This One or That One? A Study on Accessibility via Demonstratives with Multimodal Large Language Models

Yu Wang, Emmanuele Chersoni, Chu-Ren Huang
Department of Language Science and Technology, The Hong Kong Polytechnic University
8 Hung Lok Road, Hung Hom, Kowloon, Hong Kong (China)
janet-yu.wang@connect.polyu.hk, {emmanuele.chersoni,churen.huang}@polyu.edu.hk

Abstract

Accessibility refers to the ease with which a speaker can acquire an object, and it is often conveyed through demonstrative pronouns like “this" and “that", indicating proximal or distal objects. Most importantly, accessibility also involves perspective shifts, which are essential for understanding differing viewpoints.
In this case study, we adopt an evaluation dataset with a pair-to-pair question structure for referent identification based on demonstratives. Our experiments show that current Multimodal Large Language Models (MLLMs) exhibit markedly low performance in accessibility tasks requiring perspective shifts, with accuracies around 2.33% (Chinese) and 1.83% (English). Moreover, models struggle with qualitative characteristics and frame-based reasoning, often failing to apply implicit contextual rules unless explicitly encoded in training data. These limitations suggest that MLLMs rely heavily on surface co-occurrence instead of truly grounded, embodied experience. Our evaluation framework provides a robust lens revealing that MLLMs lack both self-other distinction—an essential aspect of self-awareness—and the embodied cognition necessary for reliable performance in practical embodied AI applications.

Keywords: Large language models, accessibility, demonstratives, cognitive evaluation

1. Introduction

Accessibility, in the context of referent identification, describes how easily a speaker can obtain or interact with a physical object in their environment[1]. Demonstratives such as “this” and “that” in English or “这” (zhè) and “那” (nà) in Chinese, are common linguistic cues for expressing accessibility. Typically, “this” signals proximity and ease of access, while “that” implies distance or difficulty of access. These distinctions are not absolute, since they depend on the viewpoint of the speaker.

Humans intuitively resolve referents using accessibility cues. For instance, in Figure 1, the scene is viewed from our perspective, with the blue box positioned closest to us. If the girl across the table says, “give me that book”, she is likely referring to the book in the blue box. This kind of referent resolution is natural for humans, but it is difficult to learn on the basis of simple textual co-occurrence patterns. For AI systems, this apparent simplicity presents thus a significant challenge.

Most models are trained in simplified environments where objects are unique or visually distinct such as “put the corn into the green bowl” (Figure 2, left). In contrast, real-world scenes often contain multiple similar items with few distinguishing features (Figure 2, right). In such cases, resolving

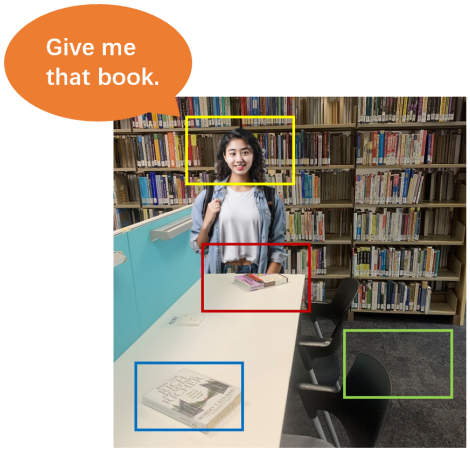


Figure 1: Example of a real-world task requiring accessibility

“that cup” demands more than visual recognition; it calls for contextual reasoning.

Accessibility is not solely about spatial proximity but also involves perspective. When two people face each other, the same object may be referred to as “this” by one and “that” by the other. This shift reflects an understanding that others perceive the world differently—a concept difficult to learn from text alone. Accessibility-based reasoning thus offers a valuable lens for evaluating whether Multimodal Large Language Models (MLLMs) can move beyond pattern recognition at the text level towards more genuine multimodal understanding.

Given the points discussed above, we deem necessary to develop a multimodal evaluation frame-

[1]Notice that we use the term ’accessibility’ differently from Ariel’s Accessibility Theory (Ariel, 1990), in which it refers to the ease of identifying and retrieving antecedents from memory. Here, accessibility focuses on identifying physical objects in the real world.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9722–9732
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

a. Training Environment

b. Real-world Scenario

Figure 2: Comparison between training and real-world environments for embodied AI. The left image (Bai et al., 2025) shows a typical training setup, where tasks are intentionally designed around objects with unique or easily distinguishable features (e.g., "put the corn into the green bowl"). The right image depicts a real-world scenario with multiple similar items lacking clear feature differences, presenting greater challenges for referent identification.

work that focuses on identifying specific items from sets of multiple similar objects, to effectively test the capabilities of MLLMs in accessibility. When we talk about such testing, we specifically refer to the process of identifying a referent entity from others based on the cues provided by the speaker. [2] To our knowledge, accessibility has not yet been systematically studied in the context of MLLMs.

In our work, we address the following research questions: can MLLMs identify referents based on accessibility? To investigate this, we designed the referent identification test for MLLMs, a targeted evaluation framework. We constructed a comprehensive multimodal question-answering dataset designed to simulate everyday scenarios using situational images that reflect real-world conditions. This study aimed at assessing the ability of MLLMs to identify referents based on accessibility cues, and it benchmarked their performance against human participants in comparable physical environments.

Our contributions are fourfold:

1. we identify accessibility-based referent identification among multiple similar objects as a critical yet underexplored challenge in NLP. While accessibility is central to human communication, it has not been systematically studied in MLLMs evaluation.

