

SPOT THE BALL: A Benchmark for Visual Social Inference

Supplementary Material

7. System Prompts

Base Prompt. Tells the model and human that the ball is removed and asks for a guess of where it would've been

```

1 The ball has been removed from this {sport} image.
  Your task is to infer the most likely location
  of the ball.
2 Respond in the following format:
3 Reasoning: <Explain where the ball is likely
  located and why.>
4 Cell: <What grid cell is the ball most likely
  located in? Respond with a label like F4.>

```

Cue-Directed Prompt In addition to Base Prompt, tells the model to consider player gaze, pose, and positions when predicting.

```

1 The ball has been removed from this {sport} image.
  Your task is to infer the most likely location
  of the ball.
2 The location of the players, where they are looking
  and their positions can help you infer the
  location of the ball.
3 Respond in the following format:
4 Reasoning: <Explain where the ball is likely
  located and why.>
5 Cell: <What grid cell is the ball most likely
  located in? Respond with a label like F4.>

```

Chain-of-Thought Prompt Contains an added intermediate social reasoning step before making the final prediction.

1. The model is asked to answer three questions to extract relevant visual information from the scene

```

1 The ball has been removed from this {sport}
  image. Your task is to infer the most
  likely location of the ball.
2 The location of the players, where they are
  looking and their positions can help you
  infer the location of the ball.
3 Respond in the following format:
4 Reasoning: <Explain where the ball is likely
  located and why.>
5 Cell: <What grid cell is the ball most likely
  located in? Respond with a label like F4.>

```

2. The model then receives both the original instruction (that the ball has been removed) and the context observations

```

1 The ball has been removed from this {sport}
  image. Here are some observations:
2 {context}
3 The above information could help you infer the
  ball's location.
4 Respond in the following format:
5 Reasoning: <Explain where the ball is likely
  located and why.>
6 Cell: <What grid cell is the ball most likely
  located in? Respond with a label like F4.>

```

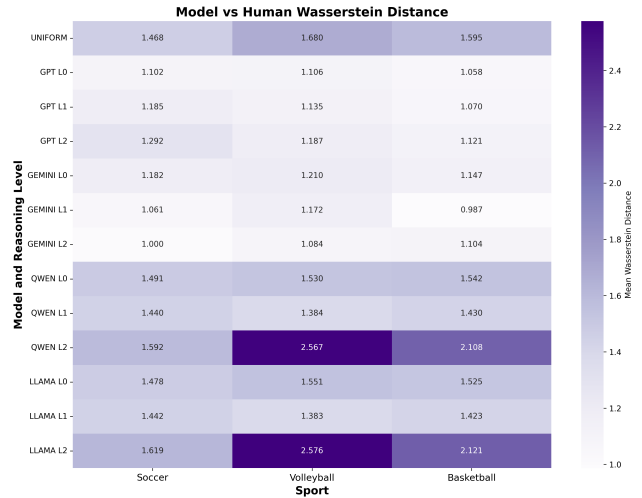


Figure 9. Wasserstein Distances with all the levels and models. L0 refers to Base prompt, L1 to cue-directed prompt and L2 to chain-of-thought prompt.

8. Evaluation Metrics Implementation

8.1. Wasserstein Distance

We used `wasserstein_distance` function from the `scipy` package with coordinate weights representing the probability mass at each grid cell to calculate the Wasserstein distances. The distances from each level of each model is shown in Figure 9.

8.2. Player Proximity

We detect players in the images via the YOLOv8 “person” class per frame and then remove audience in the background through a lightweight manual pass. The review UI shows each image with proposed boxes; annotators uncheck any non-players. The size and location of these boxes are then relevant to the proximity analysis.

Threshold fitting and robustness We determine the thresholds for the Near-Player Rate (NR) and Near-Overlap Rate (OR) metrics through a grid search over

$$\tau_{\text{near}} \in [0.04, 0.20], \quad \theta \in [0.01, 0.20],$$

where τ_{near} represents the distance threshold as a fraction of the image diagonal, and θ denotes the minimum overlap fraction between a predicted grid cell and any player box.

