

# SPOT THE BALL: A Benchmark for Visual Social Inference

Neha Balamurugan<sup>1</sup>, Sarah Wu<sup>2</sup>, Adam Chun<sup>1</sup>, Gabe Gaw<sup>1</sup>, Cristóbal Eyzaguirre<sup>1</sup>, Tobias Gerstenberg<sup>2</sup>

Stanford University

<sup>1</sup>Department of Computer Science    <sup>2</sup>Department of Psychology

{nbalamur, gerstenberg}@stanford.edu

## Abstract

Humans excel at *visual social inference*, the ability to infer hidden elements of a scene from subtle behavioral cues such as other people’s gaze, pose, and orientation. This ability drives everyday social reasoning in humans and is critical for developing more human-like AI agents. We introduce SPOT THE BALL, a challenging benchmark for evaluating visual social inference in vision–language models (VLMs) using sports as a test domain. The task is to localize a removed sports ball from soccer, basketball, and volleyball images. We present a curated evaluation set with human baselines and a scalable pipeline for generating additional test items. We evaluate four state-of-the-art VLMs (Gemini, GPT, LLaMA, Qwen) using three prompting strategies, finding that humans are consistently two to three times more accurate (20–34%) than models ( $\leq 17\%$ ) across all sports. Our analyses show that models rely on superficial spatial heuristics—such as guessing near the image center or nearby players—while humans leverage social cues like gaze direction and body pose. These findings reveal a persistent human–model gap in visual social reasoning and underscore the need for architectures that explicitly encode structured behavioral cues to achieve robust, human-like inference.

When someone scans the floor with narrowed eyes, we infer they must be searching for something. When a friend approaches us with open arms, we anticipate a hug. As humans, we readily use subtle behavioral cues, such as gaze, pose, and orientation to infer implicit information. This ability is rooted in our *theory of mind* (ToM), the capacity to reason about others’ beliefs, desires, and intentions to predict behavior (Byom and Mutlu, 2013; Dennett, 1989; Wellman, 2014). ToM is fundamental to everyday interaction with other humans, demonstrated by the challenges faced by populations with impaired social reasoning, like individuals on the autism spectrum (Baron-Cohen, Leslie, and Frith, 1985; Fletcher-Watson et al., 2014; Tager-Flusberg, 2007). AI systems that cannot reason socially may behave unsafely when interacting with humans: imagine a robot nurse that misinterprets a patient’s gesture, or an autonomous car that fails to anticipate that a pedestrian is about to cross the street. As large language models (LLMs) and vision–language models (VLMs) are increasingly deployed in socially grounded tasks, their need for accurate social reasoning is critical.

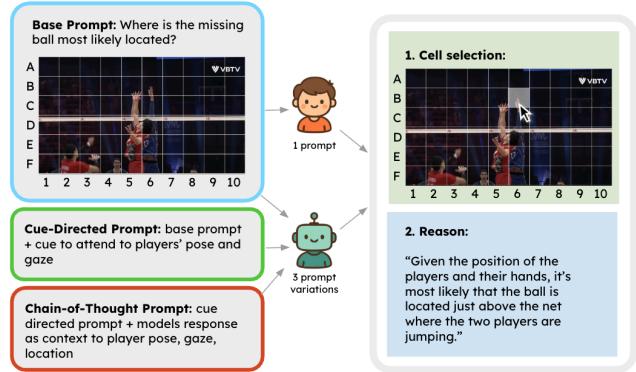


Figure 1: Overview of the SPOT THE BALL task. Given an image with the ball removed, humans and models infer the likely location by reasoning about player pose and gaze. Models are prompted under three conditions, whereas humans receive only the base prompt.

While LLMs perform some types of social reasoning based on text relatively well (Gandhi et al., 2023), such behavior may be dependent on verbal priors and pattern matching rather than robust modeling social cognition. To test whether social inference is grounded in perception, we study VLMs. Concretely, VLMs allow us to (i) force models to rely on visual cues (gaze, pose, orientation) rather than language heuristics, (ii) evaluate an AI capability that is directly relevant to embodied and safety-critical applications (assistive robots, driver assistance, collaborative tools), and (iii) pose a challenging task with a well-defined correct answer.

Some existing benchmarks for social reasoning evaluate models on fully visible scenes (Mathur et al., 2025; Xu et al., 2025), while others emphasize inanimate objects (Liu et al., 2025). Both approaches fail to capture the ecological validity of dynamic social contexts in which humans reason under conditions of partial information. To address this gap, we introduce SPOT THE BALL—inspired by a classic newspaper puzzle—to evaluate VLMs’ capacity for visual social inference in real-world, socially grounded images. Ball sports provide a challenging setting, as every player’s gaze, posture, and positioning are causally coupled to the ball and to the other

players, yielding complex but interpretable signals. To further isolate social reasoning from low-level dynamics, we employ static images rather than video, thereby removing dynamic motion cues. Our contributions are as follows:

1. A curated evaluation set of 150 modified sports images with human baselines.
2. A systematic evaluation of four leading VLMs under multiple prompting strategies.
3. A scalable pipeline for generating ball sports reasoning tasks, using which we have produced 3,000 additional soccer images for training and analysis.

Together, these contributions provide the first structured evaluation of whether VLMs can leverage social cues to infer hidden objects in complex, real-world scenes. Our results reveal large, consistent gaps between human and model performance.

## Related Work

**Human social reasoning.** Humans acquire social reasoning abilities early in life, learning from infancy to represent and infer others’ beliefs, desires, and intentions (Wellman, 2014; Sodian, 2011; Wang, Davis, and Jara-Ettinger, 2025). By one year old, infants understand not only reaching actions as goal-directed, but also actions they have not yet personally experienced (Liu and Almeida, 2023), even when there is uncertainty about the goals themselves (Woo, Liu, and Spelke, 2024). AI systems that are employed alongside humans must also have social reasoning capabilities to ensure seamless interaction (Mathur, Liang, and Morency, 2024). Adults readily ascribe and infer rich social states from visual stimuli as impoverished as moving geometrical shapes (Heider and Simmel, 1944; Tremoulet and Feldman, 2000), as well as more complex visual cues such as gaze (e.g. Kleinke, 1986), facial expression (e.g. Todorov et al., 2015), and body expression (e.g. de Gelder, de Borst, and Watson, 2015).

**Text-based social reasoning.** In NLP, a range of benchmarks examine social reasoning through text-only settings, capturing different facets of social cognition. Theory-of-mind (ToM) evaluations (Wu et al., 2023b; Ma, Gao, and Xu, 2024; Gandhi et al., 2023; Chan et al., 2024; Chen et al., 2024d; Ullman, 2023), such as false belief tasks (where observers have to predict how someone with an incorrect belief will act), test models’ ability to infer hidden beliefs, desires, and intentions. Empathy and emotional intelligence datasets assess affective understanding and perspective-taking (Chen et al., 2024c; Huang et al., 2024), while moral and ethical reasoning tasks probe alignment with human values (Marcuzzo et al., 2025; Scherrer et al., 2023; Jiao et al., 2025; Trager et al., 2025; Ji et al., 2025b). Other lines of work target deception detection (Chen et al., 2024b; Ji et al., 2025a; Krishna et al., 2025), negotiation and dialogue-based cooperation (Abdelnabi et al., 2024; Sap et al., 2019; Choi et al., 2023; Xu et al., 2025), as well as fairness, bias, and social influence (Jung et al., 2025; Gassmann, Campbell, and Edwards, 2024). While text is an important source of information, much of our social inferences rely primarily on visual information.

**Video-based social reasoning.** Social reasoning in dynamic visual contexts has been studied through a growing set of video-based benchmarks that capture motion, gaze, and causality. Theory-of-mind reasoning has been explored using both human-interaction videos (Jin et al., 2024; Shi et al., 2025; Li et al., 2025; Anonymous, 2025) and synthetic multi-agent simulations (Fan et al., 2025). A broad class of video question-answering tasks (Zhong et al., 2022) extend ToM toward causal and counterfactual inference, including temporal and explanatory reasoning datasets (Xiao et al., 2021; Pătrăucean et al., 2023; Chandrasegaran et al., 2024), as well as real-world causal understanding benchmarks (Wu et al., 2023a; Li, Niu, and Zhang, 2022; Foss et al., 2025; see also Vicol et al., 2018 for a long-form video dataset, and Ren et al., 2023; Luo et al., 2019 which focus on affective information). Further, benchmarks like Social-IQ 1.0 (Zadeh et al., 2019) and Social-IQ 2.0 (Wilf et al., 2023) directly evaluate question answering about social interactions, while newer efforts incorporate grounded knowledge or social media data (Kong et al., 2025; Niu et al., 2025). Finally, interactive, agentic environments (Puig et al., 2023; Chen et al., 2023) aim to probe multi-agent collaboration and embodied social behavior.