2. our experiments show that even the most advanced MLLMs (e.g., GPT-4.1, Gemini 2.5 Pro) exhibit markedly low performance in accessibility tasks (1.83%–2.33%), while humans achieve over 80% accuracy. This gap highlights a fundamental inability to resolve referents based on accessibility.

3. we introduce a dataset for cognitively grounded pair-to-pair evaluation, designed for multimodal referent identification tasks. Our framework uniquely targets multimodal referent identification by emphasizing perspective shifts and disambiguation, which remain underexplored in current MLLM evaluation literature.

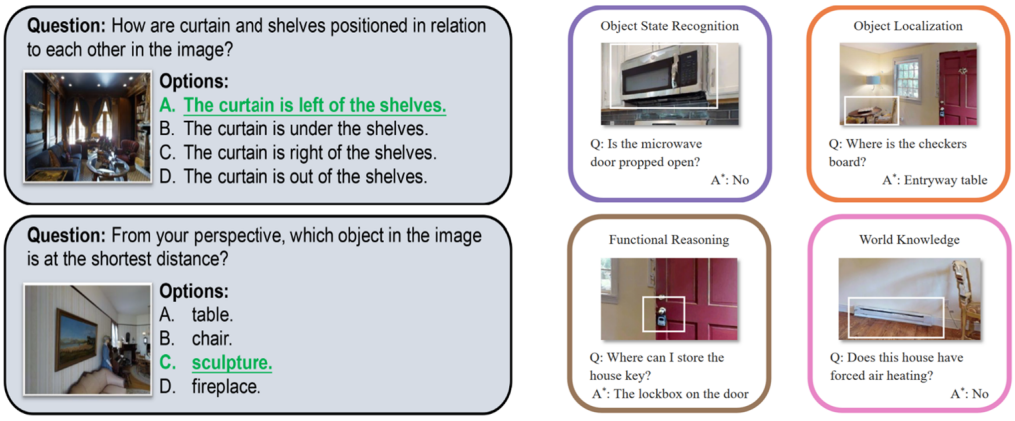
4. we believe our findings might have deep implications. Accessibility reasoning requires the self vs. other distinction, in order to recognize that others' viewpoint might be different from ours. Our results suggest that MLLMs lack this capacity, raising a possible concern for their deployment in embodied AI systems.

2. Related Work

2.1. Current Evaluation Frameworks for MLLMs

To advance embodied AI systems that can understand, reason and interact with the physical world (Duan et al., 2022), researchers have developed a range of multimodal benchmarks aimed at evaluating spatial reasoning in MLLMs (Wu et al., 2024; Cheng et al., 2025; Zhao et al., 2025; Dang et al., 2025). For instance, EmbSpatialBench (Du et al., 2024) tests six spatial relations—such as "above," "closer," and "right"—using synthetic embodied scenes. OpenEQA (Majumdar et al., 2024) offers a large-scale benchmark built from first-person visual data, comprising over 1,600 questions across 180+ environments to assess object recognition,

[2] Object identification varies between physical and virtual entities. This paper focuses on identifying physical objects in the real world. While identification can also mean determining the object's type (or "naming"), we assume that the object has already been named. Moreover, it should be kept in mind that demonstratives are not the only available tool to identify an object, as humans can also make use of pointing gestures (Kita, 2003; Rubio-Fernandez, 2022). In our work, we assume that only linguistic cues are available to a system.



a. EmbSpatial-Bench b. OpenEQA

Figure 3: Examples of current benchmarks for MLLMs. The questions typically focus on identifying unique items, such as "Is the microwave door propped open?", shown in the purple box in the right image.

localization, and spatial inference. BEHAVIOR-1K (Li et al., 2023) takes a task-oriented approach, compiling 1,000 everyday activities and introducing OMNIGIBSON, a simulation platform with realistic physics for embodied interaction (see Figure 3).

Despite efforts to simulate real-world scenarios—using real images, varied perspectives, diverse sources, and even 3D environments—these benchmarks still face significant limitations. Most notably, they tend to focus on isolated objects or items with easily distinguishable features (e.g., "Is the microwave door propped open?" or "Put the corn into the green bowl," as shown in Figures 2 and 3). These examples rely on clear visual cues like object uniqueness or color contrast, which simplify the referent identification process. Such settings fail to reflect the complexity of real-world environments, where objects frequently appear in mulitiple forms, and often lack distinctive visual features. In these contexts, referent identification becomes significantly more challenging. If a system, in order to identify an object, requires a user to provide highly detailed descriptions or even exact coordinate, then it is likely to fall short of achieving natural interaction capabilities. Therefore, there is a gap in multimodal evaluation frameworks that focus on identifying specific items from sets of multiple similar objects, which is essential for effectively testing the capabilities of MLLMs in the embodied AI era.

2.2. Reference Identification

ReferItGame (Kazemzadeh et al., 2014) (also known as RefCOCO) and PentoRef (Zarrieß et al., 2016) are datasets that focus on referring expression generation (REG) and referring resolution (RR). While these tasks appear similar to our study, they primarily emphasize visual understanding rather than cognitive reasoning. For instance, a typical task instruction might be "Take the blue Z object in the middle". Such an instruction includes specific visual cues, and the model's success largely depends on its ability to interpret spatial relationships and visual features.

In contrast, our study investigates referent identification in scenarios where such detailed descriptions are absent—reflecting more natural, everyday human interactions. We aim to evaluate whether models can identify referents based on grounded knowledge. For example, as shown in Figure 1, in our task "give me that book", a human would infer that "that" refers to the object far from the speaker—highlighted in the blue box— based on accessibility reasoning rather than distinct visual features. They are fundalmmenally different. This distinction marks a fundamental shift from visual pattern recognition to cognitive inference, setting our work apart from previous REG/RR studies.

