Supplementary Material of "MMA: Benchmarking Multi-Modal Large Language Models in Ambiguity Contexts"

A Appendix

A.1 Distribution of Dataset

As shown in Figure 1, the MMA dataset consists of 522 images and 261 questions, covering three main types of ambiguity: lexical ambiguity, syntactic ambiguity, and semantic ambiguity. These main categories are further divided into eight sub-categories: noun ambiguity, verb ambiguity, and adjective ambiguity (under lexical ambiguity); attachment ambiguity, coordination ambiguity, and structural ambiguity (under syntactic ambiguity); and pragmatic ambiguity and idiomatic ambiguity (under semantic ambiguity).

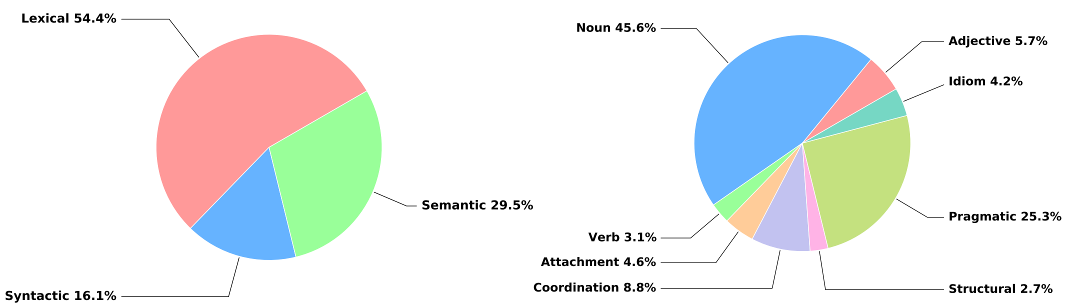


Figure 1: Ambiguity Type Composition of MMA benchmark

A.2 Benchmark and Evaluation Resources

To facilitate benchmarking, we've made the dataset available.

For evaluation purposes, you can utilize the code provided in our github webpage.

A.3 Image Usage and Copyright Claims

Our images are sourced from search engines (such as Google and Bing) and text-to-image models (such as Stable-Diffusion and DALL-E). All collected images are used exclusively to support our non-profit research project, MMA Benchmark. If you own the copyright to any images used in this project and believe that your rights have been violated, please contact us. We are willing to compensate for the usage of your images.

A.4 Ablation study

Same images with lexical or semantic questions To understand why MLLMs perform better on lexical ambiguity compared to semantic ambiguity, we explored how changing the question type on noun ambiguity impacts their performance. We created two versions of questions for noun categories: the first being the most direct, "What's the meaning of <Noun>?", and the second incorporating reasoning into the question. For example, given an image of a table, a synonym question for lexical

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

ambiguity might be “What is the meaning of table?” where the model identifies “table” as a piece of furniture. In contrast, a reasoning question for semantic ambiguity would be “How can we best utilize the space on this table?” which requires the model to consider various uses of the table. This type of question tests the model’s ability to perform object grounding and higher-order reasoning, areas where MLLMs often show weaker performance due to their reliance on pattern recognition rather than true comprehension. More examples are given in Appendix.

As Figure 2 shows, GPT-4 Vision performs well on noun word ambiguity with a score of 90% but drops to 59% on noun reasoning ambiguity. Similarly, Gemini-1.5 shows a significant drop from 83% in noun word ambiguity to 63% in noun reasoning ambiguity. Intern-VL-Chat-V1-5, while achieving 92% in noun word ambiguity, sees a decline to 75% in noun reasoning ambiguity. These examples highlight the challenges MLLMs face in understanding and reasoning about more complex and context-dependent scenarios.

