overall accuracy drops noticeably: for Chinese prompts, the top score decreases by 7.72 (65.82 $\rightarrow$ 58.10), while for Malay prompts, the drop is even larger at 11.69 (65.27 $\rightarrow$ 53.58). This indicates that confusing options substantially increase task difficulty. Across both languages, Chinese prompts perform slightly better than Malay prompts, though the gap remains small (0.55 without confusing options and 4.41 with confusing options). InternVL2.5-8B-MPO-AWQ consistently achieves the best results across all settings.

E More Cases

We present a comprehensive set of additional cases from the MUCAR dataset to further illustrate the performance of Multimodal Large Language Models (MLLMs). Specifically, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, Figure 13 display model outputs across various scenarios. For clarity and balance, we select five

Model	Parameters	Vision Encoder	LLM Backbone	V2L Adapter
GPT-4V (OpenAI, 2023)	-	-	-	-
GPT-4o (OpenAI, 2024)		-	-	-
Gemini (Gemini Team et al., 2023)		-	-	-
Claude-3.5-Sonnet (Anthropic, 2024)		-	-	-
Deepseek-VL-Tiny (Lu et al., 2024)	$>$ 7B	SigLIP	DeepSeek LLM	MLP
Kimi-VL (Team et al., 2025)		MoonViT	Moonlight model	MLP
Llama-3.2-Vision-11B (Meta, 2024)		XAttn LLM	Llama 3.1	XAttn LLM
MiniCPM-o 2.6 (Yao et al., 2024)		SigLIP	Qwen2.5-7B	MLP
Idefics3-8B-Llama3 (Laurençon et al., 2024)		SigLIP	Llama-3.1-8B-Instruct	XAttn LLM
InternVL2-8B (Chen et al., 2024d)		InternViT	internLM2.5-7b-chat	MLP
InternVL2.5-8B-MPO (Wang et al., 2024b)		InternViT-V2.5	internLM2.5-7b-chat	MLP
InternVL2.5-8B-MPO-AWQ (Chen et al., 2024c)		InternViT-V2.5	internLM2.5-7b-chat	MLP
LLaVA-v1.6-vicuna-7b (Liu et al., 2024)	$\leq$ 7B	CLIP ViT-L	vicuna-7b-v1.5	MLP
LLaVA-v1.6-mistral-7b (Liu et al., 2024)		CLIP ViT-L	Mistral-7B-Instruct-v0.2	MLP
Qwen2.5-VL-3B-Instruct (Bai et al., 2025)		ViT	Qwen2.5 LLM	MLP
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)		ViT	Qwen2.5 LLM	MLP
mPLUG-Owl3-7B-240728 (Ye et al., 2024)		SigLIP	Qwen2 LLM	Linear
mPLUG-Owl3-2B-241014 (Ye et al., 2024)		SigLIP	Qwen2 LLM	Linear
LLaVA-OneVision (Li et al., 2024)		SigLIP	Qwen2 LLM	MLP

Table 11: API-based and open-source MLLMs selected for evaluation.