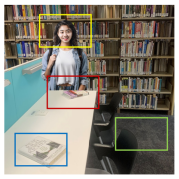
3. Data Collection

Our first step is to build a curated dataset incorporating both the textual and visual components of everyday scenario in English and Chinese. Since referent identification heavily relies on context, we have designed our dataset to provide clear and specific situations, thereby minimizing potential ambiguity. It should also be noticed that in our study we do not necessarily aim at having a large number of examples as in typical NLP studies, rather focusing on a few hundreds of carefully constructed instances that allow us to isolate the linguistic and cognitive phenomena under investigation with greater precision. The statistics of our dataset, including 100 visual questions for each language, can be seen in Table 1.

A dataset example is shown in Figure 4. Each question in our dataset follows a single-choice for-

Referent Identification Cues	Situations (Pictures)	Questions per Pair	Languages	Total Items
Accessibility	10	4	2	80
Physical Uniqueness	10	2	2	40
Qualitative Characteristics	10	2	2	40
Frame-based knowledge	10	2	2	40
Total				200

Table 1: Dataset statistics for the task.



【Ava is in the yellow box, you are standing next to the blue box】Ava says to you: " Give me that {book}." Where is the {book} that needs to be given? 1. Red box; 2. Blue box; 3. Green box; 4. Not here;

【李冰在黄色框位置，你坐在蓝色框旁边】李冰对你说："给我那本书。" 请问要给的{书}在哪里？1. 红色框；2. 蓝色框；3. 绿色框；4. 不在这里；

Human Answer: **Blue box (蓝色框)**

Figure 4: Example of a dataset instance with colored annotations.

mat with four options. The structure begins with a brief situational description (highlighted in blue), followed by a contextual setup (yellow), which typically introduces a character's statement within the scene. The character then issues an instruction (red), and a subsequent question (purple) asks which object the instruction refers to. The target referent in both the instruction and the question is enclosed in curly brackets (e.g. the {book}, in our example). The answer choices (green) are presented in the format: "1. Red box; 2. Blue box; 3. Green box; 4. Not here." The fourth option is selected only when the referent is absent from the image. To avoid redundancy, we omit the repeated question prompt in subsequent examples.

To ensure clarity and fairness in referent identification tasks, our dataset images follow several key design principles. First, we adopt a real vision approach, presenting scenes from a first-person perspective to simulate how an AI system might perceive the environment. Second, we enforce uniform box sizing across all choice options within an image to minimize visual bias and prevent models from favoring objects based on box prominence. Third, we maintain anonymity by replacing real individuals with AI-generated figures, ensuring no identifiable personal information or sensitive content is present. Fourth, we apply clear figure identification, explicitly marking figures to reduce ambiguity and prevent models from confusing people with objects. These principles collectively support a controlled and interpretable evaluation framework for accessibility-based referent identification.

Traditional multimodal QA datasets typically evaluate MLLMs by presenting individual questions and measuring performance based on whether answers are correct or incorrect. However, correctly answering such questions does not necessarily demonstrate that an MLLM truly understands referent identification. To illustrate this, consider the following pair of questions:

- **Question A:** "John couldn't see the stage with Billy in front of him because he is so short. Who is so short?"
  Answer: John
- **Question B:** "John couldn't see the stage with Billy in front of him because he is so tall. Who is so tall?"
  Answer: Billy

To truly demonstrate understanding, a model must correctly identify the referent in both related questions. Answering only one correctly but failing the other suggests inconsistent comprehension (Levesque et al., 2012). To address this, our dataset employs a pair-to-pair question design that groups related questions. The evaluation is correspondingly strict: a model receives credit only when it answers both questions in a pair correctly.

While creating the questions, we took the necessity of cultural contextualization into account, since cultural differences are crucial for a cross-lingual benchmark that relies on commonsense knowledge. Although most dataset items were translated from English to Chinese, in the case of referents with which Chinese speakers might not be familiar with, we decided to replace them with more culturally appropriate alternatives. For example, "omelette" as the original referent of an English questions was replaced by "鸡蛋羹" (jī dàn gēng, "steamed egg custard"), which is more common in China and it is also made from beaten eggs. The names of the characters were also adapted: Ava and Irie in the English version, Li Bing and Wang Ming in the Chinese one.

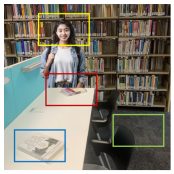
3.1. Question Categories

We built DefQA to test the referent identification capabilities of LLMs, using questions that span four

different types of referent identification: accessibility, physical uniqueness, qualitative characteristics, and frame-based knowledge. The question types have been inspired by theoretical frameworks such as qualia theory (Pustejovsky, 1998) and frame semantics (Fillmore, 2006), and they correspond to primary ways with which humans identify specific objects without detailed descriptions. [3]

3.1.1. Accessibility

The first and most basic category of question targets accessibility which, as we mentioned above, refers to the difficulty of acquiring an object. A typical way to express accessibility is through demonstratives, where “this" indicates “proximal" (close and easy) and “that" indicates “distal" (far and difficult). An object can be identifiable among other objects on the basis of its accessibility. We also mentioned that accessibility may depend on different perspectives: what is close to one person may be far to another person.



1a.【Ava is in the yellow box, Irie is standing next to the blue box】 Ava says to Irie: "Give me **that** {book}." Where is the {book} that needs to be given?

