

SPOT THE BALL: A Benchmark for Visual Social Inference

Neha Balamurugan Sarah Wu Adam Chun
Gabe Gaw Cristobal Eyzaguirre Tobias Gerstenberg
Stanford University
nbalamur@stanford.edu 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 capacity 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.

1. Introduction

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 their behavior [5, 10, 41]. ToM is fundamental to everyday human interaction [3, 12, 37], and increasingly critical for AI systems deployed in social contexts: a robot nurse that misinterprets a patient’s gesture, or an autonomous car that fails to anticipate a pedestrian’s intention, may behave

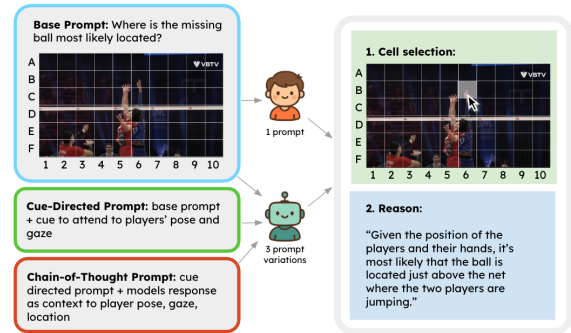


Figure 1. Overview of the SPOT THE BALL task. Given an image with the ball removed, humans and models infer the likely location by reasoning about player pose and gaze. Models are prompted under three conditions, whereas humans receive only the base prompt.

unsafely.

While LLMs perform some types of social reasoning based on text relatively well [13], such behavior may rely primarily on linguistic pattern matching rather than perceptual grounding. Further, human social reasoning rarely occurs through text alone—we continuously integrate visual cues such as gaze direction, facial expressions, body language, and spatial relationships alongside (or in the absence of) verbal information. To evaluate whether AI systems can perform similar visually-grounded social inference, we study VLMs. VLMs process the visual information available to human observers and are increasingly deployed in embodied, safety-critical contexts where robust interpretation of visual social cues is essential.

Existing benchmarks for social reasoning either evaluate models on fully visible scenes [28, 46] or emphasize inanimate objects [24], failing to capture how humans reason about others under partial information. We introduce SPOT THE BALL, inspired by a classic newspaper puzzle, to evaluate whether VLMs can infer hidden objects from visual social cues in real-world images. Ball sports provide an ideal testbed as players’ gaze, posture, and positioning

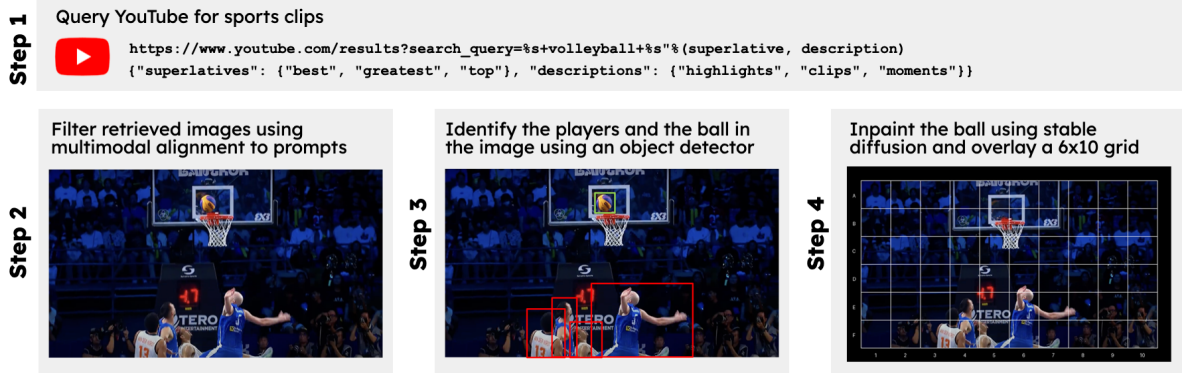


Figure 2. **Pipeline for constructing the SPOT THE BALL dataset.** We retrieve and filter sports footage from YouTube by alignment to the prompts, detect players and balls using an object detector, and inpaint the ball region with stable diffusion before overlaying a 6x10 grid for location annotation.

are causally coupled to the ball, yielding interpretable social signals. We use static images to isolate social reasoning from motion dynamics.