To guide selection, we define a *balanced objective* that

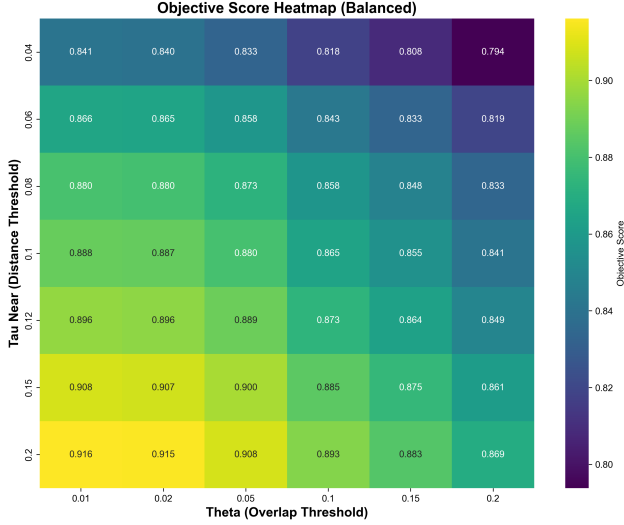


Figure 10. Balanced objective values over the grid of τ_{near} and θ . The objective peaks broadly around $\tau_{\text{near}} = 0.08$ and $\theta = 0.02$, suggesting the chosen thresholds are stable and conservative.

equally weights NR and OR:

$$\mathcal{O}(\tau, \theta) = \frac{1}{2} [\text{NR}(\tau, \theta) + \text{OR}(\tau, \theta)].$$

This formulation treats proximity to players (NR) and geometric overlap (OR) as complementary aspects of spatial behavior, ensuring thresholds that capture both close and intersecting predictions without overfitting to one metric.

The balanced objective exhibited a broad plateau centered near

$$\tau_{\text{near}} = 0.08, \quad \theta = 0.02,$$

with neighboring configurations yielding comparable scores (within 1% of the maximum; Fig. 10). Empirical distributions revealed that most predicted or true ball regions already overlap with player boxes (median normalized distance = 0; median overlap fraction ≈ 0.56), suggesting that lower thresholds would saturate the metrics. The selected values therefore represent conservative, fitted cutoffs that maintain discriminability and stability across models and sports.

9. Example model outputs

We show an example image from our dataset with the volleyball masked and grid overlaid, followed by model responses at each of the three prompting levels from Gemini and GPT (Figure 11). These examples illustrate how models interpret player positions, gaze, and team configurations to localize the ball, and how their reasoning changes with increasing prompt information.

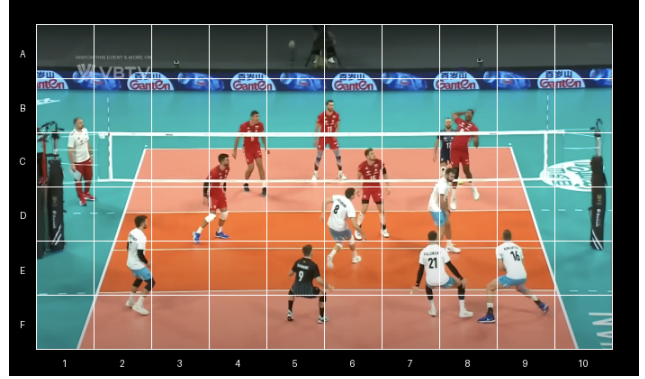


Figure 11. An example of the volleyball image shown to the models. The ball is masked, and a grid is overlaid; the responses in the following tables are generated based on this image.

10. Code and Dataset

All code required to reproduce the experiments, analyses, and figures is available at this [GitHub repository](#). The corresponding dataset, including evaluation splits and annotation metadata, can be accessed at this [HuggingFace repository](#).

Table 4. Example GPT output on the volleyball image (Base Prompt).

Reasoning	The player in the red jersey on the right side of the net is in the air with his arm extended, appearing to be in the act of spiking or hitting the ball. The other players on the red team are looking towards the same area, indicating the ball is likely near his hand. The ball would be just above the net and close to this player's hand
Grid cell guess	Cell: C9

Table 5. Example Gemini output on the volleyball image (Base Prompt).

Reasoning	The player in cell D4 is in the middle of spiking the volleyball. His arm is extended upward and towards the net. Therefore, the ball is most likely in that vicinity as well.
Grid cell guess	Cell: D4

Table 6. Example GPT output on a volleyball image (Cue-Directed Prompt).

Reasoning	The players in white are focused and positioned as if they are about to receive or set the ball. The player in white at cell E8 is looking upwards and has his hands raised, indicating he is preparing to make contact with the ball. The players in red on the opposite side are also looking towards the same area, suggesting the ball is above the net near the middle of the court. Given the positioning and gaze of the players, the ball is most likely just above the net around the center-right area.
Grid cell guess	Cell: C7

Table 7. Example Gemini output on a volleyball image (Cue-Directed Prompt).

Reasoning	The player at position 6 in the D row is clearly hitting the ball. Given that the player at the front net at the C row is looking down as well as the players in the bottom right, it would be likely the ball is in the air after the player at position 6 in the D row is hitting it.
Grid cell guess	Cell: D6

Table 8. Example GPT output for the volleyball image (Chain-of-Thought Prompt).