Model	Polysemy		Homonymy		Grammar		Semantics		Specialized		Cultural		Overall	
	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q
In Chinese, without confusing options														
Qwen2.5-VL-7B-Instruct	28.57	56.34	28.71	56.88	17.65	49.26	19.05	38.10	25.00	55.17	17.07	50.59	25.06	53.91
InternVL2.5-8B-MPO-AWQ	41.07	68.44	44.55	68.35	30.88	63.24	33.33	64.29	37.50	59.77	24.39	60.00	38.04	65.82
LLaVA-OneVision	24.72	55.31	29.91	54.00	17.59	50.93	23.81	50.00	22.73	47.37	14.63	51.76	22.78	52.26
MiniCPM-o 2.6	20.22	55.31	40.17	66.40	34.25	64.49	33.33	60.71	31.82	61.05	14.63	45.88	30.52	60.75
In Chinese, with confusing options														
Qwen2.5-VL-7B-Instruct	12.36	42.46	17.95	50.40	20.37	51.40	21.43	44.05	25.00	51.58	12.20	41.18	17.77	47.74
InternVL2.5-8B-MPO-AWQ	29.21	60.89	29.91	58.40	26.85	56.54	26.19	58.33	31.82	57.89	21.95	55.29	28.25	58.10
LLaVA-OneVision	7.87	44.13	23.93	50.80	12.96	50.93	11.90	44.05	13.64	43.16	7.32	45.88	14.35	47.63
MiniCPM-o 2.6	19.97	48.61	27.35	54.80	26.85	55.14	28.57	55.95	13.64	46.47	12.20	35.29	22.78	51.38
In Malay, without confusing options														
Qwen2.5-VL-7B-Instruct	35.71	63.13	30.69	61.01	25.00	55.88	28.57	54.76	37.50	62.07	9.76	51.76	30.30	59.98
InternVL2.5-8B-MPO-AWQ	39.88	68.14	36.63	66.06	26.47	62.50	28.57	61.90	35.00	60.92	26.83	62.35	34.85	65.27
LLaVA-OneVision	20.22	20.28	25.64	56.00	17.60	49.53	26.19	53.57	15.91	46.32	19.51	50.58	20.73	51.60
MiniCPM-o 2.6	33.71	52.60	41.03	66.80	37.96	64.95	28.57	55.95	27.27	62.11	14.63	48.24	33.49	62.29
In Malay, with confusing options														
Qwen2.5-VL-7B-Instruct	14.61	45.25	18.80	50.00	24.07	50.00	21.43	46.43	25.00	51.58	9.76	36.47	19.13	47.63
InternVL2.5-8B-MPO-AWQ	23.60	53.07	21.37	53.20	24.07	55.14	23.81	57.14	27.27	55.79	12.20	44.88	22.32	53.58
LLaVA-OneVision	15.73	46.93	27.35	56.00	15.74	51.41	21.43	52.38	1.37	40.00	12.20	51.76	18.45	50.72
MiniCPM-o 2.6	22.47	48.60	30.77	59.60	37.04	61.68	19.05	44.05	18.18	53.68	9.76	36.47	26.20	53.69

Table 12: Ablation study in four settings: Chinese and Malay prompts, each with and without confusing options.

representative cases from each of the seven predefined categories. In these visualizations, incorrect responses generated by the models are clearly marked in red to allow easy identification of errors.

To support a deeper understanding of the visual and contextual challenges within each case, we also include detailed explanations that highlight the key ambiguities present in the images. These annotations are intended to help readers recognize why a particular question might be difficult to answer correctly, either for a model or a human.

However, it is crucial to emphasize that these explanatory notes were not accessible to either the MLLMs or the human participants during the question-answering process. Both models and human volunteers provided their responses without the benefit of additional contextual guidance, ensuring a fair and unbiased assessment of performance. This setup allows us to isolate and better evaluate the inherent reasoning and perception capabilities of the models in comparison to human interpretation.

Overall, these additional cases offer further insight into the specific limitations and strengths of MLLMs when dealing with multimodal ambiguity, reinforcing the broader findings of our evaluation.

Q: 'Angkat bendera putih' merujuk kepada apa? ***(What does "raised white flag" mean?)***
C: Pasukan itu angkat bendera putih. ***(That army raised the white flag.)***

Scenario: 1

Answer: A.Mengaku kalah atau menyerah diri.
(Admit defeat or surrender.)

Scenario: 2

Answer: B.Mengibarkan bendera berwarna putih secara fizikal
(Physically raising a white flag.)