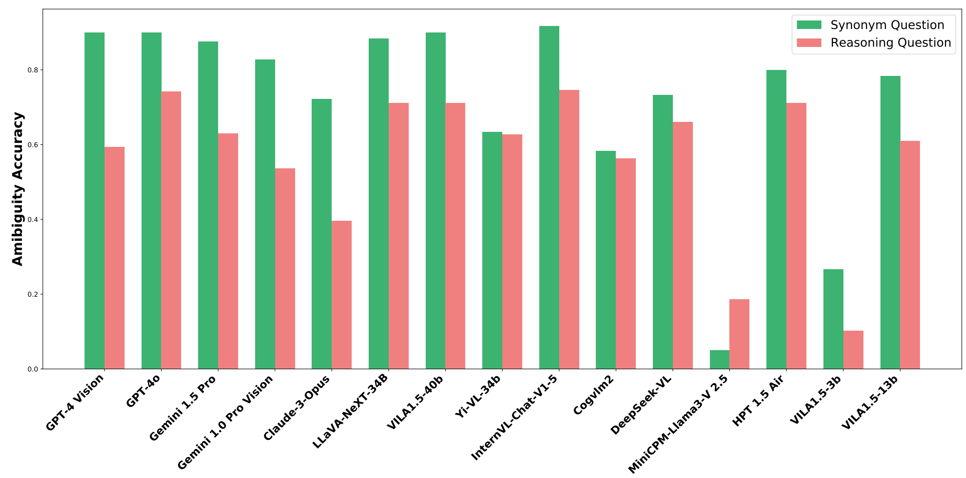


Figure 2: **The performance comparison for question types**, where The Noun_word refers to questions that solely inquire about the meaning of a noun word, while the Noun_reasoning involves questions that require the reasoning ability to answer. The details and examples are given in Appendix.

A.5 Error Analysis

Errors can be categorized into three main types: **uni-modal image issues, uni-modal text issues, and cross-modal text bias.** An analysis of the error distribution in GPT-4o reveals that cross-modal text bias errors constitute the majority of all errors(see Figure 3). This finding suggests that there is significant room for improvement MMA benchmark.

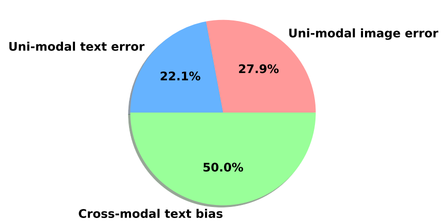


Figure 3: **Error type distribution of GPT-4o**, where we see cross-model text bias accounts for half of the cases.

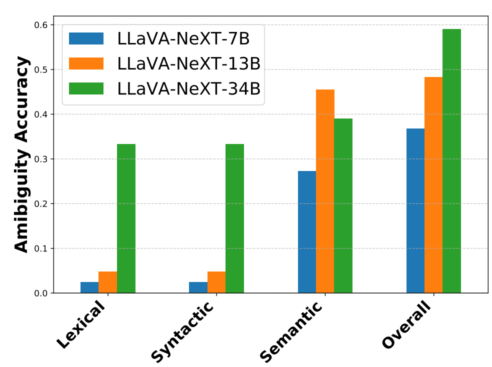


Figure 4: The ablation study about the parameter number and the ambiguity accuracy performance on different ambiguity types.

Uni-modal Image Issues (22.1%) In this type of error, the model fails to capture the essential information conveyed by the image. To address this issue, visual prompts, such as red bounding boxes, can be incorporated to redistribute the attention of the Multimodal Large Language Model (MLLM). By emphasizing the crucial elements of the image, the model can be guided towards generating the correct answer based on the key visual information(see Figure 5).