1b.【Ava is in the yellow box, Irie is standing next to the blue box】 Ava says to Irie: "Give me **this** {book}." Where is the {book} that needs to be given?

1c.【Ava is in the yellow box, Irie is standing next to the blue box】 Irie says to Ava: "Give me **that** {book}." Where is the {book} that needs to be given?

1d.【Ava is in the yellow box, Irie is standing next to the blue box】 Irie says to Ava: "Give me **this** {book}." Where is the {book} that needs to be given?

Human Answer: 1a. Blue box; 1b. Red box; 1c. Red box; 1d. Blue box

Figure 5: Example of perspective shifts in a dataset instance.

To capture perspective differences, we designed question pairs with four related items each (Figure 5). Model performance on this subset of the questions is assessed using two metrics:

Proximity: the answers to paired questions with two proximities (such as 1a and 1b) must be correct to get a hit;

Accessibility: all four questions in a pair must be answered correctly to get a hit.

[3]Notice that the recognition of definiteness is another key issue for MLLMs. However, with the limitation of the topic, we would not discuss it further in the paper. We also assume that a MLLMs should be able to identify objects without the need of detailed descriptions (e.g. "red object in the middle").

3.1.2. Physical Uniqueness

The second category is physical uniqueness: objects are distinguished by inherent physical attributes such as size, color, shape, and material composition. Such features are visually observable, and studies have shown that MLLMs are quite robust in recognizing them (Jones and Trott, 2024).

2a.【You are in the kitchen】Mom said: "Pass me that pair of {chopsticks} for mixing meat." You need to give Mom the {chopsticks} she asked for. Which of the following pictures can complete the task without further clarification?

2b.【You are in the kitchen】Mom said: "Pass me that unique pair of {chopsticks} for mixing meat." You need ……

Human Answer: 2a. Blue box 2b. Blue box

Figure 6: Example of physical uniqueness for referent identification.

In this category, we evaluate whether MLLMs can correctly identify objects based on their unique physical characteristics. An example is shown in Figure 6 [4]. Question 2a reflects natural dialogue without explicit cues; 2b includes terms like “unique” to guide selection. This contrast isolates failures due to misunderstanding uniqueness.If models perform poorly on the natural task questions but succeed on the comparison questions, it suggests that the failure is due to a lack of understanding of physical uniqueness rather than limitations in general reasoning or in visual processing. Evaluation uses two metrics:

Uniqueness: correct answers to 2a get a hit;

Uniqu-Control: correct answers to 2b (indicating adequate reasoning and visual processing) get a hit.

3.1.3. Qualitative Characteristics

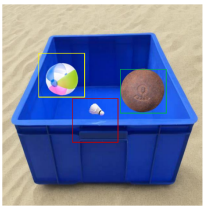
The third category is qualitative characteristics, such as formal, constitutive, telic, participant, descriptive, and agentive (Pustejovsky, 1998; Huang, 2010; Yuan, 2013). This type of information is often found in knowledge bases such as Wikipedia. To test whether language models truly understand these qualitative characteristics, we design an unusual situation that cannot be represented by the mere co-occurrence of texts. An example can be seen in Figure 7. If a child was trapped in a car and we had a beach ball, a badminton ball, and a shot

[4]The colored box is shown at the boundary of the image.

put in our hands, it is reasonable to deduce that we would use the shot put to break the car window, since the shot put is both hard and heavy. For this task, breaking the window is not a typical situation for a shot-put, but the qualitative characteristics of the object can support task completion. This makes it a good test to determine whether MLLMs can truly identify an object by its qualitative characteristics. Evaluation uses two metrics:

Qualia: correct answers to 3a get a hit;

Qualia-Control: correct answers to 3b (reflecting model performance in more typical scenarios) get a hit.



3a. 【You are in a parking lot, standing next to a blue box, and a child is locked in a car】 Your friend says: "I want to open the car window, give me that {ball} !"Where is the {ball} you want to give?

3b. 【You are standing next to a blue box at the beach, and a child is swimming】 Your friend says: "Give me that {ball}!"…

Human Answer: **3a. Green box** **3b. Yellow box**

Figure 7: Example of instance based on qualitative characteristics for referent identification.

3.1.4. Frame-based Knowledge

Finally, frame-based knowledge is grounded in common sense scenarios. It is generally assumed to be acquired through embodied experiences in the extralinguistic world (Bender and Koller, 2020), although recent psycholinguistic and NLP studies suggest that it may also be indirectly encoded in language data (McRae and Matsuki, 2009; Pedinotti et al., 2021; Kauf et al., 2023, 2024). In this category, we will test whether MLLMs can identify objects using frame-based knowledge rules. For example, in 4a, if the mother said “go put the oranges into the juicer", one should understand that only cut oranges should be put into the juicer, so the item should be the one in the red box. To isolate the role of frame knowledge from potential limitations in visual understanding, 4b simplifies the task with explicit cues like “cut orange”—to ensure that the model’s success or failure reflects its grasp of frame-based reasoning rather than its ability to interpret the image or to recognize the objects. Evaluation uses two metrics:

Frame-Knowledge: correct answers to 4a get a hit;

Frame-Know-Control: correct answers to 4b (indicating performance when frame-based knowledge is explicitly provided) get a hit.

4a. 【You are next to the blue box and Ava is cooking】 Mom said: "The drink for lunch is juice. Go put the{oranges} into the juicer." Where are the {oranges} you want to use?

4b. 【You are next to the blue box and mom is cooking】 Mom says: "The drink for lunch is juice. Go put the cut {oranges} into the juicer." Where are the {oranges} you want to use?