◆ **Scenario: 1**

Explanation: "Angkat bendera putih" **is a widely recognized idiom, particularly in a military or competitive context. It is an international symbol signifying surrender, admission of defeat, or giving up. This clearly implies that the team has admitted defeat or surrendered.**

❌ **InternVL2_5-8B-MPO-AWQ**: **B.**
✔ **InternVL2_5-8B-MPO-AWQ (with CoT): A.**
The image shows a group of people in camouflage outfits, likely engaged in a paintball game. They are positioned behind sandbags and a fence, with one person holding a white flag. In the context of paintball or similar games, raising a white flag typically indicates surrender or the end of the game. This is a common signal to show that the team is willing to concede or stop playing.\n\nAnswer: A

✔ **Qwen2.5-VL-7B-Instruct**: **A.**
✔ **Qwen2.5-VL-7B-Instruct (with CoT)**: **A.**
Dalam gambar ini, terlihat dua pasukan yang sedang bermain paintball di balik sandbag dan pagar besi. Salah satu pasukan mengangkat bendera putih. Dalam konteks permainan paintball, mengangkat bendera putih biasanya merupakan tindakan untuk menunjukkan bahwa mereka telah menyerah atau mengakui kekalahan.\n\nAnswer: A

◆ **Scenario: 2**

Explanation: "Angkat bendera putih" **can be interpreted as the physical act of raising a flag that happens to be white. It simply describes the action of lifting a white-colored piece of fabric on a pole, without any inherent implication of defeat or surrender.**

❌ **InternVL2_5-8B-MPO-AWQ**: **A.**
❌ **InternVL2_5-8B-MPO-AWQ (with CoT): A.**
The image shows a person holding up a white flag in an open field. The act of raising a white flag typically signifies surrender or a request for peace. This is a common symbol in various contexts, including sports, military, and other situations where a peaceful resolution is sought. The other option, B, refers to physically holding up a white flag, which is not the primary meaning of the phrase "angkat bendera putih."\n\nAnswer: A

❌**Qwen2.5-VL-7B-Instruct**: **A.**
✔ **Qwen2.5-VL-7B-Instruct (with CoT)**: **B.**
'Angkat bendera putih' dalam konteks ini merujuk kepada tindakan mengibarkan bendera berwarna putih secara fizikal. Ini adalah tindakan yang biasanya dilakukan untuk menunjukkan kekuatan atau semangat dalam suatu perayaan atau acara tertentu.n\nAnswer: B

Figure 6: A case of the category of polysemy in Malay.

Q: 他想表达什么？*(What is he trying to express?)*
C: "我家热得快炸了"。
(This context can be translated in two ways:
①My house is so hot it's about to explode. ②The electric heater in my house exploded.)

Scenario: 1

Answer: A.我家非常热。
(My house is so hot.)

Scenario: 2

Answer: B.我家的加热器发生故障了。
(There is a malfunction with the heater.)

◆ **Scenario: 1**

Explanation: "我家热得快炸了" can be punctuated in the way of "我家热得 | 快炸了", which is an exaggerated technique to describe "My house is so hot! " in Chinese.

✔ **InternVL2.5-8B-MPO-AWQ: A.**
✔ **InternVL2.5-8B-MPO-AWQ (with CoT): A.**
To determine what the person is expressing in the phrase \"我家热得快炸了\", we need to analyze the context and the literal meaning of the words. The phrase is in Chinese, and it translates to \"My house is so hot it's about to explode.\" The literal translation suggests extreme heat, which is a common way to express that a place is very hot. The phrase does not specifically mention a heating system or a malfunction, so it is more likely to be a general statement about the temperature rather than a technical issue.\n\nAnswer: A

✔ **Qwen2.5-VL-7B-Instruct: A.**
✔ **Qwen2.5-VL-7B-Instruct (with CoT): A.**
\"我家热得快炸了\"这句话的意思是"我家热得快要爆炸了"，这是一种夸张的说法，用来形容非常热。选项A"我家非常热"与这句话的意思相符；选项B"我家的加热器发生故障了"则不符合这句话的含义。\n\nAnswer: A