**Image-based social reasoning.** Compared to video settings, there are fewer benchmarks that study social reasoning from static images. Visual Commonsense Reasoning (VCR) (Zellers et al., 2019) introduced socially grounded question answering over images containing people and objects, requiring inference of intentions and interactions. Similarly, KiVA (Yiu et al., 2024) extends this to more complex multimodal reasoning over image-text pairs. Broader multimodal datasets (Lin et al., 2025b; Yao et al., 2025) demonstrate how static visual understanding provides social and contextual cues, though they typically operate in fully visible, non-occluded settings. Other benchmarks target specific social reasoning facets, including facial expression understanding (Mollahosseini, Hasani, and Mahoor, 2019; Kosti et al., 2017), social relationship recognition (Li et al., 2018), and moral or norm violation detection (Lin et al., 2025a; Hu et al., 2024).

**Reasoning under missing information.** Some datasets study inference about occluded or missing scene elements (Pothiraj et al., 2025), but they do not leverage social information. Others target social inference but ask predictive questions about intentions (Mathur et al., 2025) rather than reasoning about unobserved elements. Thus, no existing benchmark requires models to infer *hidden* information based on social cues.

**Our contribution.** We introduce a nonverbal, image-based benchmark where inferring the missing element requires both social and physical reasoning about the agents in the scene. Unlike prior work, our task is zero-shot, purely visual, and requires models to integrate pose, gaze, and orientation cues to infer what is absent. This makes the problem both ecologically valid and uniquely challenging.

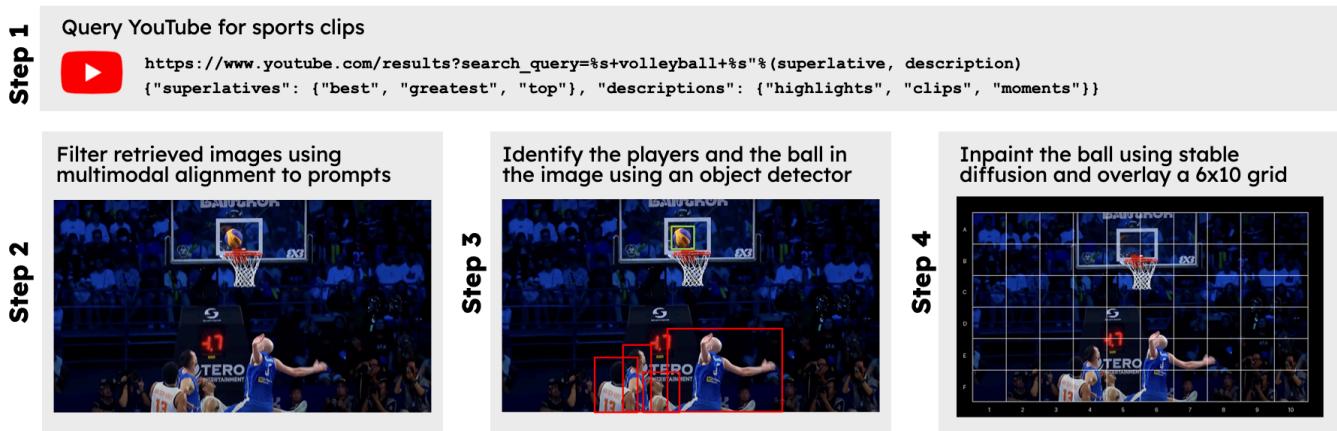


Figure 2: **Pipeline for constructing the SPOT THE BALL dataset.** We retrieve and filter sports footage from YouTube by alignment to the prompts, detect players and balls using an object detector, and inpaint the ball region with stable diffusion before overlaying a  $6 \times 10$  grid for location annotation.

### SPOT THE BALL

In SPOT THE BALL, the objective is to infer the location of a removed ball in a sports frame (see Figure 1). This task evaluates a model’s ability to localize a hidden object through reasoning over social and physical contextual cues such as players’ gaze, body orientation, and spatial positioning, rather than relying on direct visual evidence of the object itself in addition to sport specific knowledge.

#### Evaluation Set

We curated an evaluation set of 150 images drawn from publicly available broadcast footage of soccer, basketball, and volleyball from YouTube (see Figure 2). Frames were selected procedurally to maximize contextual informativeness (i.e., non-occluded presence of a ball, number and distribution of players etc.) with a light sanity check to rid any pipeline errors. The ball was removed using an inpainting procedure after recording its ground truth location. This evaluation set serves as the primary benchmark for comparing models against human reasoning.

To scale beyond the evaluation set, we developed a modular pipeline to generate realistic, inpainted sports scenes for evaluation. Below we describe the logic:

- 1. Video retrieval.** Broadcast soccer, basketball, and volleyball footage was retrieved from YouTube using sport-specific queries concatenated with keywords (e.g., “best”, “greatest”, “highlights”, “moments”) to bias towards action-focused clips. The resulting videos were decoded with OpenCV (Bradski, 2000), and frames were sampled at fixed intervals ( $\sim 1$  FPS) to ensure temporal sparsity.
- 2. Frame extraction.** To avoid collecting irrelevant or redundant material, each frame was scored with CLIP (Radford et al., 2021) against prompts such as “picture of volleyball players in action with ball”. Only frames exceeding a similarity threshold were retained.
- 3. Ball detection.** Retained frames were passed through YOLOv8 to detect both the ball and surrounding play-

ers (Jocher, Chaurasia, and Qiu, 2023). Detections were filtered by class confidence and spatial plausibility: we required exactly one ball per frame, non-overlapping and proximal to player bounding boxes in some cases. This eliminated spurious detections and ensured that contextual cues were preserved.

4. **Ball masking.** For each valid detection, the ball region was removed and filled using a Stable Diffusion inpainting model. Compared to simple masking or classical inpainting (e.g., Telea’s fast marching method (Telea, 2004)), diffusion-based inpainting generates realistic textures and lighting, avoiding visible artifacts. Player masks were applied to prevent painting over them, ensuring that body posture, gaze, and orientation cues remained intact. A light manual check was performed to ensure no visible artifacts (e.g., shadows of the ball) remained. This procedure guarantees that the ground-truth ball location is precisely known while producing naturalistic occluded scenes.

Finally, each image is overlaid with a  $6 \times 10$  alphanumeric grid (rows A–F, columns 1–10). The ground-truth label corresponds to the coordinate(s) covering the original ball location (e.g., [A5] when the ball was in a single square, or [A5, A6, B5, B6] when a ball overlapped with four squares).

The pipeline yields paired ‘original & inpainted’ images with standardized metadata. By batching across thousands of candidate frames, we generated 3,000 occluded soccer images in addition to the curated evaluation set. The modular design also allows extensions to other ball sports or alternative difficulty controls (e.g., varying player density or occlusion severity).

### Experiments

To assess visual social inference in models and humans, we test four VLMs using three prompting strategies under three sports in the SPOT THE BALL task and compare them to human performance. The participants, both humans and models,



**Figure 3: Player density and coverage across sports in the SPOT THE BALL dataset.** In Soccer frames (A) player count and area are intermediate compared to the other sports. Volleyball frames (B) feature the most players but each occupies a smaller visual area, providing weaker pose and gaze cues. Basketball (C) has fewer yet larger players, offering clearer postural and gaze information.

select one grid cell (e.g., ‘‘B6’’) in addition to a text reasoning. Predictions are evaluated against the ground-truth set of valid cells. Multiple adjacent cells may be considered correct if they overlap with the ball region. Then, we compare model and human performance across several quantitative and behavioral metrics to identify accuracy trends and reasoning patterns.

## Models

We evaluate four multi-modal instruction-following models:

1. Gemini-2.0-flash-001
2. GPT-4.1-mini
3. LLaMA-3.2-11B-Vision-Instruct
4. Qwen-2.5-VL-7B-Instruct

This set spans both closed and open-weight paradigms, all supporting high-resolution vision inputs and free-form text reasoning.

While the specifics of the proprietary models are not publicly disclosed, available evidence suggests that both Gemini-2.0-flash-001 and GPT-4.1-mini adopt unified transformer backbones that fuse visual and textual representations through shared cross-modal attention layers.

LLaMA-3.2-Vision-Instruct and Qwen-2.5-VL-7B-Instruct both pair pretrained language models with Vision Transformer-based encoders that extract image features and integrate them into the text model using adapters to enable reasoning across modalities. Qwen-2.5 handles native-resolution inputs by using convolutional and windowed-attention blocks.(Bai et al., 2025). Together, these architectures represent a spectrum from fully integrated multimodal transformers to adapter-based and hierarchical fusion strategies.