<p>Where are the players located?</p>	<p>The players are located at the following grid positions:</p> <ul style="list-style-type: none"> - Player in white jersey near the left side: E2 - Player in white jersey near the center: E6 - Player in white jersey near the right side: E8 - Player in black jersey near the center: E5 - Player in black jersey near the right side: E7 - Player in red jersey near the left side: C3 - Player in red jersey near the center-left: C5 - Player in red jersey near the center: C6 - Player in red jersey near the center-right: C7 - Player in red jersey near the right side: C8
<p>Where are the players looking?</p>	<p>The players are looking in the following directions:</p> <ul style="list-style-type: none"> - The player in white at E2 is looking towards the net, slightly to the right. - The player in white at E6 is looking upwards and slightly to the right. - The player in white at E7 is looking upwards and slightly to the left. - The player in white at E9 is looking upwards and slightly to the left. - The player in black at E5 is looking upwards and slightly to the right. - The player in red at C4 is looking upwards and slightly to the right. - The player in red at C6 is looking upwards and slightly to the left. - The player in red at C7 is looking upwards and slightly to the left. - The player in red at C8 is looking upwards and slightly to the left. - The player in red at B9 is looking upwards and slightly to the left. <p>Overall, most players are looking upwards, likely tracking the volleyball in play.</p>
<p>How are the players positioned?</p>	<p>The players are positioned as follows:</p> <p><i>Team in red (top side of the net):</i></p> <ul style="list-style-type: none"> - One player at B4 - One player at B6 - One player at B8 - One player at C3 - One player at C5 - One player at C7 <p><i>Team in white and blue (bottom side of the net):</i></p> <ul style="list-style-type: none"> - One player at E2 - One player at E6 - One player at E7 - One player at E8 - One player at F5 - One player at F6
<p>Reasoning</p>	<p>Most players are looking upwards, indicating the ball is in the air. The players in white and black near the center (E5, E6, E7) are looking upwards and slightly to the right or left, suggesting the ball is above or near the net. The red team players near the net (C5, C6, C7, C8) are also looking upwards and slightly to the left, indicating the ball is likely above the net or just past it on the white team's side. Given the positioning and gaze directions, the ball is most likely just above the net around the center, near the grid cell between E5 and F5.</p>
<p>Grid cell guess</p>	<p>E5</p>

Table 9. Example Gemini output on the volleyball image (Chain-of-Thought Prompt).

<p>Where are the players located?</p>	<p>Here are the locations of the players in the image:</p> <ul style="list-style-type: none"> * **1, C***: Player in white and red. * **2, D***: Player in white and blue. * **3, D***: Player in red. * **4, C***: Player in red. * **5, E***: Player in dark blue and white. * **5, B***: Player in red. * **6, D***: Player in white and blue. * **6, C***: Player in red. * **7, E***: Player numbered 21 in white and blue. * **8, C***: Player in red. * **8, D***: Player in white and blue. * **9, E***: Player numbered 16 in white and blue. * **9, B***: Player in red. * **10, C***: Player in white and blue.
<p>Where are the players looking?</p>	<p>Here's an analysis of where the players are likely looking in the image, keeping in mind that it's difficult to be 100% certain without seeing their eyes clearly:</p> <p>Players on the left side (likely about to receive the ball): They are likely looking at the ball or the opposing players about to hit the ball. Some of them might be tracking the anticipated trajectory of the ball.</p> <p>Players on the right side (likely having just hit the ball): They are most likely watching the ball after hitting it and looking at the opposing team to see how the receiving team will handle it.</p>
<p>How are the players positioned?</p>	<p>Here is the approximate positioning of the players based on the grid:</p> <p>Team 1 (Left Side):</p> <ul style="list-style-type: none"> * Player 1: A-1/2 * Player 2: D-2/3 * Player 3: C-4 * Player 4: D-6 <p>Team 2 (Right Side):</p> <ul style="list-style-type: none"> * Player 1: B-9 * Player 2: D-8 * Player 3: E-7 * Player 4: F-9 <p>Mid-Court Players:</p> <ul style="list-style-type: none"> * Player 5: E-5 * Player 6: B-5 * Player 7: C-6 * Player 8: B-8 * Player 9: C-4
<p>Reasoning</p>	<p>Considering the positions of the players on both sides of the net and how they are looking, it appears the ball was hit from the right side (Team 2). Player E7 is watching the ball and is looking at it coming towards his side of the court. The players on the left side (Team 1) are getting ready to receive. With that knowledge I can assume that the ball is in the E4 grid position.</p>
<p>Grid cell guess</p>	<p>Cell: E4</p>