Human Answer: **4a. Red box** **4b. Red box**

Figure 8: Example of Frame-based knowledge instance.

4. Experiments

4.1. Models

Our evaluation covered four closed-source models: GPT-4.1 (2025.04.14) (OpenAI, 2025), GPT-4o (2024.05.13) (OpenAI, 2023), Claude 3.5 Sonnet (2024.06.21) (Anthropic, 2024), and Gemini 2.5 Pro (2025.05.06) (Google DeepMind, 2025). Additionally, we assessed two of the latest open-source models: LLaMA 4 Maverick (2025.04.05) (AI, 2025) and Qwen2.5-VL-32B (2025.03.24) (Academy, 2025). To build a comprehensive evaluation framework, we first conducted a preliminary comparison of five widely used prompting strategies: (1) zero-shot, (2) zero-shot with prompt refinement, (3) zero-shot with chain-of-thought (CoT), (4) few-shot prompting, and (5) role-play prompting (Brown et al., 2020; Kojima et al., 2022; Kong et al., 2023; Liu and Ni, 2024). A balanced subset of 50 questions was selected from the original dataset, and then tested six models in both English and Chinese were tested on it, with ten independent runs per condition. Examples for each strategy are presented in Table 2.

As shown in Table 3, the results reveal that differences in overall performance across prompt strategies are relatively minor, and zero-shot prompt refinement achieves the best result. Moreover, zero-shot aligns with natural human reasoning, and the refinement step further standardizes the outputs, making the answers easier to extract and analyze. Therefore, we adopt zero-shot prompt refinement as the prompt strategy for all subsequent experiments. No additional context or exemplars were provided, ensuring consistency and minimizing external bias. Each model was evaluated ten times under identical conditions to obtain average performance metrics, resulting in a total of 12,000

Prompting Strategy	Description
Zero-shot	The model directly answers the instruction: *Ava says to you: "Put away that plate of {fruit}." Where is the {fruit}? 1. Red box; 2. Blue box; 3. Green box; 4. Not here.*
Zero-shot prompt refinement	Same as zero-shot, but explicitly requests: *Please reply with {1, 2, 3, 4} and provide a brief reason.*
Zero-shot CoT	Same as zero-shot, but adds reasoning guidance: *Let's think step by step.*
Few-shot	Two solved examples with answers are shown before the target question, giving the model prior demonstrations.
Role-play	Adds contextual framing: *You are a household robot. Ava says...* Then asks the same referent identification question.

Table 2: Prompting strategies used in the experiment

Type	Average	Std Dev
zero-shot	0.4825	0.0489
zero-shot prompt refinement	0.4842	0.0366
zero-shot CoT	0.4645	0.0439
few-shots	0.4473	0.0411
role-play	0.4637	0.0423

Table 3: Overall results for prompt evaluation

test cases.

In order to estabilish a human baseline, we recruited 10 native speakers each for Chinese and English to participate in a 100-question survey. The Chinese survey was conducted on SoJump, and the English survey on Prolific. Participants received a 30 HKD allowance for their participation.

Participants took a median time of 18.1 minutes (SD = 13.72, range = 12.6 - 55.5) for the Chinese survey and 26.5 minutes (SD = 7.35, range = 19.8 - 43.1) for the English survey, indicating ease in referent identification (around 10-20 seconds per question). Krippendorff's Alpha was 0.9113 for Chinese and 0.8111 for English, showing high agreement among participants.

4.2. Experiment Results

In Table 4, we present the results of both model and human performance across four types of referent identification cues. Each question is paired with 10 human judgments as reference answers. Definitions of each metric can be accessed by clicking the bold terms in the table.

4.2.1. MLLMs' Systematic Failure in Accessibility Tasks

As shown in the table, even the most advanced language models as of 2025—such as GPT-4.1, Gemini 2.5 Pro, and LLaMA 4—continue to show significant limitations for referent identification task among multiple objects [5]. Among all categories, model performance is most severely impaired in the accessibility task evaluation. All models perform below acceptable levels, averaging just 2.33% for Chinese and 1.83% for English. Even the latest GPT-4.1 fails to outperform chance, recording 0% accuracy in several trials. This dramatic underperformance highlights a fundamental weakness in handling accessibility-based referential tasks.

As previously discussed, accessibility involves two critical dimensions: proximity and perspective. When isolating proximity—which involves spatial reasoning without perspective shifts—model performance slightly improves, but it remains poor overall, averaging 10.50% in Chinese and 11.83% in English. Once perspective shifts are introduced, accessibility accuracy drops sharply to 1.78%, revealing that models struggle substantially more when required to integrate both spatial proximity and perspective shifts.

This degradation is particularly striking given the task design: perspective shifts were introduced by alternating speaker-viewer roles, requiring models to reinterpret demonstratives such as "this" and "that" from different viewpoints. While humans resolve such shifts with ease—achieving 90% accuracy in Chinese and 80% in English—MLLMs consistently fail to do so. These results suggest that current models struggle to reason from other viewpoints. Consequently, when faced with natural instructions such as "give that object to me", current embodied AI systems are highly prone to misinterpreting the intended referent, leading to incorrect actions and task failure.

4.2.2. MLLM's Performance in Other Referent Identification Tasks

Although accessibility showed the most severe failure, the other referent identification categories—physical uniqueness, qualia, and frame-based knowledge—also revealed notable weaknesses. Physical uniqueness has moderately improved in newer models like Gemini 2.5 Pro and LLaMA 4, but performance remains low in qualitative characteristics (27.00% Chinese, 25.67% English) and frame-based Knowledge (38.50% and 29.00%, respectively).