## Domains

We selected soccer, volleyball, and basketball because they are ball sports that are present in high-frequency in pretraining corpora, allowing the models to have an understanding of the game mechanics. Further, these sports differ from each other uniquely in how many players they contain (we use clips of 3v3 basketball) and how long the ball can be with a player and these differences in the mechanics lead to the footage of these

sports having variations in the amount of information and visual density. This variation allows us to analyze how models generalize under different types of visual ambiguity which might be relevant to model downfalls (Table 1, Figure 3).

## Prompting Strategies

We test the models on three variations of prompts that are provided in addition to the encoded image:

- **Base Prompt:** Instruction to provide the cell location of the missing ball.
- **Cue-Directed Prompt:** The Base Prompt with the additional cue to focus on players’ pose and gaze.
- **Chain-of-Thought Prompt (CoT):** First, we ask one-shot questions about the players’ location, pose and gaze (3 questions total). Next, the responses to these, are provided as context and asked to predict the grid cell.

To estimate distributional behavior, we sample  $n = 50$  predictions per image at Base Prompt and Cue-Directed Prompt, and  $n = 20$  at Chain-of-Thought Prompt, all at temperature  $T = 0.6$  for all models. These strategies are motivated by prior findings that CoT improves performance on spatial and visual reasoning (Li et al., 2024; Chen et al., 2024a), and that auxiliary cues can enhance localization (Pothiraj et al., 2025). Appendix contains system prompts.

## Human experiment

We developed a human baseline on each sport by running behavioral experiments on Prolific, the online crowd-sourcing platform. The experiment for the soccer domain was pre-

**Table 1: Statistics across 3 sports in the evaluation set.**

	Soccer	Volleyball	Basketball
Avg. grid cells ball spans	2.27	2.08	2.42
Avg. ball pixel area	785.37	904.86	1631.27
Avg. distance of ball from center	122.81	188.02	163.30
Avg. players in scene	4.26	9.92	5.46
Player coverage (pixels)	13718.87	6600.26	20067.43

registered on the [Open Science Framework](#).<sup>1</sup> The volleyball and basketball conditions were identical.

Participants were first guided through task instructions with two versions of an example image: one showing the ball, and one with the ball removed. For each image in the test set, participants made three guesses for the ball location by clicking on grid cells on the screen. They were allowed to click on the same cell multiple times. For attention checks, we included two additional images where the ball was *not* masked, and instructed participants to simply click on the ball for those. Any participant that clicked outside of the cells containing the ball was excluded. In total, each participant saw 52 images (50 test and 2 attention check) from one domain in a randomized order. The average time to complete the task was 17.3 minutes (SD = 7.4).

We recruited 176 participants and compensated them at a base rate of \$12/hour. Each participant earned an additional bonus of up to \$1 based on the accuracy of their responses. They were randomly assigned to one of the three domains: soccer, basketball, or volleyball. 26 participants were excluded for failing attention checks, leaving a final sample of  $N = 150$  with 50 for each sport. The demographic distribution of the final sample was: *age*: M = 42, SD = 13; *gender*: 74 male, 74 female, 2 non-binary; and *race*: 104 White, 37 Black/African American, 5 Asian, 1 Multiracial, and 1 undisclosed.

## Evaluation Metrics

We organize our evaluation metrics into three categories: (1) **Task Performance** assesses how accurately models localize the ball; (2) **Alignment with Humans** assesses whether model predictions resemble human reasoning; and (3) **Behavioral Strategies** assesses what strategies models and people rely on for solving the task.

### Task performance

1. Accuracy: For each image  $i$ , let  $\mathcal{G}_i$  be the set of ground-truth grid cells and  $\hat{y}_i$  the predicted cell. Accuracy is the proportion of predictions landing in any valid cell:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[\hat{y}_i \in \mathcal{G}_i].$$

2. Euclidean Error: We measure the minimum distance in pixels from the predicted cell center  $c(\hat{y}_i)$  to the nearest ground-truth cell:

$$d_i = \min_{g \in \mathcal{G}_i} \|c(\hat{y}_i) - c(g)\|_2.$$

The reported score is the mean of  $d_i$  across all images.

### Alignment with humans

1. Wasserstein Distance: To assess whether the distribution of model guesses resembles that of human guesses, we

---

<sup>1</sup>The model prompts were updated from the pre-registration based on additional piloting.

compare their empirical distributions  $P$  and  $Q$  over the 6×10 grid with the Earth Mover’s Distance:

$$\begin{aligned} W(P, Q) &= \min_{\gamma \geq 0} \sum_{j,k} \gamma_{jk} D_{jk} \\ \text{s.t. } \sum_k \gamma_{jk} &= P_j, \quad \sum_j \gamma_{jk} = Q_k, \end{aligned}$$

where  $D_{jk}$  is the Euclidean distance between cell centers  $j, k$ . This was calculated using the `wasserstein_distance` method from `scipy` package in python. Lower values indicate distributions more similar to human strategies.

## Behavioral strategies

1. Distance to players: For image  $i$ , let  $\mathcal{B}_i = \{b_{i,m}\}$  be the player boxes, with (axis-aligned) box centers  $u_{i,m}$  and support regions  $\text{supp}(b_{i,m})$ . Let  $p_{i,t} = c(\hat{y}_{i,t})$  be the predicted cell center. We use the standard point-to-rectangle distance:

$$\text{dist}(x, b) = \begin{cases} 0, & x \in \text{supp}(b), \\ \text{Eucl. dist. from } x \text{ to } \text{supp}(b), & \text{otherwise.} \end{cases}$$

- Near Player Rate (NR): With threshold  $\tau \in (0, 1)$  expressed as a fraction of image diagonal  $D$ ,

$$NR = \frac{1}{\sum_i T_i} \sum_{i,t} \mathbb{1} \left[ \min_{b \in \mathcal{B}_i} \text{dist}(p_{i,t}, b) \leq \tau D \right]$$

We use  $\tau = 0.08$  (8% of the diagonal).

- Near Overlap Rate (OR): A prediction “overlaps a player” if its grid cell region  $c(\hat{y}_{i,t})$  intersects any player box by at least a fraction  $\theta$  of the cell area:

$$OR = \frac{1}{\sum_i T_i} \sum_{i,t} \mathbb{1} \left[ \max_{b \in \mathcal{B}_i} \frac{\text{area}(c(\hat{y}_{i,t}) \cap b)}{\text{area}(c(\hat{y}_{i,t}))} \geq \theta \right]$$

We use  $\theta = 0.02$ .

2. Center Ratio and Entropy. Let  $p$  be the distribution of model guesses and  $q$  the empirical prior of ground-truth ball locations. For central window  $\mathcal{S} = \{(r, c) : r \in \{2, 3, 4\}, c \in \{3, 4, 5, 6, 7\}\}$  (3×5 block,  $|\mathcal{S}| = 15$ ), the center ratio is

$$CR = \frac{\sum_{j \in \mathcal{S}} p_j}{\sum_{j \in \mathcal{S}} q_j},$$

where  $R > 1$  indicates more predictions in the central window than the ground truth. Prediction spread is quantified with normalized entropy:

$$\hat{H}(p) = \frac{-\sum_{j=1}^{60} p_j \log p_j}{\log 60}.$$

Higher values  $\hat{H}(p)$  indicate broader exploration across grid cells; lower values indicate concentration on a smaller subset.

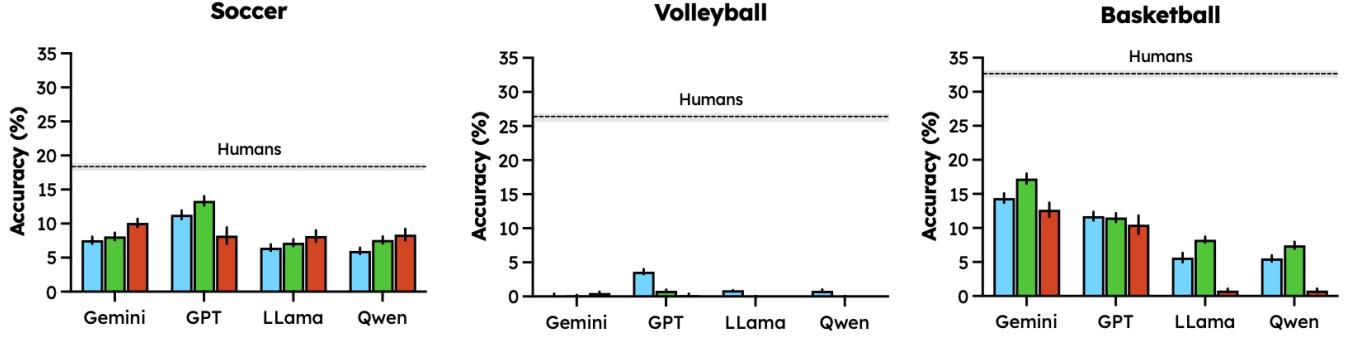


Figure 4: **Accuracy**. Model accuracy in each sport under different prompting strategies (blue = base prompt, green = cue-directed prompt, red = chain-of-thought prompt). The dashed line shows human accuracy using the base prompt in the given sport. Error bars and gray ribbon show 95% bootstrapped confidence intervals.