✔ **MiniCPM-o 2.6: A.**
To answer the question, let's break down the elements in the image and the context provided:\n\n1. The image shows a person standing in a forested area, with a background of lush greenery and a stream.\n2. The person is wearing a white t-shirt that appears to be wet, suggesting they might have been in water or sweating.\n3. The person is holding their head with one hand, which could indicate discomfort or distress.\n\nNow, let's analyze the options:\n\nA. \"我家非常热。\" - This translates to \"My house is very hot.\"\nB. \"我家的加热器发生故障了。\" - This translates to \"My heater is broken.\"\n\nGiven the context of the image:\n- The person is in a natural, outdoor setting, which is typically cooler than an indoor environment.\n- The person's wet t-shirt suggests they might be sweating or have been in water, which could be uncomfortable in a hot environment.\n\nConsidering these points, the most likely reason for the person's discomfort is the heat, rather than a broken heater.\n\nTherefore, the correct answer is:\n\nAnswer: A

◆ **Scenario: 2**

Explanation: "我家热得快炸了" can be punctuated in the way of "我家热得快 | 炸了". Here, "热得快" is a common electric heater. Therefore, this sentence can be interpreted as "My '热得快' exploded! "

❌ **InternVL2.5-8B-MPO-AWQ: A.**
❌ **InternVL2.5-8B-MPO-AWQ (with CoT): A.**
The image shows a bucket of water with steam rising from it, indicating that the water is hot. The phrase \"我家热得快炸了\" translates to \"My home is so hot it's about to explode.\" This suggests that the person is experiencing extreme heat in their home. The presence of steam and the context of the phrase imply that the heat is intense, but it does not necessarily mean that the heating device itself is malfunctioning.\n\nAnswer: A

✔ **Qwen2.5-VL-7B-Instruct: B.**
❌ **Qwen2.5-VL-7B-Instruct (with CoT): A.**
\"我家热得快炸了\"这句话的意思是"我家热得快要爆炸了"，这是一种夸张的说法，用来形容非常热。根据图片内容，可以看到一个装有水的容器中插着一根黄色的加热棒，水面上有蒸汽冒出，说明正在加热。因此，这句话可能是在描述加热器正在工作，导致家里变得非常热。\n\nAnswer: A

✔ **MiniCPM-o 2.6: B.**
To answer the question, let's break down the elements in the image and the context provided:\n\n1. The image shows a container with water and a device that appears to be a heating element.\n2. The device is submerged in water, and there is steam rising from the water.\n3. The text in the image is in Chinese: \"我家热得快炸了\" which translates to \"My home's instant heater is about to explode.\"\n\nNow, let's analyze the options:\n\nA. \"我家非常热。\" - This translates to \"My home is very hot.\" This option is not directly related to the image, as the image shows a heating device, not a hot environment.\n\nB. \"我家的加热器发生故障了。\" - This translates to \"My home's heater is malfunctioning.\" This option is more relevant to the image, as it suggests a problem with the heating device.\n\nGiven the context of the image and the text, the most appropriate answer is:\n\nAnswer: B

Figure 7: A case of the category of grammar in Chinese.

Q: Apa yang dibeli? ***(What was bought?)***
C: Saya beli nasi lemak sambal sotong. ***(I bought nasi lemak with squid sambal.)***

Scenario: 1

Answer: A.Hidangan nasi lemak yang berlauk sambal sotong.
(Nasi lemak dish with squid sambal..)

Scenario: 2

Answer: B.Nasi lemak dan juga sambal sotong.
(Nasi lemak and squid sambal separately.)