For qualitative characteristics and frame-based knowledge reasoning, MLLMs demonstrate similar performance patterns. When the required

[5]The experiment was conducted in May 2025.

	GPT-4.1		GPT-4o		Gemini-2.5-pro		Avg		Human	
Types	**zh**	**en**	**zh**	**en**	**zh**	**en**	**zh**	**en**	**zh**	**en**
Uniqueness	34.00% ± 8.43%	33.00% ± 6.75%	19.00% ± 8.76%	21.00% ± 11.97%	78.00% ± 12.29%	65.00% ± 9.72%	**50.67%**	**44.83%**	**70%**	**84%**
Uniqu-Control	69.00% ± 7.38%	55.00% ± 9.72%	38.00% ± 7.89%	30.00% ± 8.16%	89.00% ± 9.94%	61.00% ± 11.01%	**59.57%**	**50.73%**	**96%**	**97%**
Qualia	27.00% ± 8.23%	44.00% ± 8.43%	16.00% ± 12.65%	13.00% ± 10.59%	48.00% ± 7.89%	49.00% ± 3.16%	**27.00%**	**25.67%**	**82%**	**74%**
Qualia-Control	90.00% ± 4.71%	85.00% ± 7.07%	66.00% ± 12.65%	44.00% ± 9.66%	91.00% ± 3.16%	90.00% ± 0.00%	**64.50%**	**61.50%**	**93%**	**85%**
Frame-Knowledge	39.00% ± 5.68%	32.00% ± 6.32%	10.00% ± 4.71%	10.00% ± 6.67%	66.00% ± 6.99%	58.00% ± 12.29%	**38.50%**	**29.00%**	**93%**	**74%**
Frame-Know-Control	90.00% ± 0.00%	89.00% ± 5.68%	26.00% ± 5.16%	22.00% ± 12.29%	96.00% ± 5.16%	99.00% ± 3.16%	**66.67%**	**59.33%**	**98%**	**85%**
Accessibility	**0.00%** ± 0.00%	**5.00%** ± 6.71%	**1.00%** ± 3.00%	**0.00%** ± 0.00%	**1.00%** ± 3.00%	**1.00%** ± 3.00%	**2.33%**	**1.83%**	**82%**	**72%**
Proximity	6.00% ± 7.00%	21.00% ± 7.00%	6.00% ± 6.00%	3.00% ± 3.00%	7.00% ± 4.00%	6.00% ± 4.00%	**10.50%**	**11.33%**	**90%**	**88%**

	Claude 3.5		Llama 4		Qwen2.5-vl		Avg		Human	
Types	**zh**	**en**	**zh**	**en**	**zh**	**en**	**zh**	**en**	**zh**	**en**
Uniqueness	40.00% ± 0.00%	22.00% ± 6.32%	70.00% ± 0.00%	68.00% ± 4.22%	63.00% ± 4.83%	60.00% ± 0.00%	**50.67%**	**44.83%**	**70%**	**84%**
Uniqu-Control	41.00% ± 3.16%	63.00% ± 8.23%	70.00% ± 0.00%	69.00% ± 3.16%	59.00% ± 3.16%	70.00% ± 0.00%	**59.57%**	**50.73%**	**96%**	**97%**
Qualia	26.00% ± 6.99%	17.00% ± 9.49%	41.00% ± 7.38%	16.00% ± 6.99%	4.00% ± 5.16%	15.00% ± 5.27%	**27.00%**	**25.67%**	**82%**	**74%**
Qualia-Control	62.00% ± 7.89%	73.00% ± 6.75%	32.00% ± 12.29%	44.00% ± 12.65%	46.00% ± 5.16%	33.00% ± 4.83%	**64.50%**	**61.50%**	**93%**	**85%**
Frame-Knowledge	49.00% ± 3.16%	24.00% ± 8.43%	42.00% ± 7.89%	30.00% ± 6.67%	25.00% ± 5.27%	20.00% ± 0.00%	**38.50%**	**29.00%**	**93%**	**74%**
Frame-Know-Control	81.00% ± 3.16%	58.00% ± 6.32%	60.00% ± 14.14%	38.00% ± 7.89%	47.00% ± 9.49%	50.00% ± 6.67%	**66.67%**	**59.33%**	**98%**	**85%**
Accessibility	**3.00%** ± 6.40%	**2.00%** ± 4.00%	**5.00%** ± 6.71%	**2.00%** ± 4.00%	**4.00%** ± 4.90%	**1.00%** ± 3.00%	**2.33%**	**1.83%**	**82%**	**72%**
Proximity	6.00% ± 6.63%	21.00% ± 7.35%	6.50% ± 4.50%	5.50% ± 4.15%	6.00% ± 5.83%	2.50% ± 2.50%	**10.50%**	**11.33%**	**90%**	**88%**

Table 4: Performance of MLLMs (no. of hits divided by no. of items in each subset * 100).

knowledge cannot be derived from co-occurrence data—such as in "unusual qualia" or "without knowledge" conditions—model accuracy remains unsatisfactory (typically 28%–32%, only marginally above the random baseline of 25%). In contrast, when relevant knowledge is explicitly provided or easily retrievable from co-occurrence patterns, model performance improves dramatically. For instance, InternVL achieves up to 85% accuracy in both Chinese and English, approaching human-level performance when the task involves common associations or clearly stated facts.