## Results and Discussion

We evaluate humans and four VLMs across three sports and three prompting strategies based on accuracy, spatial error, and distributional analyses. The overarching finding is a large human–model gap in both accuracy and approach. Because ball sports are highly structured and heavily represented in web-scale pretraining, a lack of generic world knowledge is an unlikely driver of errors. Instead, we aim to discern if the models fail at either **identifying** the relevant social cues, **extracting** them or **composing** them to localize the location of the missing ball.

## Quantitative Performance

**Humans outperform all models by a large margin.** Humans consistently outperform models in predicting the ball’s location. Across sports, human accuracy ranges from 19–34%, while all models remain at or below 17% (see Figure 4). The

accuracy gap is not due to the models being more likely to produce close misses. The Euclidean errors in Table 2 show that model predictions are often far from the true location. The distances from the correct locations are larger for models than for humans. In volleyball, where humans are most precise ( $72.0 \pm 40.1$  pixels), the models’ error is about twice as large on average. Moreover, performance does not always improve with richer prompts. In fact, for models like Llama and Qwen, Chain-of-Thought prompts amplify errors in certain cases (e.g., LLaMA reaching  $272.6 \pm 50.7$  pixels in volleyball).

**Models and humans find different sports challenging.** Performance across sports differs between humans and models. Humans perform best in basketball, worse in volleyball, and worst in soccer, while models perform similarly in basketball and soccer but struggle most in volleyball. This discrepancy suggests that humans and models rely on different visual cues to infer ball location.

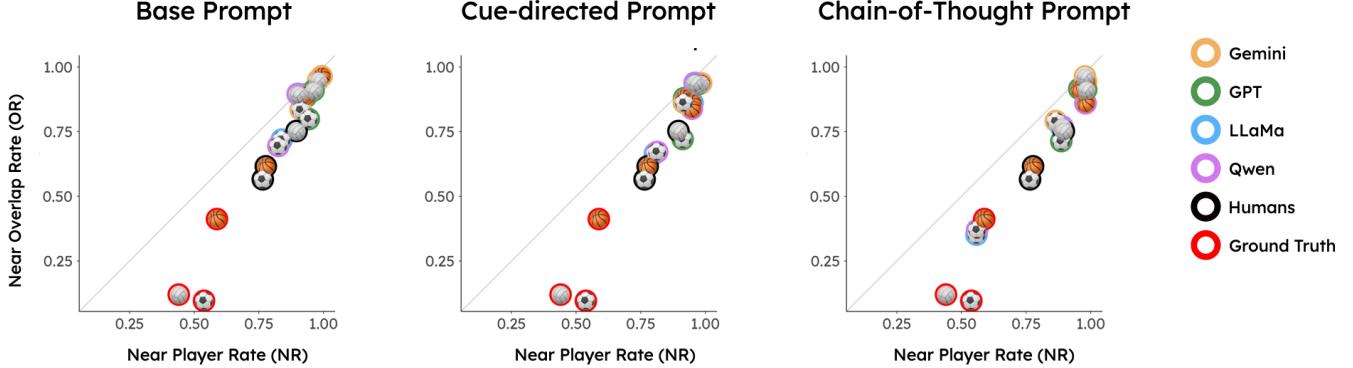
As shown in Table 1, basketball scenes feature fewer players (5.5 on average) who occupy the largest proportion of the frame (~20,000 pixels per player), making pose and gaze cues clearer and likely contributing to the highest human accuracy. Volleyball, by contrast, includes nearly twice as many players (9.9 on average) but with much lower per-player pixel density (~6,600), reducing the salience of individual cues. Nevertheless, humans may still aggregate directional information across multiple players, leading to intermediate performance compared to soccer, where both the number of players and their coverage fall in between.

Models, however, perform poorly in volleyball partly because the ball is rarely in contact with players as it is struck rather than held, making a “guess-near-player” heuristic unreliable. As shown in Figure 5, models (90%) are more likely than humans (65–75%) to predict that the ball lies near a player, a bias that fails in volleyball where the ball often travels away from them.

**Both model and human guesses are center biased.** The percentages of ground truth cells in the center window are: soccer (36.3%), volleyball (46.2%), basketball (56.3%). As

Table 2: **Mean Euclidean error** ( $\pm$  std) for humans and models across three sports and prompting types. Base = base prompt, Cue = cue-directed prompt, and CoT = chain-of-thought prompt. Lower scores reflect closer predictions to the true ball location.

Model	Prompts	Soccer	Volleyball	Basketball
Human	Base	$113.4 \pm 65.1$	$72.0 \pm 40.1$	$68.5 \pm 40.8$
	Base	$139.1 \pm 79.2$	$151.9 \pm 54.9$	$132.2 \pm 81.4$
	Cue	$133.3 \pm 75.3$	$151.5 \pm 51.5$	$119.3 \pm 75.7$
Gemini	CoT	$141.1 \pm 72.9$	$150.5 \pm 48.2$	$134.7 \pm 73.4$
	Base	$135.6 \pm 79.4$	$142.7 \pm 58.5$	$127.7 \pm 69.8$
	Cue	$139.6 \pm 88.6$	$148.6 \pm 52.2$	$125.3 \pm 70.2$
GPT	CoT	$146.1 \pm 68.3$	$155.3 \pm 56.8$	$137.1 \pm 68.8$
	Base	$143.3 \pm 79.9$	$172.8 \pm 61.4$	$161.6 \pm 95.3$
	Cue	$147.0 \pm 83.1$	$163.7 \pm 67.2$	$147.2 \pm 90.0$
LLaMA	CoT	$140.2 \pm 87.1$	$272.6 \pm 50.7$	$211.4 \pm 82.6$
	Base	$142.6 \pm 80.2$	$170.9 \pm 60.7$	$162.9 \pm 95.9$
	Cue	$147.6 \pm 81.9$	$162.9 \pm 66.6$	$147.2 \pm 90.7$
Qwen	CoT	$139.0 \pm 81.0$	$271.5 \pm 52.9$	$211.0 \pm 82.5$



**Figure 5: Player Proximity Analysis.** Each point corresponds to a model–sport combination. The  $x$ -axis shows the fraction of guesses within a fixed distance threshold of any player (Near Player Rate), while the  $y$ -axis shows the fraction of guesses whose predicted cell overlaps a player bounding box (Near Overlap Rate). 52.2% of ground truth balls are near players, 20.9% of the ground truth balls are near players by overlap.

noted in Table 3, models exhibit a strong bias toward predicting central grid cells, reflected in elevated center ratios ( $R > 1.2$ ). Humans also show a high center bias ( $R = 1.63$ ) and exhibit higher normalized entropy ( $\hat{H} = 0.855$ ) than all models (0.698–0.808). So even though humans are more likely to guess that the ball is in the center compared to models, their answers also exhibit more variance, too. This suggests that humans consider more possibilities than models do. Models may rely on simpler strategies like “guess near a player” or “guess near the center”, despite a temperature of 0.6 and repeated sampling.

**Models and humans distribute guesses differently.** Figure 6 reveals that the overall structure of the models’ predictions diverges from that of humans. The distribution of guesses by open source models (Qwen and Llama) are less similar to human response distributions compared to those of proprietary models (GPT and Gemini). All models perform better than a baseline which predicts a uniform distribution of guesses.

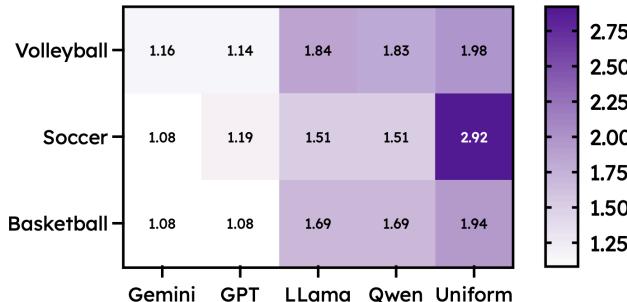
Generally, humans exhibit higher-entropy, cue-driven distributions that place probability across multiple plausible regions

(as reflected in their textual explanations). Models, by contrast, often collapse mass onto narrower regions, yielding lower entropy and, correspondingly, higher Wasserstein distance from humans. Importantly, entropy also helps understand why Wasserstein is high: (i) when model entropy  $\ll$  human entropy, divergence stems from under-dispersion (mass collapse); (ii) when entropies are comparable yet Wasserstein remains high, divergence reflects misplaced mass (systematic bias, e.g., center or near-player), not just spread. Taken together, entropy contextualizes whether distributional mismatch is due to how much probability is spread versus where it is placed, clarifying that models follow strategies distinct from humans rather than behaving as merely “noisier” variants.