We contribute: (1) 150 curated sports images with human baselines, (2) systematic evaluation of four leading VLMs under multiple prompting strategies, and (3) a scalable pipeline for generating ball sports reasoning tasks, with which we have produced 3,000 additional soccer images for training and analysis. We provide the first structured evaluation of VLMs’ ability to leverage social cues for inferring hidden objects in real-world scenes. Our results reveal large, consistent gaps between human and model performance.

2. 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 [36, 40, 41]. Adults readily infer rich social states from minimal visual cues—moving geometric shapes [14, 39], gaze [18], facial expressions [38], and body language [9]. As AI systems are increasingly deployed in social contexts, they must exhibit similar reasoning capabilities for seamless human interaction [27].

Social reasoning benchmarks for AI. Most social reasoning benchmarks for AI operate in text-only settings, evaluating ToM [13, 25, 44], empathy [8, 15], moral reasoning [26, 34], deception detection [7, 20], and negotiation [1, 33]. While informative, these benchmarks cannot assess whether models extract social cues from visual perception—the primary modality through which humans perform social inference.

Video-based benchmarks extend social reasoning evaluation to dynamic visual contexts, including theory-of-mind

in human interactions [16, 23, 35] and synthetic simulations [11], causal and counterfactual video QA [43, 45], and social interaction understanding [42, 47]. Static image benchmarks are less common but include Visual Commonsense Reasoning [48], facial expression understanding [19, 29], and social relationship recognition [22]. However, existing visual benchmarks either present fully observable scenes [28, 46] or focus on physical occlusion without social cues [24, 30]. No prior work evaluates whether models can infer hidden information purely from visual social cues.

Our contribution. We introduce a zero-shot, image-based benchmark requiring models to infer a hidden object by integrating pose, gaze, and orientation cues from human agents. Unlike prior work, our task isolates visually-grounded social reasoning in naturalistic scenes with partial information, reflecting how humans actually perform social inference in everyday contexts.

3. 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.

3.1. Evaluation Set

We curated 150 images from publicly available soccer, basketball, and volleyball footage on YouTube (see Figure 2). Frames were procedurally selected to maximize contextual informativeness (non-occluded players, clear ball presence, spatial distribution) and manually verified for quality. The ball was removed via inpainting after recording its ground

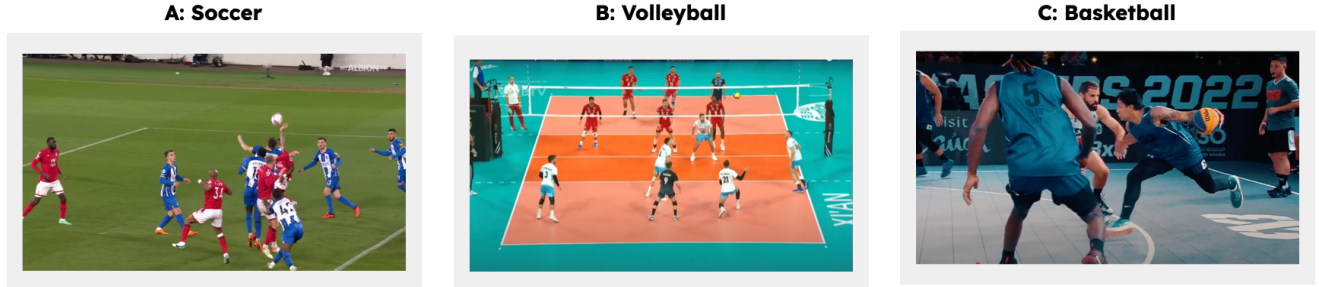


Figure 3. **Player density and coverage across sports in the SPOT THE BALL dataset.** In Soccer frames (A) player count and area are intermediate compared to the other sports. Volleyball frames (B) feature the most players but each occupies a smaller visual area, providing weaker pose and gaze cues.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

truth location. This evaluation set is used in our experiments for comparing models against human reasoning.