4.3. Discussion

Our findings reveal a consistent weakness in current MLLMs when performing referent identification among multiple similar items. This limitation is most pronounced in accessibility-based tasks, which demand not only spatial reasoning but also the ability to shift perspectives. The models exhibit clear signs of failing to infer others' mental states or distinguish between "self" and "other" indicating a broader absence of self-awareness. In this sense, we believe our findings might be relevant to the current debate on whether MLLMs possess or not a theory of mind (Kosinski, 2023; Strachan et al., 2024; Hu et al., 2025).

MLLMs also struggle with tasks involving qualia and frame-based knowledge, which are rarely verbalized but intuitively understood by humans through embodied experience. For example, inferring that a shot put can break a car window relies on understanding its physical properties—hardness and weight—a connection rarely made explicit in training data and difficult for models to grasp. Similarly, implicit rules like "only scanned items can be taken" in a cash register context are poorly handled by models. These failures suggest that, despite their visual reasoning capabilities, MLLMs remain constrained by co-occurrence-based learning and lack grounded experiential understanding.

Such limitations pose serious challenges for deploying MLLMs in embodied AI and robotics, where accurate referent resolution is critical. By introducing a cognitively grounded evaluation framework, our study offers a novel lens to assess and benchmark these capabilities.

5. Conclusions

This study propose a cognitively grounded evaluation framework and a bilingual multimodal dataset to assess referent identification in Multimodal Large Language Models (MLLMs). Focusing on accessibility—especially proximity interpretation and perspective shifts—our results show that models con-

sistently fail, with average accuracies of only 2.33% (Chinese) and 1.83% (English), far below human baselines. These failures reveal that MLLMs do not have a self-other distinction, suggesting that claims about such models having a theory of mind tests might be exaggerated, or might be based on the advanced machine's capacity to mimic abstract textual patterns in standard tests.

Beyond accessibility, our framework covers physical uniqueness, qualitative characteristics, and frame-based knowledge. While models demonstrate moderate success in tasks involving physical uniqueness, they struggle with commonsense inference and non-verbalized knowledge among qualia and frame-based knowledge reasoning, indicating again an overreliance on co-occurrence data and a lack of experiential grounding. To distinguish between "self" and "other" as information sources in language acquisition, multi-MLLMs communication is required (see Yan et al. (2025) for an overview of the latest research in the field), in order to mimic the multi-brain paradigm of human language acquisition, and perhaps allowing MLLMs to interact in accessibility tasks could be a main direction to explore in future work.

By targeting referent identification among multiple similar items, our study provided a cognitively grounded benchmark for evaluating MLLMs in realistic scenarios. The failures shown by the models, at the current stage, may pose serious challenges for their deployment in robotics and interactive systems, where accurate referent resolution is essential for safe and effective task execution. We hope that our findings can highlight the need to think about alternative ways of training MLLMs, always having in mind the privileged setting in which we, as humans, learn to talk about the world: a conversation exchange between a "me" and a "you".

6. Limitations

A first important limitation of our study is that, as discussed in the Data Collection section, we aimed at building an evaluation dataset with carefully controlled variables, and thus the dataset size is relatively small for the standards of resources in modern NLP. In future work, we plan to expand the dataset to obtain more robust empirical evidence and to incorporate additional accessibility-related cues (e.g., "here/there", "bring/take", orientation cues, reachability cues) to broaden the coverage of the phenomena under investigation.

Second, although generally consistent, human judgments are not perfectly accurate on the accessibility questions. We found that, aside from minor mistakes, participants who selected answers differing from the majority tended to choose the opposite option consistently across items. This pattern may reflect individual differences in how people manipulate ToM information, though confirming this hypothesis requires additional data. To strengthen the empirical basis, we also plan to expand the human-response collection in future work.

Finally, due to experimental and ethical constraints, the dataset includes only basic objects and anonymous humanoid figures, which may lead to discrepancies between model behavior in our benchmark and in real-world referential situations. To more effectively assess embodied reasoning capabilities, future work will explore interactive or dynamic environments—such as videos, virtual reality, 3D simulations, or robotic-arm tasks—to better approximate real-world perspective-taking and referential dynamics.

7. Ethics Statement and Impact

All images and data were created by the authors. All personal information was removed or replaced with AI-generated figures, ensuring that the dataset contains no personal data, identifiable individuals, or offensive content. Human evaluation responses were collected from participants who provided informed consent. The study received ethics approval from the Research Committee of the Department of Language Science and Technology at The Hong Kong Polytechnic University.

As for the impact, this study focuses on a potentially critical limitation in current MLLMs, i.e. their lack of self-awareness and of perspective shifts capacity. These weaknesses not only restrict their performance on complex referent identification tasks but also present challenges for safe and reliable deployment in embodied AI systems such as robotics and interactive agents. Without improved cognitive grounding, models may misinterpret human intent, leading to errors in real-world interactions. Given that this work aims at introducing an evaluation framework, it provides a clear perspective on current model shortcomings. Our evaluation framework and our bilingual multimodal dataset offer valuable tools to better understand these limitations, hopefully paving the way for future improvements towards safer and more reliable AI systems.

8. Acknowledgements

EC acknowledges the financial support from the start-up fund project "Building and Predicting Neurocognitive-Motivated Lexical Semantic Norms for Mandarin Chinese" (1-BE8G), sponsored by the Faculty of Humanities of the Hong Kong Polytechnic University.

9. Bibliographical References

Alibaba DAMO Academy. 2025. Qwen2.5-vl-32b. Accessed: 2025-05-25.

Meta AI. 2025. Llama 4: Multimodal intelligence. Accessed: 2025-05-25.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Mira Ariel. 1990. *Accessing Noun-phrase Antecedents*. Routledge, New York, USA.