**Richer prompting does not lead to consistent improvements.** Figure 4 shows that Cue-Directed prompting (explicitly instructing models to attend to player gaze and orientation) yields some improvement over Base prompting in some cases. However, these gains are inconsistent and don’t close the large gap with human performance. Interestingly, performance sometimes degrades under Chain-of-Thought prompting compared to both Base and Cue-Directed prompting (GPT in soccer and basketball). Moreover, there are no clear overarching patterns of CoT effectiveness across models: while Gemini performs best when prompted using Chain-of-Thought in soccer, the model performs worst using the same prompting in basketball. These results suggest that the models have fundamental limitations in social understanding, in that they fail to use the relevant information even when it’s explicitly pointed out to them.

### Qualitative Performance

We analyze the semantic content of model reasoning through embedding-based similarity comparison (we omitted the reasoning step in Chain-of-Thought prompting in Qwen and Llama from this analysis due to compute constraints). For each model’s reasoning text  $r_i$  from dataset  $\mathcal{D} = \{r_1, r_2, \dots, r_n\}$ , we generate high-dimensional embed-



**Figure 6: Wasserstein Distance.** Calculated between model and human predictions across sports. Lower values indicate distributions closer to human guesses.

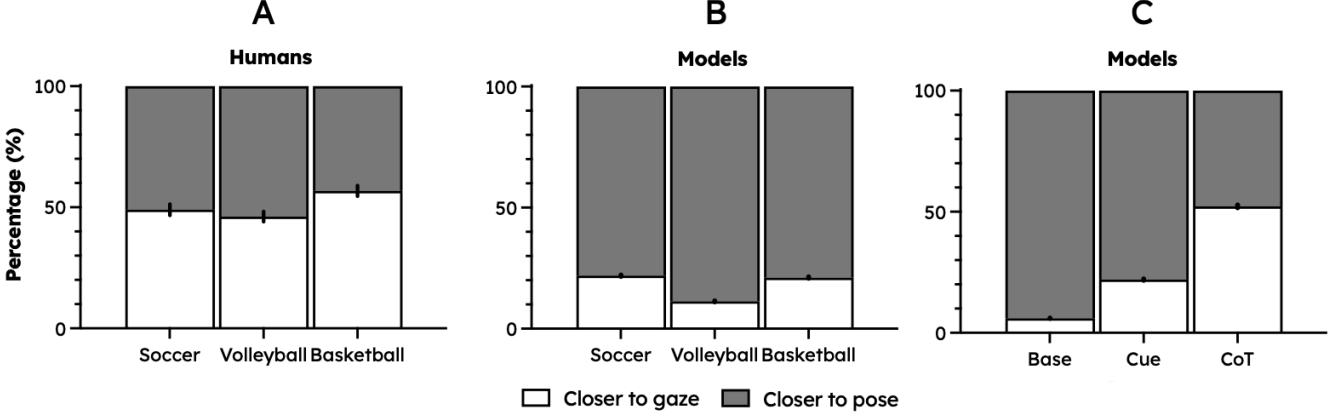


Figure 7: **Embedding analysis of reasoning similarity.** Each bar shows the proportion of model rationales whose sentence embeddings are closer to pose-like versus gaze-like reasoning templates, separated by sports and prompting strategies. Error bars indicate 95% bootstrapped confidence interval. Higher percentages toward one category indicate stronger reliance on that cue type in the model’s explanations. Panel A shows results for humans separated by sport, and Panel B results for models separated by sport, and Panel C results for models separated by prompting strategy.

dings  $\mathbf{e}_i = \text{Embed}(r_i)$  using Google’s Gemini embedding model (models/embedding-001). We then compare these reasoning embeddings against two predefined template sets: pose templates  $\mathcal{P} = \{p_1, p_2, \dots, p|\mathcal{P}|\}$  and gaze templates  $\mathcal{G} = \{g_1, g_2, \dots, g|\mathcal{G}|\}$ , where each template represents pro-

totypical reasoning patterns for pose-based or gaze-based ball location prediction (see Appendix for templates). The similarity between reasoning  $r_i$  and template set  $\mathcal{T} \in \{\mathcal{P}, \mathcal{G}\}$  is computed as  $s_i^{\mathcal{T}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \cos(\mathbf{e}_i, \mathbf{e}_t)$ , where  $\cos(\cdot, \cdot)$  denotes cosine similarity. Each reasoning is classified as pose-aligned if  $s_i^{\mathcal{P}} > s_i^{\mathcal{G}}$ , otherwise as gaze-aligned.

Table 3: **Center ratio ( $R$ ) and normalized entropy ( $\hat{H}$ )** for each model across sports, with aggregate values across all sports.  $R > 1$  indicates a center bias (meaning, the ball is predicted to be closer to the center than it actually is). Higher  $\hat{H}$  indicates broader distribution of predictions.

Sport	Model	Center Ratio $R$	Norm. Entropy $\hat{H}$
Soccer	Gemini	0.989	0.763
	GPT	0.732	0.792
	LLaMA	1.150	0.607
	Qwen	1.131	0.611
	Human	1.164	0.817
Volleyball	Gemini	1.697	0.721
	GPT	1.487	0.710
	LLaMA	0.945	0.420
	Qwen	0.953	0.422
	Human	1.602	0.768
Basketball	Gemini	0.801	0.736
	GPT	0.883	0.737
	LLaMA	0.510	0.515
	Qwen	0.518	0.517
	Human	1.093	0.801
Aggregate	Gemini	1.378	0.795
	GPT	1.356	0.808
	LLaMA	1.262	0.698
	Qwen	1.257	0.702
	Human	1.628	0.855

**Models attend more to pose than gaze in textual reasoning.** Models disproportionately rely on pose cues relative to gaze across all sports (Figure 7b). This imbalance highlights a preference for coarse, body-level orientation signals rather than fine-grained gaze information. While humans exploit both gaze and pose relatively evenly across all sports (Figure 7a), especially in less structured cues (Base and Cue-Directed Prompts), models default to pose, which may partly explain their systematic under-performance in sports, such as volleyball where the low size of the players make the pose cues harder to extract.

**CoT prompting leads to more attention on gaze, but doesn’t improve accuracy.** Chain-of-thought prompting shifts model behavior. Compared to direct prompting, CoT outputs refer to gaze cues more frequently (Figure 7c). This suggests that explicit reasoning steps help models distribute attention across multiple social cues rather than disproportionately relying on pose alone. However, this attention difference only affected the textual reasoning. When the model guesses are evaluated on accuracy and distributional similarity to humans, CoT does not yield consistent gains, and in some cases, it even degrades performance. That is, prompting helps models describe a more human-like reasoning process – one that mentions both pose and gaze cues – but without leading to better predictions of where the ball is.

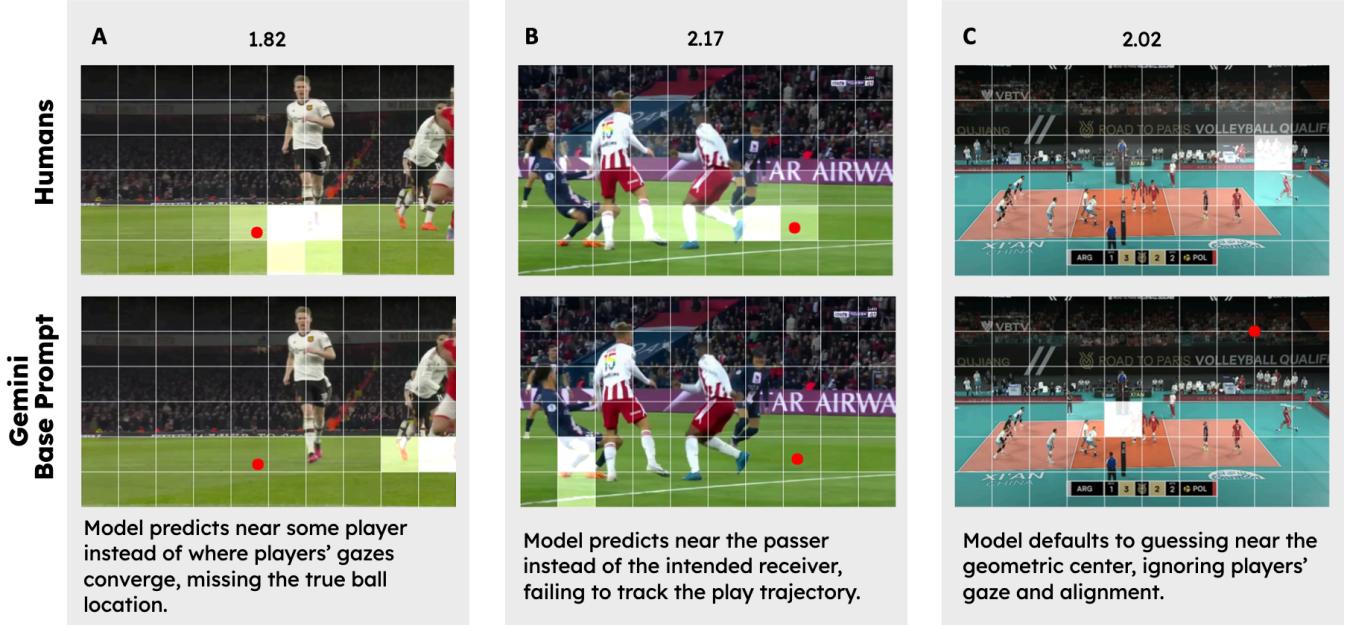


Figure 8: **Qualitative failure modes** demonstrated using Gemini using Base Prompt and human predictions. Red dots mark ground-truth ball locations, white squares mark model-predicted cells, and shaded heatmaps show prediction density. Wasserstein distances between the model and human for each example are shown in the center above each panel.