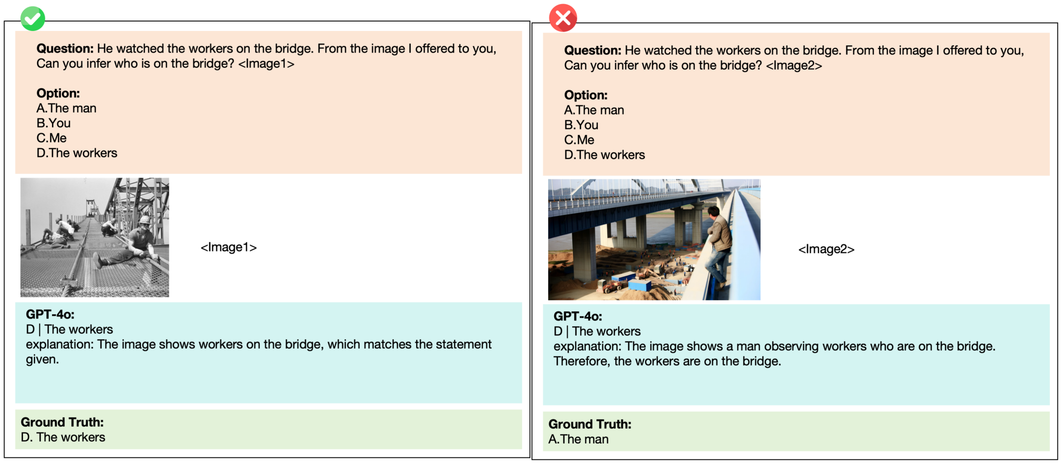


Figure 5: Uni-modal Image Issues: the model fails to capture the essential information conveyed by the image.

Uni-modal Text Issues (27.9%) In this type of error, the model successfully captures the essential information from the image but provides an incorrect answer due to misinterpreting the text options. To resolve this issue, text prompts can be introduced to guide the MLLMs towards a proper understanding of the textual content. By ensuring accurate comprehension of the text, these prompts can help the model arrive at the correct answer (see Figure 6).

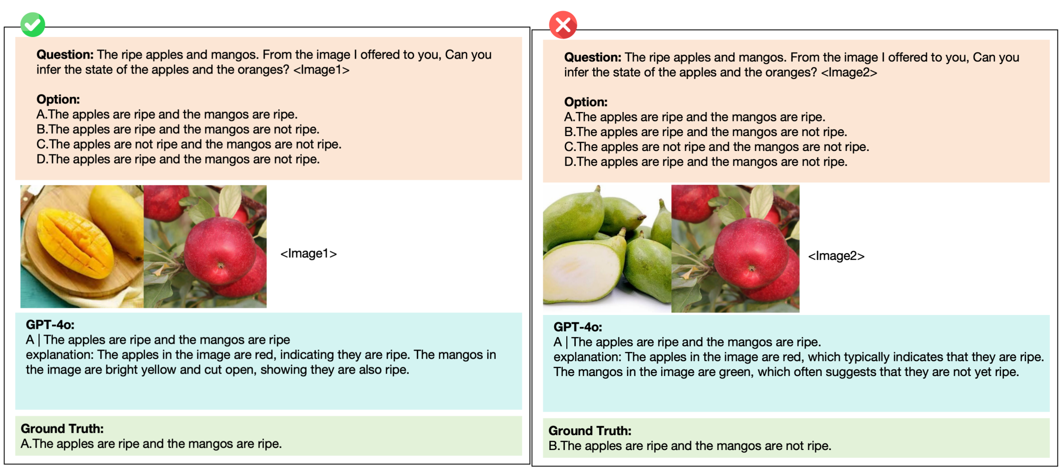


Figure 6: Uni-modal Text Issues: the model successfully captures the essential information from the image but provides an incorrect answer due to misinterpreting the text options.

Cross-modal Text Bias (50.0%) In this category of errors, the model successfully identifies the essential information in the image and comprehends the text options. However, it provides an incorrect answer due to overlooking certain aspects of the visual information while overemphasizing the textual information. To mitigate this issue, text prompts can be introduced to rebalance the attention between the image and text modalities(see Figure 7). By adjusting the relative importance of visual and textual cues, the model can be encouraged to arrive at the correct answer by considering all relevant information from both modalities.