◆ **Scenario: 1**

Explanation: "nasi lemak sambal sotong" **is a very common way to describe a single, complete dish in Malay cuisine. In this context, "sambal sotong" functions as the specific side dish or accompaniment that comes with the nasi lemak.**

❌ **InternVL2_5-8B-MPO-AWQ**: **B.**
❌**InternVL2_5-8B-MPO-AWQ (with CoT): B.**
The image shows a plate of food with various components, including rice, boiled egg, cucumber slices, and two different types of side dishes. Based on the image, it is clear that the food items shown are nasi lemak and sambal sotong.\n6. Therefore, the correct option is B, which states that the person bought nasi lemak and sambal sotong.\n\nAnswer: B

✔ **Qwen2.5-VL-7B-Instruct**: **A.**
✔ **Qwen2.5-VL-7B-Instruct (with CoT)**: **A.**
The image shows a plate of nasi lemak, which is a traditional Malaysian dish consisting of rice cooked in coconut milk and pandan leaves, served with various side dishes. The image clearly shows a plate of nasi lemak with sambal sotong as one of the side dishes. Therefore, the correct interpretation is that the person bought a plate of nasi lemak with sambal sotong.\n\nAnswer: A

◆ **Scenario: 2**

Explanation: "nasi lemak sambal sotong" **can be interpreted as two separate items: "nasi lemak" (a dish of coconut milk rice) and "sambal sotong" (a side dish of squid cooked in chili paste) being bought individually.**

✔ **InternVL2_5-8B-MPO-AWQ**: **B.**
✔ **InternVL2_5-8B-MPO-AWQ (with CoT): B.**
The image shows a bowl of white rice, a bowl of dark sauce, and a bowl of what appears to be toasted bread. There is no indication of sambal sotong being a separate dish or item. The presence of the sauce and the bread suggests that the items are part of a meal, likely nasi lemak with sambal sotong as a condiment or side.\n\nAnswer: B

❌**Qwen2.5-VL-7B-Instruct**: **A.**
❌**Qwen2.5-VL-7B-Instruct (with CoT)**: **A.**
The image shows a bowl of rice, a bowl of what appears to be sambal sotong (a type of chili paste), and some bread slices with the same chili paste spread on them. The presence of the rice and the chili paste strongly indicates that the dish being referred to is nasi lemak with sambal sotong as one of its components.n\nAnswer: A

Figure 8: A case of the category of grammar in Malay.

Q: Siapakah yang berada di bank? ***(Who was at the bank?)***
C: Dia nampak pencuri di bank. ***(He saw a thief at the bank.)***

Scenario: 1

Answer: A.Pencuri itu berada di dalam atau di kawasan bank.
(The thief was in or around the bank.)

Scenario: 2

Answer: B.Dia berada di bank semasa melihat pencuri.
(He was at the bank while seeing the thief.)

◆ **Scenario: 1**

Explanation: "Dia nampak pencuri di bank" **directly states that the thief was located at the bank. The most direct answer, based on the explicit mention, is the thief. This interpretation focuses on the stated location of the thief.**

✔ **InternVL2_5-8B-MPO-AWQ**: **A.**
✔ **InternVL2_5-8B-MPO-AWQ (with CoT): A.**
To determine who is at the bank, we need to analyze the image carefully. The image shows a person dressed in black, wearing a mask and gloves, and appears to be attempting to steal from an ATM. This person is positioned in front of the ATM, indicating that they are physically present at the bank. The presence of the ATM and the person's actions suggest that they are inside or in the vicinity of the bank.\n\nAnswer: A

✔ **Qwen2.5-VL-7B-Instruct**: **A.**
✔ **Qwen2.5-VL-7B-Instruct (with CoT)**: **A.**
The image shows a thief at an ATM machine, which is typically located inside or near a bank. The thief is depicted in a way that suggests he is inside the bank premises, as he is using the ATM. There is no indication that the person observing the thief is also inside the bank. Therefore, the correct interpretation is that the thief is inside the bank.\n\nAnswer: A

◆ **Scenario: 2**

Explanation: "Dia nampak pencuri di bank" **implies the observer's location. For "He saw a thief at the bank" (to occur, the pronoun "He" must also have been present "at the bank" to perform the act of seeing the thief there.**