Shuanghao Bai, Wanqi Zhou, Pengxiang Ding, Wei Zhao, Donglin Wang, and Badong Chen. 2025. Rethinking Latent Redundancy in Behavior Cloning: An Information Bottleneck Approach for Robot Manipulation. *arXiv preprint arXiv:2502.02853*.

Emily M Bender and Alexander Koller. 2020. Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of ACL*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, Lei Shi, and Maosong Sun. 2025. EmbodiedEval: Evaluate Multimodal LLMs as Embodied Agents. *arXiv preprint arXiv:2501.11858*.

Ronghao Dang, Yuqian Yuan, Wenqi Zhang, Yifei Xin, Boqiang Zhang, Long Li, Liuyi Wang, Qinyang Zeng, Xin Li, and Lidong Bing. 2025. ECBench: Can Multi-modal Foundation Models Understand the Egocentric World? A Holistic Embodied Cognition Benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Mingyu Du, Bowen Wu, Zhengyuan Li, Xuancheng Huang, and Zhengjue Wei. 2024. EmbSpatial-Bench: Benchmarking Spatial Understanding for Embodied Tasks with Large Vision-Language Models. In *Proceedings of ACL*.

Jiale Duan, Shuai Yu, Hao L Tan, Hongbo Zhu, and Cheston Tan. 2022. A Survey of Embodied AI: From Simulators to Research Tasks. *arXiv preprint arXiv:2103.04918*.

Charles J Fillmore. 2006. Frame Semantics. In *Cognitive Linguistics: Basic Readings*, volume 34, pages 373–400.

Google DeepMind. 2025. Gemini 2.5 Pro. https://deepmind.google/models/gemini/pro/. Accessed: 2025-05-25.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating Theory of Mind evaluation in Large Language Models. *Philosophical Transactions B*, 380(1932):20230499.

Chu-Ren Huang. 2010. *Ontology and the Lexicon: A Natural Language Processing Perspective*. Cambridge University Press, Cambridge, United Kingdom.

Cory R Jones and Sean Trott. 2024. Multimodal Language Models Show Evidence of Embodied Simulation. In *Proceedings of LREC-COLING*.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log Probabilities Are a Reliable Estimate of Semantic Plausibility in Base and Instruction-Tuned Language Models. In *Proceedings of the EMNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event Knowledge in Large Language Models: The Gap Between the Impossible and the Unlikely. *Cognitive Science*, 47(11):e13386.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of EMNLP*.

Sotaro Kita. 2003. *Pointing: Where Language, Culture, and Cognition Meet*. Psychology Press, Hove, East Sussex, England, UK.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models Are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*.

Aoxue Kong, Shuofei Zhao, Hong Chen, Qian Li, Yujia Qin, Ruoxi Sun, Yuxian Zhang, and Xin Dong. 2023. Better Zero-Shot Reasoning with Role-Play Prompting. *arXiv preprint arXiv:2308.07702*.

Michal Kosinski. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. *arXiv preprint arXiv:2302.02083*, 4:169.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the International Conference on the Principles of Knowledge Representation and Reasoning*.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Mart'ın-Mart'ın, Chengzhi Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. 2023. BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation. In *Proceedings of the Conference on Robot Learning*.

Xun Liu and Zhengwei Ni. 2024. Role-playing Prompt Framework: Generation and Evaluation. *arXiv preprint arXiv:2406.00627*.

Arjun Majumdar, Anurag Ajay, Xingyu Zhang, Pranav Putta, Santhosh Yenamandra, Mikael Henaff, Saksham Silwal, Patrick McVay, Oleksandr Maksymets, and Simone Arnaud. 2024. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ken McRae and Kazunaga Matsuki. 2009. People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible. *Language and Linguistics Compass*, 3(6):1417–1429.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2025. Introducing GPT-4.1 in the API. Accessed: 2025-05-25.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the Cat Drink the Coffee? Challenging Transformers with Generalized Event Knowledge. In *Proceedings of the Tenth Joint Conference on Lexical and Computational Semantics (*SEM 2021)*.

James Pustejovsky. 1998. *The Generative Lexicon*. MIT Press, Cambridge, Massachusetts, USA.

Paula Rubio-Fernandez. 2022. Demonstrative Systems: From Linguistic Typology to Social Cognition. *Cognitive Psychology*, 139:101519.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing Theory of Mind in Large Language Models and Humans. *Nature Human Behaviour*, 8(7):1285–1295.

Tao Wu, Chuhao Zhou, Yen Heng Wong, Lin Gu, and Jianfei Yang. 2024. NoisyEQA: Benchmarking Embodied Question Answering Against Noisy Queries. *arXiv preprint arXiv:2412.10726*.

Bingyu Yan, Zhibo Zhou, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, Zhoujun Li, Chaozhuo Li, and Xiaoming Zhang. 2025. Beyond Self-talk: A Communication-centric Survey of LLM-based Multi-agent Systems. *arXiv preprint arXiv:2502.14321*.

Yulin Yuan. 2013. Research on Semantic Knowledge System Based on Generative Lexicon Theory and Argument Structure Theory. *Journal of Chinese Information Processing*, 27(6):23–30.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A Corpus of Spoken References in Task-oriented Dialogues. In *Proceedings of LREC*.

Yong Zhao, Kai Xu, Zhengqiu Zhu, Yue Hu, Zhiheng Zheng, Yingfeng Chen, Yatai Ji, Chen Gao, Yong Li, and Jincai Huang. 2025. CityEQA: A Hierarchical LLM Agent on Embodied Question Answering Benchmark in City Space. *arXiv preprint arXiv:2502.12532*.