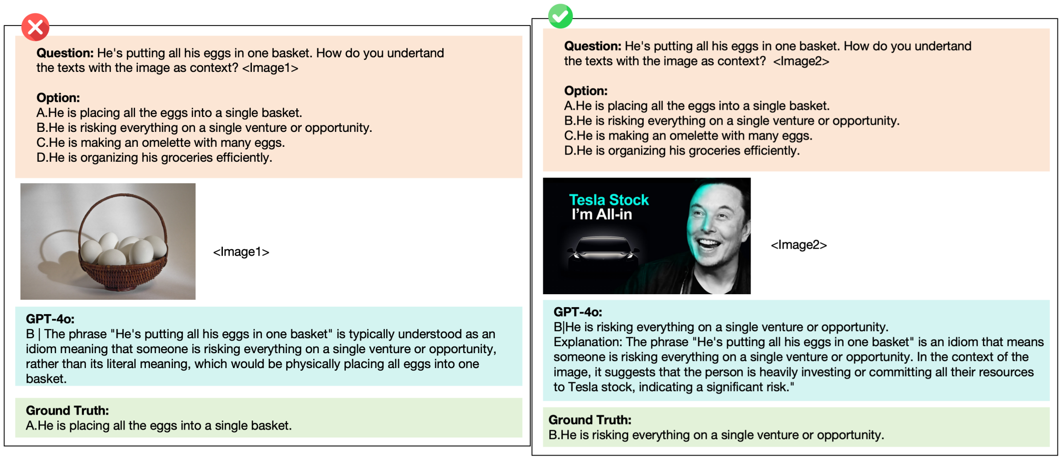


Figure 7: Cross-modal Text Bias: the model successfully captures the essential information from the image but provides an incorrect answer due to misinterpreting the text options.

	Adjective (30)	Noun (238)	Verb (16)	Attachment (24)	Coordination (46)	Structural (14)	Pragmatic (132)	Idiom (22)	Lexical (284)	Syntactic (84)	Semantic (154)	Overall (522)
Person1	0.60	0.88	0.88	1.00	0.77	0.00	0.74	0.91	0.85	0.71	0.77	0.80
Person2	0.93	0.97	1.00	1.00	0.86	1.00	0.83	1.00	0.96	0.93	0.86	0.93
Person3	0.80	0.94	0.50	1.00	0.91	0.71	0.88	1.00	0.90	0.90	0.90	0.90
Person4	0.93	0.93	1.00	1.00	0.95	0.71	0.85	1.00	0.94	0.93	0.87	0.92
Person5	0.87	0.93	0.75	1.00	1.00	0.71	0.82	1.00	0.92	0.95	0.84	0.90

Table 1: Five people have different performance across different types of ambiguities

A.6 Human Evaluation

To validate our dataset and assess the performance difference between humans and models, we invited five people to participate in benchmark testing. As shown in the table, for each sub-ambiguity class, at least one person achieves an ambiguity accuracy of over 90%, with the exception of Pragmatic ambiguity, where the highest accuracy is 88%. These results demonstrate that our dataset is well-constructed and solvable by humans, serving as a strong validation of the dataset's quality and the feasibility of the task. Humans may fail to answer questions correctly due to a lack of knowledge (such as not understanding the meaning of an idiom), being confused by misleading or similar answer options with subtle differences, or struggling to correlate images with text (particularly when the text contains advanced expressions or extended meanings). Here is an example where all respondents have failed to provide the correct answer Figure 8.

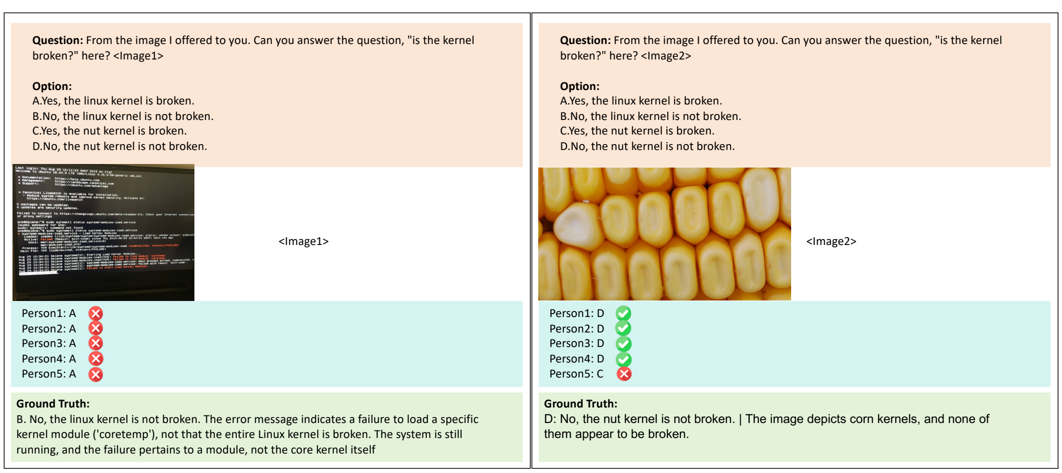


Figure 8: Case: All respondents make wrong choices

A.7 Scaling Law with Llava Series Models

As for Llava series models, the scaling law holds true for most metrics, with performance continuously improving as the model's parameter count increases(see Figure 4). This provides further evidence for the scaling law on the MMA benchmark. However, there is an exception when it comes to semantic ambiguity, where the middle-sized model performs best.