3.2. Scalable Pipeline

To scale beyond the evaluation set, we developed a modular pipeline to generate realistic inpainted sports scenes:

1. **Video retrieval.** Broadcast footage was retrieved from YouTube using sport-specific queries with action-focused keywords (“best”, “highlights”, “moments”). Videos were decoded with OpenCV [4] and sampled at ~ 1 FPS.
2. **Frame filtering.** Each frame was scored with CLIP [32] against prompts like “picture of volleyball players in action with ball”. Only frames exceeding a similarity threshold were retained.
3. **Ball and player detection.** Frames were passed through YOLOv8 [17] to detect balls and players. We filtered by confidence and spatial plausibility, requiring exactly one ball per frame, non-overlapping with and proximal to players. This eliminated spurious detections while preserving contextual cues.
4. **Ball inpainting.** Ball regions were removed and filled using Stable Diffusion inpainting, which generates realistic textures and lighting while avoiding visible artifacts. Player masks ensured body posture and gaze cues remained intact. Images were manually checked to remove any remaining ball shadows or artifacts.

Each image is overlaid with a 6×10 alphanumeric grid (rows A–F, columns 1–10). Ground-truth labels correspond to coordinates covering the original ball location (e.g., [A5] for single-cell coverage, or [A5, A6, B5, B6] for multi-cell overlap).

This pipeline generated 3,000 additional soccer images beyond the curated evaluation set. The modular design allows extensions to other ball sports or difficulty controls (e.g., varying player density or occlusion severity).

4. 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, 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.

4.1. Models

We evaluate four multi-modal instruction-following models: Gemini-2.0-flash-001, GPT-4.1-mini, LLaMA-3.2-11B-Vision-Instruct and 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.[2]. Together, these architectures represent a spectrum from fully integrated multimodal transformers to adapter-based and hierarchical fusion strategies.

4.2. Domains

We selected soccer, volleyball, and basketball because they are ball sports that are present in high-frequency in pretrain-

ing 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).

4.3. 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 [6, 21], and that auxiliary cues can enhance localization [30]. Supplementary material contains system prompts.

4.4. Human experiment

To collect human baselines, recruited 176 participants from Prolific and compensated them for their time (\$12/hour base + up to \$1 accuracy bonus). After excluding 26 for failed attention checks, the final sample was $N = 150$ (50 per sport). The experiments were pre-registered on the [Open Science Framework](#)¹ and approved by the [REDACTED] Institutional Review Board. Participants received instructions with example images (ball visible and removed) before making three guesses per test image by clicking grid

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

¹The volleyball and basketball conditions were identical to the pre-registered soccer experiment. Model prompts were updated from pre-registration based on piloting.

cells. Each participant saw 52 images (50 test, 2 attention checks with visible balls) from one sport in randomized order. Participants who clicked outside ball-containing cells on attention checks were excluded. Completion time averaged 17.3 minutes (SD = 7.4).

4.5. Evaluation Metrics

We evaluate models across three dimensions: **task performance** (localization accuracy), **alignment with humans** (similarity to human reasoning), and **behavioral strategies** (prediction patterns).

Task performance. (1) **Accuracy:** For image i with ground-truth cells $\mathcal{G}_i$ and prediction $\hat{y}_i$:

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

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

$$d_i = \min_{g \in \mathcal{G}_i} \|c(\hat{y}_i) - c(g)\|_2. \quad (2)$$

Alignment with humans. (1) **Wasserstein Distance:** We compare model and human prediction distributions P and Q using Earth Mover’s Distance:

$$\begin{aligned} W(P, Q) &= \min_{\gamma \geq 0} \sum_{j,k} \gamma_{jk} D_{jk} \\ \text{s.t. } \sum_k \gamma_{jk} &= P_j, \quad \sum_j \gamma_{jk} = Q_k, \end{aligned} \quad (3)$$

where D_{jk} is Euclidean distance between cell centers. Lower values indicate closer alignment with human strategies.

Behavioral strategies. Let $\mathcal{B}_i = \{b_{i,m}\}$ be player bounding boxes in image i , with $p_{i,t} = c(\hat{y}_{i,t})$ the predicted cell center. Define point-to-box distance:

$$\text{dist}(x, b) = \begin{cases} 0, & x \in \text{supp}(b), \\ d(x, \text{supp}(b)), & \text{otherwise,} \end{cases} \quad (4)$$

where $d(x, \text{supp}(b))$ is Euclidean distance to the box.