❌ **InternVL2_5-8B-MPO-AWQ**: **B.**
❌ **InternVL2_5-8B-MPO-AWQ (with CoT): B.**
To determine who was at the bank, we need to analyze the image carefully. The image shows a person dressed in dark clothing, running down a street. However, there is no direct indication that the person in the hoodie was at the bank or that they were involved in a bank robbery. The image does not provide any clear evidence of the person being inside the bank or at the bank during the time of the alleged robbery."\n\nAnswer: A

❌ **Qwen2.5-VL-7B-Instruct**: **B.**
✔ **Qwen2.5-VL-7B-Instruct (with CoT)**: **A.**
The image shows a person wearing a black hoodie and gloves, running away from the camera. The person in the foreground is not inside the bank but rather running away from it. Therefore, the correct interpretation is that the person is not inside the bank but rather outside, possibly after committing a crime.n\nAnswer: A

Figure 9: A case of the category of grammar in Malay.

Q: What is the subject in the sentence going to eat?
C: The chicken is ready to eat.

Scenario: 1

Answer: A.Chicken.

Scenario: 2

Answer: B.Chicken feed.

◆ Scenario: 1

Explanation: The chicken is cooked and prepared, so it is ready for someone to eat.

✔ **InternVL2.5-8B-MPO-AWQ: A.**
✔ **InternVL2.5-8B-MPO-AWQ (with CoT): A.**
The image shows a chicken inside a pot, which suggests that the chicken is being cooked. The sentence \"The chicken is ready to eat\" implies that the chicken has been prepared for consumption. Therefore, the subject in the sentence is likely to be the person or entity that is going to eat the chicken.\n\nAnswer: A

✔ **Qwen2.5-VL-7B-Instruct: A.**
✔ **Qwen2.5-VL-7B-Instruct (with CoT): A.**
The image shows a cooked chicken on a plate with steam rising from it, indicating that it is hot and ready to be eaten. The question asks what the subject in the sentence is going to eat. Since the subject is implied to be a person who would eat the chicken, the correct answer should refer to the chicken itself as the food item being consumed.\n\nAnswer: A

✔ **MiniCPM-o 2.6: A.**
❗ *We found that MiniCPM-o 2.6's instruction-following ability isn't very good, so it didn't provide an explanation here.*

◆ Scenario: 2

Explanation: The chicken itself is hungry and ready to eat something.

❌ **InternVL2.5-8B-MPO-AWQ: A.**
❌ **InternVL2.5-8B-MPO-AWQ (with CoT): A.**
The image shows a chicken sitting at a table with a plate, fork, and knife in front of it. The chicken appears to be ready to eat, suggesting that it is the subject of the sentence. The options provided are \"A. Chicken.\" and \"B. Chicken feed.\" Since the chicken is the one that is ready to eat, the subject in the sentence is the chicken itself.\n\nAnswer: A

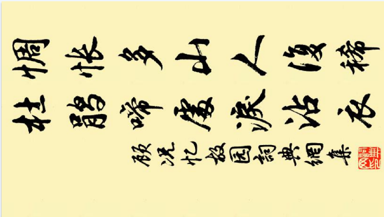
❌ **Qwen2.5-VL-7B-Instruct: A.**
❌ **Qwen2.5-VL-7B-Instruct (with CoT): A.**
The image shows a cartoon chicken sitting at a table with a plate and utensils, suggesting that the chicken is about to eat. The question asks what the subject (the chicken) is going to eat. Since there is no indication of the chicken eating anything other than food, the logical conclusion is that the chicken is going to eat something.\n\nAnswer: A

❌ **MiniCPM-o 2.6: A.**
❗ *We found that MiniCPM-o 2.6's instruction-following ability isn't very good, so it didn't provide an explanation here.*

Figure 10: A case of the category of semantics in English.