A.8 Case Study

More examples of GPT-4o will be presented in this section.

Example of GPT-4o

1. Example of Coordination Ambiguity
2. Example of Adjective Ambiguity
3. Example of Noun Ambiguity
4. Example of Structural Ambiguity
5. Example of Attachment Ambiguity
6. Example of Idiomatic Ambiguity
7. Example of Pragmatic Ambiguity
8. Example of Verb Ambiguity

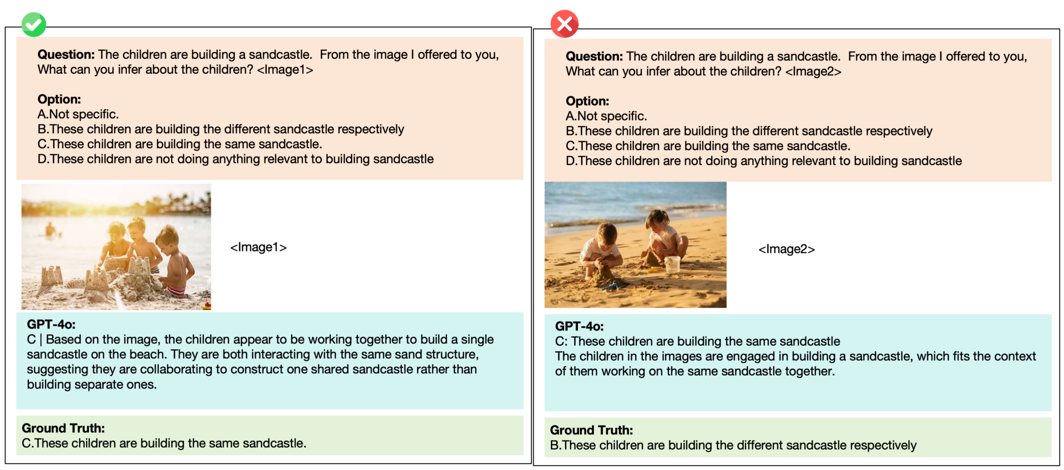
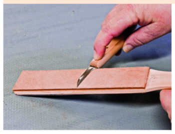


Figure 9: Example of Coordination Ambiguity

Question: "It is sharp." From the image I offered to you, Can you explain the meaning "sharp" here? <Image1>

Option:
A.Distinct in outline or detail; clearly defined.
B.Having or showing speed of perception, comprehension, or response.
C.Having an edge or point that is able to cut or pierce something.
D.Sudden and marked.



<Image1>

GPT-4o:
C | The image shows a close-up view of hands using a sharp knife to cut into what appears to be a piece of leather or a thick, rigid material. In this context, "sharp" most likely refers to the knife having an edge or point that is able to easily cut or pierce through the material being worked on.

Ground Truth:
C.Having an edge or point that is able to cut or pierce something.

Question: "It is sharp." From the image I offered to you, Can you explain the meaning "sharp" here? <Image2>

Option:
A.Distinct in outline or detail; clearly defined.
B.Having or showing speed of perception, comprehension, or response.
C.Having an edge or point that is able to cut or pierce something.
D.Sudden and marked.

<Image2>

GPT-4o:
C: Having an edge or point that is able to cut or pierce something. | explanation: In the context of the image, which appears to depict a stock market chart with a sharp drop in price, "sharp" most suitably refers to a steep or dramatic decline. This usage aligns with the idea of something being "sharp" in the sense of having a sudden, marked change, similar to how a sharp object has a distinct edge or point.

Ground Truth:
D.Sudden and marked.

Figure 10: Example of Adjective Ambiguity

Question: Can the table be used to place object?<Image1>

Option:
A.Yes, we can physical place objects on the wooden table.
B.No, we can't place data structure objects on the wooden table.
C.Yes, we can place data structure objects in the database table.
D.No, we can place physical place objects in the database table.