(1) **Near Player Rate (NR):** Fraction of predictions within threshold τD of any player ($\tau = 0.08$, D = image diagonal):

$$NR = \frac{1}{\sum_i T_i} \sum_{i,t} \mathbb{1} \left[\min_{b \in \mathcal{B}_i} \text{dist}(p_{i,t}, b) \leq \tau D \right]. \quad (5)$$

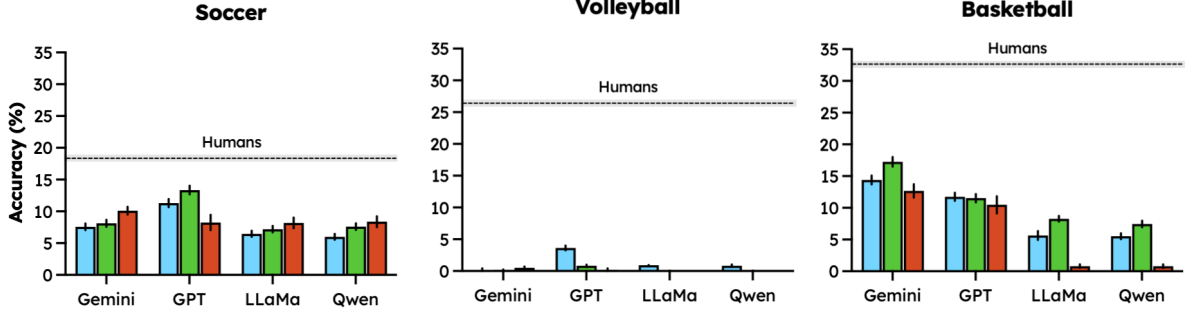


Figure 4. **Accuracy.** Model accuracy in each sport under different prompting strategies (blue = base prompt, green = cue-directed prompt, red = chain-of-thought prompt). The dashed line shows human accuracy using the base prompt in the given sport. Error bars and gray ribbon show 95% bootstrapped confidence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

(2) **Overlap Rate (OR):** Fraction of predictions whose grid cell intersects a player box by $\geq \theta$ of cell area ($\theta = 0.02$):

$$OR = \frac{1}{\sum_i T_i} \sum_{i,t} \mathbb{I} \left[\max_{b \in \mathcal{B}_i} \frac{\text{area}(c(\hat{y}_{i,t}) \cap b)}{\text{area}(c(\hat{y}_{i,t}))} \geq \theta \right]. \quad (6)$$

(3) **Center Ratio (CR):** Ratio of prediction mass to ground-truth mass in central window $\mathcal{S} = \{(r, c) : r \in \{2, 3, 4\}, c \in \{3, \dots, 7\}\}$ (3×5 region):

$$CR = \frac{\sum_{j \in \mathcal{S}} p_j}{\sum_{j \in \mathcal{S}} q_j}, \quad (7)$$

where p is the model distribution and q is the ground-truth prior. Values > 1 indicate center bias.

(4) **Entropy:** Normalized entropy measures prediction spread:

$$\hat{H}(p) = \frac{-\sum_{j=1}^{60} p_j \log p_j}{\log 60}. \quad (8)$$

Higher values indicate broader exploration; lower values indicate concentration.

5. 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.

5.1. 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 ± 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 ± 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 vol-

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
Gemini	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
	CoT	141.1 $\pm$ 72.9	150.5 $\pm$ 48.2	134.7 $\pm$ 73.4
GPT	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
	CoT	146.1 $\pm$ 68.3	155.3 $\pm$ 56.8	137.1 $\pm$ 68.8
LLaMA	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
	CoT	140.2 $\pm$ 87.1	272.6 $\pm$ 50.7	211.4 $\pm$ 82.6
Qwen	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
	CoT	139.0 $\pm$ 81.0	271.5 $\pm$ 52.9	211.0 $\pm$ 82.5