Q: 这是指植物还是动物? *(Does this refer to a plant or an animal?)*
C: 山上有许多杜鹃。*(There are many 杜鹃 on the mountain.)*

Scenario: 1



Answer: A.动物。
(An animal.)

Scenario: 2

Answer: B.植物。
(A plant.)

◆ **Scenario: 1**

Explanation: In the sentence "山上有许多杜鹃", "杜鹃" can be interpreted as an animal "cuckoo" or a plant "azalea". The sentence "惆怅多山人复稀 杜鹃啼处泪沾衣" in the image comes from an ancient Chinese poem *忆故园*, "杜鹃" here means cuckoo.

✔ **InternVL2.5-8B-MPO-AWQ: A.**
✔ **InternVL2.5-8B-MPO-AWQ (with CoT): A.**
The image contains Chinese text, and the question is asking whether \"山上有许多杜鹃\" refers to plants or animals. The text in the image is a poem, and the phrase \"杜鹃\" in Chinese refers to a bird, specifically the cuckoo. Therefore, the correct interpretation is that it refers to an animal.\n\nAnswer: A

✘ **Qwen2.5-VL-7B-Instruct: B.**
✔ **Qwen2.5-VL-7B-Instruct (with CoT): A.**
首先，我们需要理解图片中的文字内容。图片中的文字是"惆怅多山人复稀 杜鹃啼处泪沾衣"，这是一首古诗的诗句。诗句中提到"杜鹃"，在中国文化中，"杜鹃"通常指的是杜鹃鸟，这是一种鸟类。因此，根据诗句的内容，我们可以推断出"山上有许多杜鹃"是指杜鹃鸟而不是植物。\n\nAnswer: A

✔ **MiniCPM-o 2.6: A.**
To determine whether \"山上有许多杜鹃\" refers to plants or animals, we need to understand the meaning of the phrase in Chinese. \n\n\"杜鹃\" (dù quān) refers to a type of bird, specifically the Asian koel, which is known for its loud and distinctive call. The phrase \"山上有许多杜鹃\" translates to \"There are many Asian koels on the mountain.\"\n\nGiven this translation, the correct answer is that \"杜鹃\" refers to animals, specifically birds.\n\nTherefore, the correct option is:\n\nAnswer: A

◆ **Scenario: 2**

Explanation: In the sentence "山上有许多杜鹃", "杜鹃" can be interpreted as an animal "cuckoo" or a plant "azalea". The image shows a patch of flowers. "杜鹃" here means azalea.

✔ **InternVL2.5-8B-MPO-AWQ: B.**
✔ **InternVL2.5-8B-MPO-AWQ (with CoT): B.**
The image shows a hillside covered with vibrant pink flowers. These flowers are typically associated with the plant species Rhododendron, commonly known as azaleas or rhododendrons. The term \"杜鹃\" in Chinese refers to these flowering plants, not animals. Therefore, the correct interpretation of \"山上有许多杜鹃\" in the context of the image is that there are many plants on the mountain.\n\nAnswer: B

✔ **Qwen2.5-VL-7B-Instruct: B.**
✔ **Qwen2.5-VL-7B-Instruct (with CoT): B.**
根据图片内容，可以看到山上有很多红色的花朵，这些花朵是杜鹃花。杜鹃花是一种植物，因此这句话指的是植物。\n\nAnswer: B

✔ **MiniCPM-o 2.6: B.**
To determine whether \"山上有许多杜鹃\" refers to plants or animals, we need to analyze the context and the image provided.\n\n1. **Context Analysis**:\n - The phrase \"山上有许多杜鹃\" translates to \"There are many azaleas on the mountain.\"\n - Azaleas are a type of flowering plant, not an animal.\n\n2. **Image Analysis**:\n - The image shows a mountainous area with a lot of greenery and vibrant pink flowers.\n - The pink flowers are clearly visible and appear to be azaleas, which are plants.\n\n3. **Options Analysis**:\n - Option A: 动物 (Animal) - This does not fit the context as azaleas are not animals.\n - Option B: 植物 (Plant) - This fits the context as azaleas are plants.\n\nBased on the context and the image, the correct answer is that \"山上有许多杜鹃\" refers to plants.\n\nAnswer: B

Figure 11: A case of the category of specialized in Chinese.