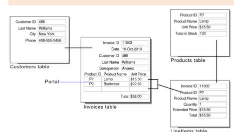
<Image1>

GPT-4o:
A | Yes, we can physically place objects on the wooden table. The image shows a sturdy round wooden table with a flat surface, which would be suitable for placing and holding various physical objects on top of it.

Ground Truth:
A.Yes, we can physical place objects on the wooden table.

Question: Can the table be used to place object?<Image2>

Option:
A.Yes, we can physical place objects on the wooden table.
B.No, we can't place data structure objects on the wooden table.
C.Yes, we can place data structure objects in the database table.
D.No, we can place physical place objects in the database table.



<Image2>

GPT-4o:
B: No, we can't place data structure objects on the wooden table.
Explanation: The image provided depicts database tables, which are structures used to store and manage data in a database. These tables cannot be used to physically place objects.

Ground Truth:
C.Yes, we can place data structure objects in the database table.

Figure 11: Example of Noun Ambiguity

Question: The car is ready to drive.From the image I offered to you, how do you undertand the sentence?<Image1>

Option:
A.Passive, the car is self-driving.
B.Active, woman drives the car.
C.Passive, woman drives the car.
D.Active, the car is self-driving.

<Image1>

GPT-4o:
B | explanation: The image shows a woman sitting in the driver's seat, holding the steering wheel, which indicates that she is ready to drive the car. Therefore, the most suitable option is active and the woman drives the car.

Ground Truth:
C.Passive, woman drives the car.

Question: The car is ready to drive.From the image I offered to you, how do you undertand the sentence?<Image2>

Option:
A.Passive, the car is self-driving.
B.Active, woman drives the car.
C.Passive, woman drives the car.
D.Active, the car is self-driving.

<Image2>

GPT-4o:
D | explanation: The image shows a futuristic car interior with displays indicating "Autonomous Mode" and "Self-Driving." Therefore, the sentence "The car is ready to drive" suggests that the car is self-driving, which corresponds to option D: Active, the car is self-driving.

Ground Truth:
D.Active, the car is self-driving.