**Repeated failure modes.** Figure 8 illustrates common failure modes, using Gemini under Base Prompt as an example:

1. **Neglect of gaze.** Models fail to incorporate gaze cues into their predictions (as evidenced earlier by the lack of gaze reasoning in the embedding analysis), even when such cues provide strong evidence of ball location. For example, in Figure 8a, the model places its guess near the players' feet off to the right rather than recognizing the central player's gaze which the human guesses seem to attend to.
2. **Role confusion.** Models frequently misidentify which player has possession or is about to act. Instead of reasoning about roles within the play, they often resort to simply guessing near a player, as reflected in the proximity to players noted in our metrics. And as in Figure 8b, when the player identified is incorrect the low entropy further exacerbates the impact on the accuracy.
3. **Default-to-center heuristic.** Figure 8c provides a clear example of center bias: the model places its prediction directly in the middle of the image, at the net. This is an unlikely location for the ball, since play would have already terminated if the volleyball had struck the net. The prediction reflects a tendency to default to the geometric center rather than incorporating contextual cues from the scene.

## Future Directions

Our results highlight fundamental limitations of current vision-language models in socially-grounded object inference. Prompting strategies are insufficient to close the performance

gap, pointing to deeper issues in how models perceive and reason about social cues such as pose, and gaze.

It is worth noting that the goal of this benchmark is not simply to train a neural network to localize balls in images, but to probe whether models possess the social understanding necessary to do so. Success requires reasoning under uncertainty—inferring hidden states from body orientation, gaze, and contextual cues that humans naturally exploit. The challenge therefore lies less in visual detection itself and more in whether current architectures and training objectives can capture these social priors and deploy them in a one-shot setting.

The aim of this work is to expose systematic blind spots in current models and to motivate progress on this problem. By releasing both our dataset and evaluation code, we hope to establish a foundation that the community can build upon in developing models with more human-like capacities for socially grounded inference.

Several future directions emerge from this work:

1. **Architectural innovations.** New model designs explicitly tailored to capture social reasoning in visual scenes may be required, as current architectures appear biased toward spatial heuristics rather than agent dynamics.
2. **Broader domains.** Expanding evaluation across additional sports and domains would test the robustness and generality of these findings.
3. **Controlled environments.** Moving beyond static images, 3D interactive settings (e.g., Unity simulations) would allow systematic manipulation of gaze, pose, and player configurations to probe how models adapt to changes in

pose and gaze while holding other aspects of the scene constant.

4. **Inferring player positions.** Testing how models predict (inpainted) player positions given information about the ball and other players could provide more insights into the models’ social reasoning capabilities.

## Conclusion

We introduced SPOT THE BALL— a new benchmark for evaluating how vision-language models and humans infer hidden objects in socially grounded sports settings. Across soccer, volleyball, and basketball humans consistently outperform models. Models exhibit systematic biases in their guesses, particularly a collapse toward central and near player regions.

Progress may require integrating perceptual priors, temporal reasoning, and architectures explicitly designed to capture agentive and relational dynamics. By identifying both where models succeed and where they fail, our work lays the groundwork for future approaches to building AI systems capable of more robust and human-like visual social inference.

## Acknowledgments

We thank Jiajun Wu for helpful feedback and discussion. TG was supported by research grants from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and Cooperative AI.

## References

- Abdelnabi, S.; Gomaa, A.; Sivaprasad, S.; Schönherr, L.; and Fritz, M. 2024. Cooperation, competition, and maliceousness: Llm-stakeholders interactive negotiation. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomeczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 83548–83599. Curran Associates, Inc.
- Anonymous. 2025. Mind the motions: Benchmarking theory-of-mind in everyday body language. In *Submitted to ACL Rolling Review - May 2025*. under review.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-vl technical report.
- Baron-Cohen, S.; Leslie, A. M.; and Frith, U. 1985. Does the autistic child have a "theory of mind"? *Cognition* 21(1):37–46.
- Bradski, G. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
- Byom, L. J., and Mutlu, B. 2013. Theory of mind: mechanisms, methods, and new directions. *Frontiers in Human Neuroscience* 7:413.
- Chan, C.; Jiayang, C.; Yim, Y.; Deng, Z.; Fan, W.; Li, H.; Liu, X.; Zhang, H.; Wang, W.; and Song, Y. 2024. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding.
- Chandrasegaran, K.; Gupta, A.; Hadzic, L. M.; Kota, T.; He, J.; Eyzaguirre, C.; Durante, Z.; Li, M.; Wu, J.; and Fei-Fei, L. 2024. Hourvideo: 1-hour video-language understanding.
- Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Chan, C.-M.; Yu, H.; Lu, Y.; Hung, Y.-H.; Qian, C.; Qin, Y.; Cong, X.; Xie, R.; Liu, Z.; Sun, M.; and Zhou, J. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors.
- Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Driess, D.; Florence, P.; Sadigh, D.; Guibas, L.; and Xia, F. 2024a. Spatialvilm: Endowing vision-language models with spatial reasoning capabilities.
- Chen, K.; Lian, Z.; Sun, H.; Liu, R.; Yi, J.; Liu, B.; and Tao, J. 2024b. Can deception detection go deeper? dataset, evaluation, and benchmark for deception reasoning.
- Chen, Y.; Wang, H.; Yan, S.; Liu, S.; Li, Y.; Zhao, Y.; and Xiao, Y. 2024c. Emotionqueen: A benchmark for evaluating empathy of large language models.
- Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; and Huang, M. 2024d. Tombench: Benchmarking theory of mind in large language models.
- Choi, M.; Pei, J.; Kumar, S.; Shu, C.; and Jurgens, D. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11370–11403. Singapore: Association for Computational Linguistics.
- de Gelder, B.; de Borst, A.; and Watson, R. 2015. The perception of emotion in body expressions. *WIREs Cognitive Science* 6(2):149–158.
- Dennett, D. C. 1989. *The Intentional Stance*. The MIT Press.
- Fan, X.; Zhou, X.; Jin, C.; Nottingham, K.; Zhu, H.; and Sap, M. 2025. Somi-tom: Evaluating multi-perspective theory of mind in embodied social interactions.
- Fletcher-Watson, S.; McConnell, F.; Manola, E.; and McConachie, H. 2014. Interventions based on the theory of mind cognitive model for autism spectrum disorder (asd). *Cochrane Database of Systematic Reviews* 2014(3):CD008785.
- Foss, A.; Evans, C.; Mitts, S.; Sinha, K.; Rizvi, A.; and Kao, J. T. 2025. Causalvqa: A physically grounded causal reasoning benchmark for video models.
- Gandhi, K.; Fränken, J.-P.; Gerstenberg, T.; and Goodman, N. 2023. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Gassmann, L.; Campbell, J.; and Edwards, M. 2024. Influence reasoning capabilities of large language models in social environments. *Proceedings of the AAAI Symposium Series* 4(1):40–47.
- Heider, F., and Simmel, M. 1944. An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57(2):243–259. Publisher: University of Illinois Press.
- Hu, Z.; Ren, Y.; Li, J.; and Yin, Y. 2024. Viva: A benchmark for vision-grounded decision-making with human values.
- Huang, J.-t.; Lam, M. H.; Li, E. J.; Ren, S.; Wang, W.; Jiao, W.; Tu, Z.; and Lyu, M. R. 2024. Apathetic or empathetic? evaluating llms' emotional alignments with humans. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomeczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 97053–97087. Curran Associates, Inc.
- Ji, J.; Chen, W.; Wang, K.; Hong, D.; Fang, S.; Chen, B.; Zhou, J.; Dai, J.; Han, S.; Guo, Y.; and Yang, Y. 2025a. Mitigating deceptive alignment via self-monitoring.
- Ji, J.; Chen, Y.; Jin, M.; Xu, W.; Hua, W.; and Zhang, Y. 2025b. Moralbench: Moral evaluation of llms.
- Jiao, J.; Afrooghi, S.; Murali, A.; Chen, K.; Atkinson, D.; and Dhurandhar, A. 2025. Llm ethics benchmark: A three-dimensional assessment system for evaluating moral reasoning in large language models.
- Jin, C.; Wu, Y.; Cao, J.; Xiang, J.; Kuo, Y.-L.; Hu, Z.; Ullman, T.; Torralba, A.; Tenenbaum, J. B.; and Shu, T. 2024. Mmtom-qa: Multimodal theory of mind question answering.
- Jocher, G.; Chaurasia, A.; and Qiu, J. 2023. Ultralytics yolov8.