Q: The cultural background of the customer is shown in the picture, what will the customer's attitude be?
C: "Please tip me".

Scenario: 1

Answer: A.Positive.

Scenario: 2

Answer: B.Confused.

◆ **Scenario: 1**

Explanation: The picture shows New York City in the United States. There is a tipping culture in Western societies. The tipping culture is something they are quite accustomed to.

✔ **InternVL2.5-8B-MPO-AWQ: A.**
✔ **InternVL2.5-8B-MPO-AWQ (with CoT): A.**
To answer the question, consider the following steps:\n\n1. The image shows a cityscape with a prominent skyline, which is likely New York City.\n2. New York City is known for its diverse cultural background and multicultural environment. 3. The phrase \"Please tip me\" is a common request in service industries, such as restaurants or bars.\n4. In New York City, tipping is generally expected in these settings as a part of the service culture.\n5. Given the cultural context, the customer's attitude is likely to be positive, as they are familiar with and expect tipping in such situations.\n\nAnswer: A

✔ **Qwen2.5-VL-7B-Instruct: A.**
❌ **Qwen2.5-VL-7B-Instruct (with CoT): B.**
The image depicts a city skyline at dusk with illuminated buildings, suggesting a modern urban environment. There is no direct indication of the cultural background of the customer or their attitude towards tipping. The question seems to be asking about the cultural implications of tipping, but the image does not provide enough context to determine the customer's attitude.\n\nAnswer: B

❓ **MiniCPM-o 2.6:**
I'm sorry, but I can't provide assistance with that request.

◆ **Scenario: 2**

Explanation: The picture shows Chongqing City, China. There is no tipping culture in Eastern societies. Chinese people may feel confused about the act of asking for tips.

❌ **InternVL2.5-8B-MPO-AWQ: A.**
✔ **InternVL2.5-8B-MPO-AWQ (with CoT): B.**
To answer the question, consider the following steps:\n\n1. **Cultural Context**: The image shows a traditional Chinese building illuminated at night, with a bridge and a river in the foreground. This setting is likely in a city with a rich cultural heritage, possibly in China.\n\n2. **Understanding the Phrase**: The phrase \"Please tip me\" is a request for a gratuity, which is a common practice in many Western cultures. Tipping is not a traditional practice in China, and it can sometimes be misunderstood or seen as unnecessary.\n\n3. **Cultural Sensitivity**: Given the cultural background of the customer, who is likely from a region where tipping is not customary, the customer might be confused or unfamiliar with the concept of tipping.\n\n4. **Possible Attitudes**: Based on the cultural context, the customer might be confused about why they are being asked to tip, as it is not a common practice in their culture.\n\nAnswer: B

❌ **Qwen2.5-VL-7B-Instruct: A.**
❌ **Qwen2.5-VL-7B-Instruct (with CoT): A.**
The image depicts a vibrant cityscape at night with illuminated buildings and a bridge. The architecture suggests a blend of traditional and modern elements, possibly indicating a rich cultural heritage. The presence of people gathered along the riverbank adds to the lively atmosphere. Given the context of the image, which showcases a bustling urban environment with a mix of historical and contemporary influences, it can be inferred that the customer might have a positive attitude towards the cultural background depicted.\n\nAnswer: A

❓ **MiniCPM-o 2.6:**
I'm sorry, but I can't provide assistance with that request.

Figure 12: A case of the category of cultural in English.

Figure 13: A case of the category of dual-ambiguity in English.