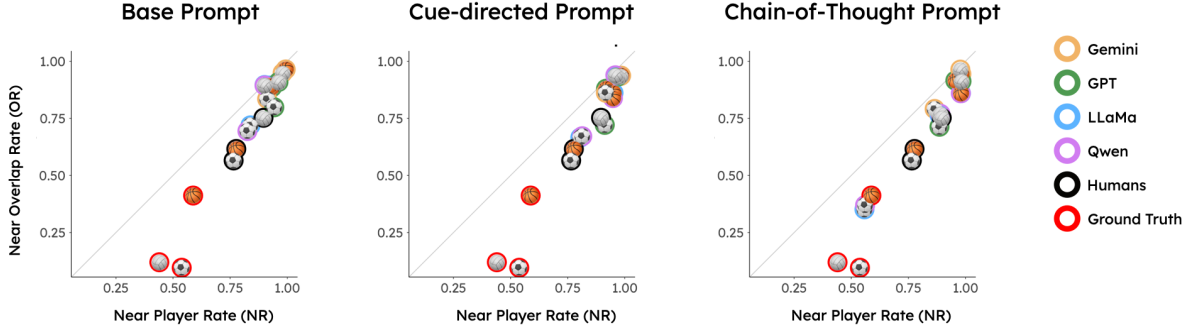


Figure 5. **Player Proximity Analysis.** Each point corresponds to a model–sport combination. The x -axis shows the fraction of guesses within a fixed distance threshold of any player (Near Player Rate), while the y -axis shows the fraction of guesses whose predicted cell overlaps a player bounding box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

leyball, 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 ($\sim 20,000$ 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 ($\sim 6,600$), 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.

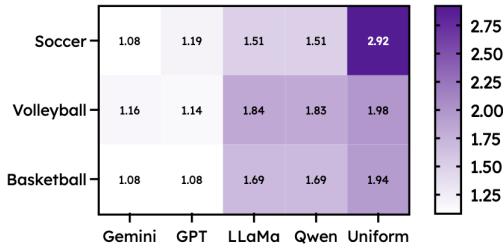


Figure 6. **Wasserstein Distance.** Calculated between model and human predictions across sports. Lower values indicate distributions closer to human guesses.

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 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 is much less than human entropy, divergence could stem from under-dispersion of guesses; (ii) when entropies are comparable yet Wasserstein distance remains high, divergence could reflect misplaced mass (e.g. systematic center or near-player bias), not just spread. Taken together, entropy contextualizes whether distributional mis-

match 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.

5.2. Qualitative Performance

We analyze the semantic content of model reasoning through embedding-based similarity comparison (excluding Chain-of-Thought reasoning from Qwen and Llama due to compute constraints). For each model’s reasoning text, we generate embeddings using Google’s Gemini embedding model (models/embedding-001) and compare them

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

against predefined pose and gaze reasoning templates (see Appendix). We compute the average cosine similarity between each reasoning and all templates in each category, then classify reasoning as pose-aligned if it is more similar to pose templates than gaze templates, and gaze-aligned otherwise.

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. While prompting helps models describe a more human-like reasoning process – one that mentions both pose and gaze cues – it doesn’t lead to better predictions of where the ball is.

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.

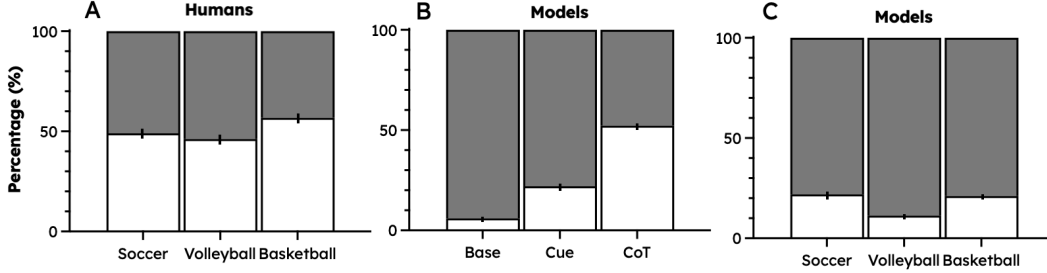


Figure 7. **Embedding analysis of reasoning similarity.** Each bar shows the proportion of model rationales whose sentence embeddings are closer to pose-like versus gaze-like reasoning templates, separated by sports and prompting strategies. Error bars indicate 95% bootstrapped confidence interval. The distribution of explanations with higher presence of gaze-like vs. pose-like language is shown for (A) humans separated by sport, (B) models separated by sport, and (C) models separated by prompting strategy.