- Jung, D.; Lee, S.; Moon, H.; Park, C.; and Lim, H. 2025. Flex: A benchmark for evaluating robustness of fairness in large language models.
- Kleinke, C. L. 1986. Gaze and eye contact: A research review. *Psychological Bulletin* 100(1):78–100.
- Kong, F.; Zu, W.; Chen, X.; Yang, Y.; Zhu, S.-C.; and Feng, X. 2025. Siv-bench: A video benchmark for social interaction understanding and reasoning.
- Kosti, R.; Alvarez, J. M.; Recasens, A.; and Lapedriza, A. 2017. Emotic: Emotions in context dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2309–2317.
- Krishna, S.; Zou, A.; Gupta, R.; Jones, E. K.; Winter, N.; Hendrycks, D.; Kolter, J. Z.; Fredrikson, M.; and Matsoukas, S. 2025. D-rex: A benchmark for detecting deceptive reasoning in large language models.
- Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2018. Visual social relationship recognition.
- Li, C.; Zhang, C.; Zhou, H.; Collier, N.; Korhonen, A.; and Vulić, I. 2024. Topviewrs: Vision-language models as top-view spatial reasoners.
- Li, Y.; Veerabadran, V.; Iuzzolino, M. L.; Roads, B. D.; Celikyilmaz, A.; and Ridgeway, K. 2025. Egotom: Benchmarking theory of mind reasoning from egocentric videos.
- Li, J.; Niu, L.; and Zhang, L. 2022. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering.
- Lin, X.; Liu, Z.; Yang, Z.; Li, G.; Qiu, R.; Wang, S.; Liu, H.; Li, H.; Keswani, S.; Pardeshi, V.; Zhao, H.; Fan, W.; and Tong, H. 2025a. Moralise: A structured benchmark for moral alignment in visual language models.
- Lin, Z.; Xu, Z.; Song, X.; Wan, Y.; Yao, X.; Lin, T.-H.; Song, S.; Subbaraman, P.; Zhou, B.; Chang, K.-W.; and Sun, Y. 2025b. V-ALPHASOCIAL: Benchmark and self-reflective chain-of-thought generation for visual social commonsense reasoning. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 19025–19047. Vienna, Austria: Association for Computational Linguistics.
- Liu, S., and Almeida, M. 2023. Knowing before doing: Review and mega-analysis of action understanding in prereaching infants. *Psychological Bulletin* 149(5-6):294–310.
- Liu, Z.; Gao, K.; Liang, S.; Xiao, B.; Qiao, L.; Ma, L.; and Jiang, T. 2025. Beyond the visible: Benchmarking occlusion perception in multimodal large language models.
- Luo, Y.; Ye, J.; Adams, R. B.; Li, J.; Newman, M. G.; and Wang, J. Z. 2019. Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International Journal of Computer Vision* 128(1).
- Ma, X.; Gao, L.; and Xu, Q. 2024. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind.
- Marcuzzo, M.; Zangari, A.; Albarelli, A.; Camacho-Collados, J.; and Pilehvar, M. T. 2025. Morables: A benchmark for assessing abstract moral reasoning in llms with fables.
- Mathur, L.; Qian, M.; Liang, P. P.; and Morency, L.-P. 2025. Social genome: Grounded social reasoning abilities of multimodal models.
- Mathur, L.; Liang, P. P.; and Morency, L.-P. 2024. Advancing social intelligence in AI agents: Technical challenges and open questions. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20541–20560. Miami, Florida, USA: Association for Computational Linguistics.
- Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2019. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10(1):18–31.
- Niu, L.; Li, J.; Yu, X.; Wang, S.; Feng, R.; Wu, B.; Wei, P.; Wang, Y.; and Fan, L. 2025. R<sup>3</sup>-vqa: “read the room” by video social reasoning.
- Pothiraj, A.; Stengel-Eskin, E.; Cho, J.; and Bansal, M. 2025. Capture: Evaluating spatial reasoning in vision language models via occluded object counting.
- Puig, X.; Undersander, E.; Szot, A.; Cote, M. D.; Yang, T.-Y.; Partsey, R.; Desai, R.; Clegg, A. W.; Hlavac, M.; Min, S. Y.; Vondruš, V.; Gervet, T.; Berges, V.-P.; Turner, J. M.; Maksymets, O.; Kira, Z.; Kalakrishnan, M.; Malik, J.; Chaplot, D. S.; Jain, U.; Batra, D.; Rai, A.; and Mottaghi, R. 2023. Habitat 3.0: A co-habitat for humans, avatars and robots.
- Pătrăucean, V.; Smaira, L.; Gupta, A.; Continente, A. R.; Markeeva, L.; Banarse, D.; Koppula, S.; Heyward, J.; Malinowski, M.; Yang, Y.; Doersch, C.; Matejovicova, T.; Sulsky, Y.; Miech, A.; Frechette, A.; Klimczak, H.; Koster, R.; Zhang, J.; Winkler, S.; Aytar, Y.; Osindero, S.; Damen, D.; Zisserman, A.; and Carreira, J. 2023. Perception test: A diagnostic benchmark for multimodal video models.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision.
- Ren, Z.; Ortega, J.; Wang, Y.; Chen, Z.; Guo, Y.; Yu, S. X.; and Whitney, D. 2023. Veatic: Video-based emotion and affect tracking in context dataset.
- Sap, M.; Rashkin, H.; Chen, D.; LeBras, R.; and Choi, Y. 2019. Socialiqqa: Commonsense reasoning about social interactions.
- Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. M. 2023. Evaluating the moral beliefs encoded in llms.
- Shi, H.; Ye, S.; Fang, X.; Jin, C.; Isik, L.; Kuo, Y.-L.; and Shu, T. 2025. Muma-tom: Multi-modal multi-agent theory of mind.
- Sodian, B. 2011. Theory of Mind in Infancy. *Child Development Perspectives* 5(1):39–43.

- Tager-Flusberg, H. 2007. Evaluating the theory-of-mind hypothesis of autism. *Current Directions in Psychological Science* 16(6):311–315.
- Telea, A. 2004. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools* 9(1):23–34.
- Todorov, A.; Olivola, C. Y.; Dotsch, R.; and Mende-Siedlecki, P. 2015. Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology* 66(1):519–545.
- Trager, J.; Vargas, F.; Alves, D.; Guida, M.; Ngueajio, M. K.; del Arco, F. P.; Daryanai, Y.; Karimi-Malekabadi, F.; and Agrawal, A. 2025. Mftcxplain: A multilingual benchmark dataset for evaluating the moral reasoning of llms through hate speech multi-hop explanations.
- Tremoulet, P. D., and Feldman, J. 2000. Perception of Animacy from the Motion of a Single Object. *Perception* 29(8):943–951.
- Ullman, T. 2023. Large language models fail on trivial alterations to theory-of-mind tasks.
- Vicol, P.; Tapaswi, M.; Castrejon, L.; and Fidler, S. 2018. Moviegraphs: Towards understanding human-centric situations from videos.
- Wang, Z.; Davis, I.; and Jara-Ettinger, J. 2025. Modeling Other Minds: A Computational Account of Social Cognition and Its Development. *Annual Review of Developmental Psychology*.
- Wellman, H. M. 2014. *Making minds: How theory of mind develops*. Making minds: How theory of mind develops. New York, NY, US: Oxford University Press.
- Wilf, A.; Mathur, L.; Mathew, S.; Ko, C.; Kebe, Y.; Liang, P. P.; and Morency, L.-P. 2023. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. *GitHub repository*.
- Woo, B. M.; Liu, S.; and Spelke, E. S. 2024. Infants rationally infer the goals of other people’s reaches in the absence of first-person experience with reaching actions. *Developmental Science* 27(3):e13453.
- Wu, T.-L.; Dou, Z.-Y.; Hu, Q.; Hou, Y.; Chandra, N. R.; Freedman, M.; Weischedel, R. M.; and Peng, N. 2023a. Acquired: A dataset for answering counterfactual questions in real-life videos.
- Wu, Y.; He, Y.; Jia, Y.; Mihalcea, R.; Chen, Y.; and Deng, N. 2023b. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10691–10706. Singapore: Association for Computational Linguistics.
- Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa:next phase of question-answering to explaining temporal actions.
- Xu, Z.; Wang, Y.; Huang, Y.; Ye, J.; Zhuang, H.; Song, Z.; Gao, L.; Wang, C.; Chen, Z.; Zhou, Y.; Li, S.; Pan, W.; Zhao, Y.; Zhao, J.; Zhang, X.; and Chen, X. 2025. Socialmaze: A benchmark for evaluating social reasoning in large language models.
- Yao, R.; Zhang, B.; Huang, J.; Long, X.; Zhang, Y.; Zou, T.; Wu, Y.; Su, S.; Xu, Y.; Zeng, W.; Yang, Z.; Li, G.; Zhang, S.; Li, Z.; Chen, Y.; Xiong, S.; Xu, P.; Zhang, J.; Zhou, B.; Clifton, D.; and Gool, L. V. 2025. Lens: Multi-level evaluation of multimodal reasoning with large language models.
- Yiu, E.; Qraitem, M.; Wong, C.; Majhi, A. N.; Bai, Y.; Ginosar, S.; Gopnik, A.; and Saenko, K. 2024. Kiva: Kid-inspired visual analogies for testing large multimodal models. *ArXiv* abs/2407.17773.
- Zadeh, A.; Chan, M.; Liang, P. P.; Tong, E.; and Morency, L.-P. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8799–8809.
- Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning.
- Zhong, Y.; Xiao, J.; Ji, W.; Li, Y.; Deng, W.; and Chua, T.-S. 2022. Video question answering: Datasets, algorithms and challenges.

