

Beyond Recognition: Evaluating Visual Perspective Taking in Vision Language Models

Supplementary Material

A. Data

A.1. Gold-Standard Answer

In Table 3 we list the gold-standard answers distribution.

Table 3. Gold-standard answer distribution for each question type (Q1–Q7).

Q	Gold-Standard Answers (% of items)
Q1	1 (100.0)
Q2	1 (100.0)
Q3	yes (100.0)
Q4	east (15.3), east, north (10.4), east, south (8.3), north (14.6), north, west (10.4), south (18.1), south, west (6.9), west (16.0)
Q5	east (15.3), east, south (9.7), north (25.0), south (25.0), south, west (7.6), west (17.4)
Q6	no (50.0), yes (50.0)
Q7	back (31.9), back, left (4.9), back, right (13.2), front (39.6), front, left (7.6), front, right (2.8)

A.2. Prediction Correctnesses

Models sometimes generated compound answers – for instance, *northeast* in Q4 and Q5, or *back and slightly to the left* in Q7 (see Figure 9 and Table 4). Because these responses contained multiple components, our evaluation needed to acknowledge partial as well as fully correct answers.

To tackle this, we employed a precision-based metric that rewarded models for each correctly identified component while tolerating omissions. Assume that we evaluate R responses for a given diagnostic questions and a model, like in Table 2. The score P reported in the table is called *prediction correctness* and is defined as the mean precision across all R responses:

$$P = \frac{1}{R} \sum_{i=1}^R P_i, \quad \text{where } P_i = P(M_i, G_i) = \frac{|M_i \cap G_i|}{|M_i|},$$

where M_i (resp. G_i) is the set of components in the model’s prediction (resp. gold-standard answer) for the i response.

The value of P ranges from 0 to 1. We note that for the questions with single answers, like e.g., Q6, this metric is equivalent to standard accuracy. Moreover, we experimented with several other metrics that take into account partial correctness (e.g., the Jaccard index), and they all yielded similar results.

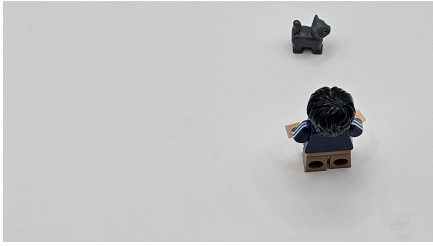
Table 4. Frequency of answer types for Q4, Q5, and Q7 across models. *Comp.* (compound) denotes compound answers (e.g., *northwest* to Q5); *Sing.* (Single) denotes single-word answers (e.g., *back* to Q7); *Disc.* (Disclaimer) marks instances in which the model failed to provide a relevant answer (e.g., claiming no object is present when one is).

Model	Comp.	Sing.	Disc.
<i>Question 4</i>			
GPT-4-o	21	123	0
GPT-4-Turbo	0	144	0
Claude 3.5 Sonnet	21	117	6
Claude 3 Sonnet	0	127	17
Llama-3.2-11B-Vision	33	111	0
Qwen3-2B-Instruct	0	144	0
Qwen3-4B-Instruct	0	144	0
Qwen3-8B-Instruct	0	144	0
Gemini Robotics ER 1.5	0	144	0
<i>Question 5</i>			
GPT-4-o	0	144	0
GPT-4-Turbo	0	144	0
Claude 3.5 Sonnet	0	144	0
Claude 3 Sonnet	0	144	0
Llama-3.2-11B-Vision	1	143	0
Qwen3-2B-Instruct	0	144	0
Qwen3-4B-Instruct	0	144	0
Qwen3-8B-Instruct	0	144	0
Gemini Robotics ER 1.5	0	144	0
<i>Question 7</i>			
GPT-4-o	10	134	0
GPT-4-Turbo	16	127	1
Claude 3.5 Sonnet	68	76	0
Claude 3 Sonnet	1	121	22
Llama-3.2-11B-Vision	2	139	3
Qwen3-2B-Instruct	0	144	0
Qwen3-4B-Instruct	0	144	0
Qwen3-8B-Instruct	0	144	0
Gemini Robotics ER 1.5	0	144	0

A.3. Random Baseline

To establish a chance performance benchmark, we defined a random baseline classifier selecting one answer uniformly at random from the N_q distinct choices available for question type q . For a given instance of question type q , the probability of this random choice being evaluated as correct is $\text{Prob}(\text{success}_q) = \frac{k_q}{N_q}$, where k_q is the number ($k_q \geq 1$) of acceptable answers specified in the gold-standard answers for that instance out of N_q total options. This formulation accurately quantifies the expected success rate of uninformed random guessing under our evaluation proto-

GPT-4-Turbo



From the perspective of the humanoid minifigure, where is the cat located relative to it? Please use terms such as front, left, right, or back.

From the perspective of the humanoid minifigure, the cat is **located to its front and slightly to the right**.

Claude 3.5 Sonnet



Assuming the top of the image is north, in which cardinal direction (i.e., north, west, east, or south) is the plant located relative to the humanoid minifigure?

The plant is **located southeast** of the LEGO minifigure. The minifigure is positioned in the upper right (northeast) part of the image, while the small plant or sprout is in the lower right (southeast) corner, diagonally down and to the left from the minifigure's position.

Figure 9. Example model responses to Q5 and Q7 questions. Left: GPT-4 Turbo answers *front and slightly to the right* (gold-standard answer: front). Right: Claude 3.5 Sonnet answers *southeast* (gold-standard answer: south).

col, accommodating multiple correct answers, based on the empirical distribution of k_q values observed in our gold-standard answers dataset (detailed in A.1, Table 3). For instance, for Q4 ($N_4 = 4$), where 64% of instances have $k_4 = 1$ (Prob = $\frac{1}{4}$) and 36% have $k_4 = 2$ (Prob = $\frac{1}{2}$), the weighted average yields a chance level for this question type of $(0.64 \times \frac{1}{4}) + (0.36 \times \frac{1}{2}) = 0.16 + 0.18 = 0.34$.

Category-level chance performance was computed by averaging the chance levels of the constituent question types. For example, the scene understanding category comprises Q1 (chance = $\frac{1}{3}$), Q2 (chance = $\frac{1}{3}$), and Q3 (chance = $\frac{1}{2}$), resulting in an average category chance level of $(\frac{1}{3} + \frac{1}{3} + \frac{1}{2})/3 = \frac{7}{18} \approx 0.389$. Following this methodology across all categories yields the following random baseline classifier levels: scene understanding (0.389), spatial reasoning (0.317), and visual perspective taking (0.411).

A.4. Co-occurrence Matrix

Our co-occurrence matrices presented in the result section show how the model's predictions line up with each gold-standard answer. To build it, we take all questions whose gold-standard answers include a particular label – for example, *north* in Q5, and within that subset simply count how often the model produced *north*, *east*, *south*, or *west*. Those four counts become the row for *north*. Because a single question can have several gold answers and the model may mention several answers at once (such as *northeast* in Q5 or *back and slightly to the left* in Q7), one question can be counted in more than one row or column, so the values in a row may exceed the total number of questions. When every question has exactly one gold label and the model also

outputs exactly one label, this co-occurrence table collapses to the ordinary single-label confusion matrix, with each row summing to the number of items.