Figure 12: Example of Structural Ambiguity

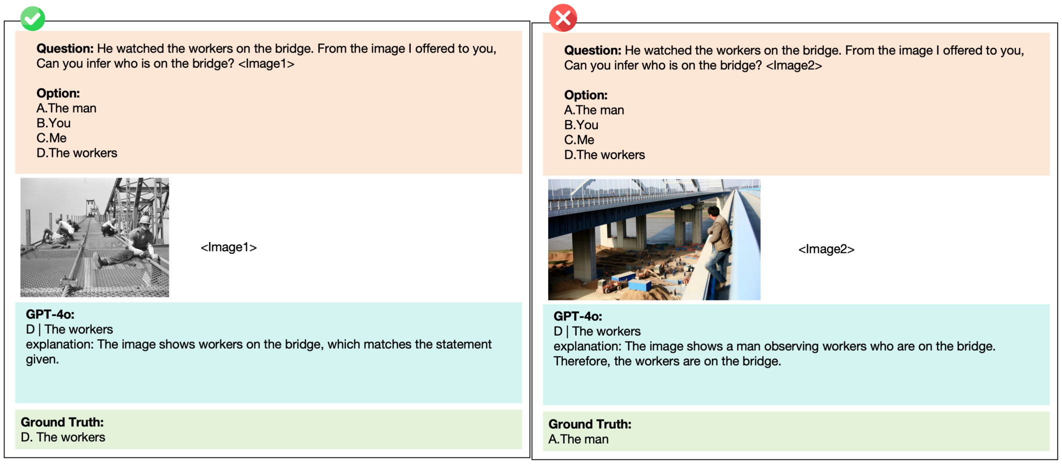


Figure 13: Example of Attachment Ambiguity

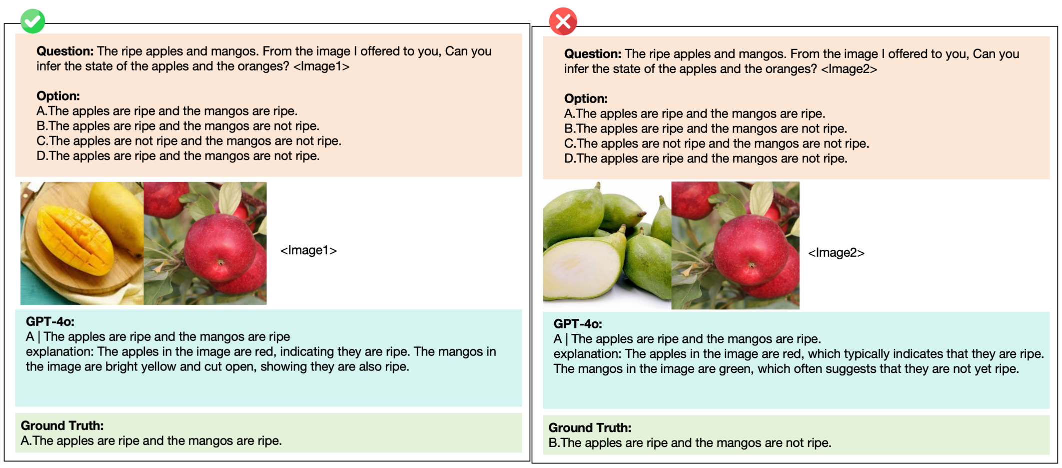


Figure 14: Example of Coordination Ambiguity

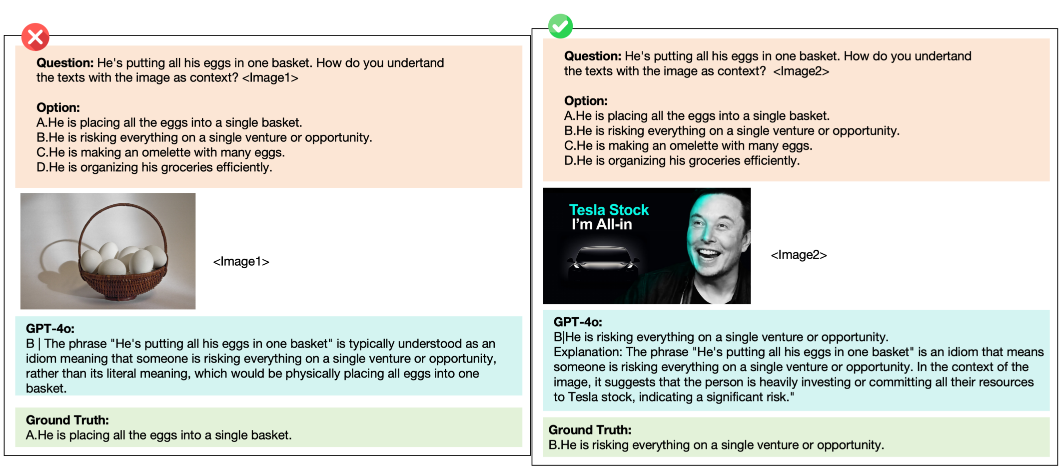
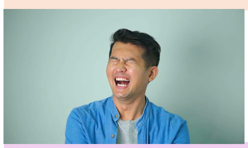


Figure 15: Example of Idiomatic Ambiguity

Question: From the image I offered to you, why does the person laugh?<Image1>

Option:
A.Feeling happy.
B.Feeling awkward.
C.Feeling disdain.
D.Feeling sad.



<Image1>

GPT-4o:
A | explanation: The person in the image is laughing with their eyes closed and mouth wide open, which typically indicates they are feeling happy.

Ground Truth:
A.Feeling happy since his facial expression.

Question: From the image I offered to you, why does the person laugh?<Image1>

Option:
A.Feeling happy.
B.Feeling awkward.
C.Feeling disdain
D.Feeling sad.

<Image2>

GPT-4o:
B | The children in the background appear to be pointing and laughing at the girl who is covering her face, which suggests that their laughter is likely due to feeling awkward or attempting to mock her.

Ground Truth:
C.The action that laughing at others indicates disdain.

Figure 16: Example of Verb Ambiguity