## 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.>
```

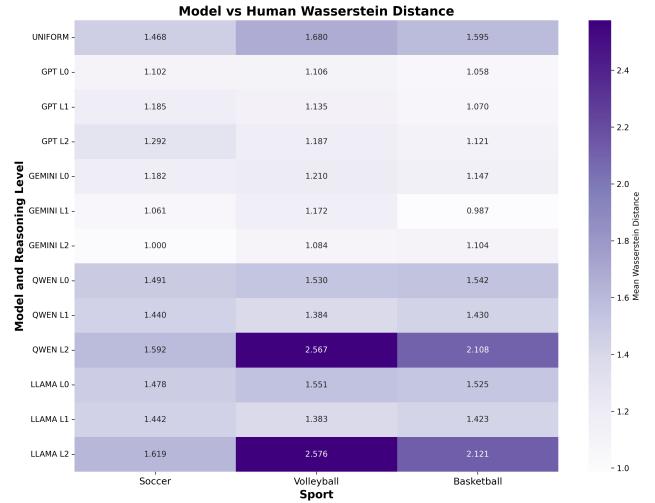


Figure 9: Wasserstein Distances with all the levels and models

## Evaluation Metrics Implementation

### 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.

### 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  $\tau_{\text{near}}$  represents the distance threshold as a fraction of the image diagonal, and  $\theta$  denotes the minimum overlap fraction between a predicted grid cell and any player box.

To guide selection, we define a *balanced objective* that equally weights NR and OR:

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

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



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

revealed that most predicted or true ball regions already overlap with player boxes (median normalized distance = 0; median overlap fraction  $\approx 0.56$ ), 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.

### 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.

### Code and Dataset

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

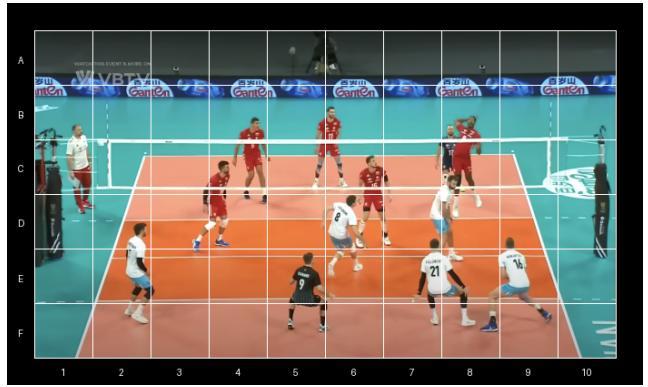


Figure 11: An example of the volleyball image shown to the models. The ball is masked, and a grid is overlaid; the responses in the following tables are generated based on this image.

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

<b>Reasoning</b>	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
<b>Grid cell guess</b>	Cell: C9

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

<b>Reasoning</b>	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.
<b>Grid cell guess</b>	Cell: D4

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

<b>Reasoning</b>	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.
<b>Grid cell guess</b>	Cell: C7

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

<b>Reasoning</b>	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.
<b>Grid cell guess</b>	Cell: D6

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

<b>Where are the players located?</b>	<p>The players are located at the following grid positions:</p> <ul style="list-style-type: none"> <li>- Player in white jersey near the left side: E2</li> <li>- Player in white jersey near the center: E6</li> <li>- Player in white jersey near the right side: E8</li> <li>- Player in black jersey near the center: E5</li> <li>- Player in black jersey near the right side: E7</li> <li>- Player in red jersey near the left side: C3</li> <li>- Player in red jersey near the center-left: C5</li> <li>- Player in red jersey near the center: C6</li> <li>- Player in red jersey near the center-right: C7</li> <li>- Player in red jersey near the right side: C8</li> </ul>
<b>Where are the players looking?</b>	<p>The players are looking in the following directions:</p> <ul style="list-style-type: none"> <li>- The player in white at E2 is looking towards the net, slightly to the right.</li> <li>- The player in white at E6 is looking upwards and slightly to the right.</li> <li>- The player in white at E7 is looking upwards and slightly to the left.</li> <li>- The player in white at E9 is looking upwards and slightly to the left.</li> <li>- The player in black at E5 is looking upwards and slightly to the right.</li> <li>- The player in red at C4 is looking upwards and slightly to the right.</li> <li>- The player in red at C6 is looking upwards and slightly to the left.</li> <li>- The player in red at C7 is looking upwards and slightly to the left.</li> <li>- The player in red at C8 is looking upwards and slightly to the left.</li> <li>- The player in red at B9 is looking upwards and slightly to the left.</li> </ul> <p>Overall, most players are looking upwards, likely tracking the volleyball in play.</p>
<b>How are the players positioned?</b>	<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"> <li>- One player at B4</li> <li>- One player at B6</li> <li>- One player at B8</li> <li>- One player at C3</li> <li>- One player at C5</li> <li>- One player at C7</li> </ul> <p><i>Team in white and blue (bottom side of the net):</i></p> <ul style="list-style-type: none"> <li>- One player at E2</li> <li>- One player at E6</li> <li>- One player at E7</li> <li>- One player at E8</li> <li>- One player at F5</li> <li>- One player at F6</li> </ul>
<b>Reasoning</b>	<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>
<b>Grid cell guess</b>	E5

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

<b>Where are the players located?</b>	<p>Here are the locations of the players in the image:</p> <ul style="list-style-type: none"> <li>* **1, C**: Player in white and red.</li> <li>* **2, D**: Player in white and blue.</li> <li>* **3, D**: Player in red.</li> <li>* **4, C**: Player in red.</li> <li>* **5, E**: Player in dark blue and white.</li> <li>* **5, B**: Player in red.</li> <li>* **6, D**: Player in white and blue.</li> <li>* **6, C**: Player in red.</li> <li>* **7, E**: Player numbered 21 in white and blue.</li> <li>* **8, C**: Player in red.</li> <li>* **8, D**: Player in white and blue.</li> <li>* **9, E**: Player numbered 16 in white and blue.</li> <li>* **9, B**: Player in red.</li> <li>* **10, C**: Player in white and blue.</li> </ul>
<b>Where are the players looking?</b>	<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>
<b>How are the players positioned?</b>	<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"> <li>* Player 1: A-1/2</li> <li>* Player 2: D-2/3</li> <li>* Player 3: C-4</li> <li>* Player 4: D-6</li> </ul> <p>**Team 2 (Right Side):**</p> <ul style="list-style-type: none"> <li>* Player 1: B-9</li> <li>* Player 2: D-8</li> <li>* Player 3: E-7</li> <li>* Player 4: F-9</li> </ul> <p>**Mid-Court Players:**</p> <ul style="list-style-type: none"> <li>* Player 5: E-5</li> <li>* Player 6: B-5</li> <li>* Player 7: C-6</li> <li>* Player 8: B-8</li> <li>* Player 9: C-4</li> </ul>
<b>Reasoning</b>	<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>
<b>Grid cell guess</b>	Cell: E4