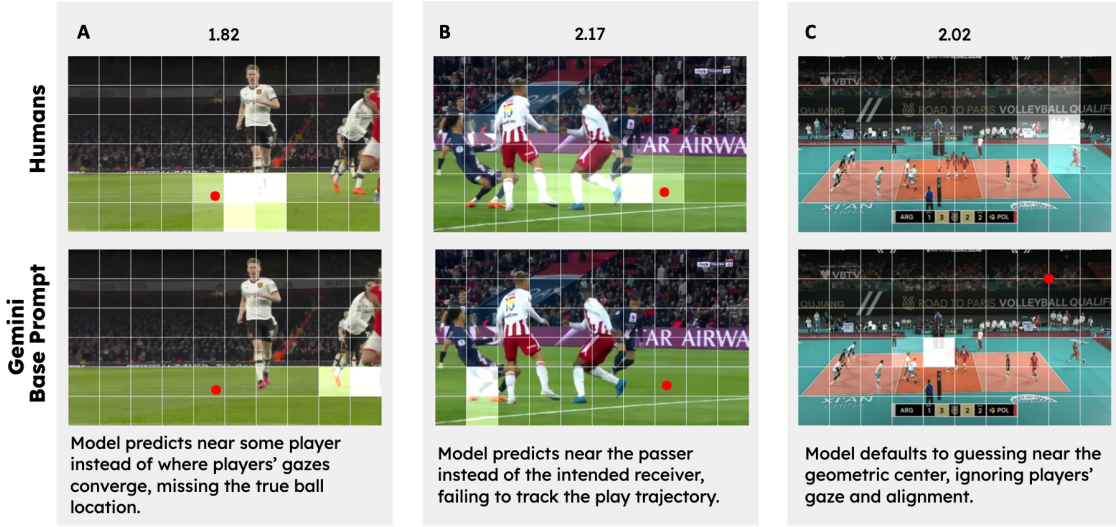


Figure 8. **Qualitative failure modes** demonstrated using Gemini using Base Prompt and human predictions. Red dots mark ground-truth ball locations, white squares mark model-predicted cells, and shaded heatmaps show prediction density. Wasserstein distances between the model and human for each example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

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.

6. Conclusion

We introduced SPOT THE BALL, a benchmark evaluating how vision-language models infer hidden objects from social cues in sports scenes. Across soccer, volleyball, and basketball, humans consistently outperform models by substantial margins. Models exhibit systematic biases like collapsing toward central regions and player proximity, and prompting strategies fail to close this gap. This points to

deeper limitations in how current architectures perceive and reason about social cues like pose and gaze [see also 31]. Progress may require integrating perceptual priors, temporal information, or architectures explicitly designed to capture agentive and relational dynamics.

The need to understand these limitations is increasingly urgent. As VLMs are used to interpret complex scenes in interactive, visually rich environments like in embodied AI, their ability to make inferences from human pose, gaze, and spatial structure becomes central to evaluating what kinds of reasoning they can or cannot perform. Indeed, SPOT THE BALL isolates a core component of visual social inference that humans use effortlessly but current models struggle with. By releasing our dataset, pipeline, and evaluation code, we aim to make these limitations visible and to support systematic progress on this capability.

References

- [1] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems*, pages 83548–83599. Curran Associates, Inc., 2024. [2](#)
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. [3](#)
- [3] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21(1): 37–46, 1985. [1](#)
- [4] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. [3](#)
- [5] Lindsey J. Byom and Bilge Mutlu. Theory of mind: mechanisms, methods, and new directions. *Frontiers in Human Neuroscience*, 7:413, 2013. [1](#)
- [6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. [4](#)
- [7] Kang Chen, Zheng Lian, Haiyang Sun, Rui Liu, Jiangyan Yi, Bin Liu, and Jianhua Tao. Can deception detection go deeper? dataset, evaluation, and benchmark for deception reasoning, 2024. [2](#)
- [8] Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. Emotionqueen: A benchmark for evaluating empathy of large language models, 2024. [2](#)
- [9] B. de Gelder, A.w. de Borst, and R. Watson. The perception of emotion in body expressions. *WIREs Cognitive Science*, 6(2):149–158, 2015. [2](#)
- [10] Daniel C. Dennett. *The Intentional Stance*. The MIT Press, 1989. [1](#)
- [11] Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. Somi-tom: Evaluating multi-perspective theory of mind in embodied social interactions, 2025. [2](#)
- [12] Sue Fletcher-Watson, Fiona McConnell, Eleni Manola, and Helen McConachie. Interventions based on the theory of mind cognitive model for autism spectrum disorder (asd). *Cochrane Database of Systematic Reviews*, 2014(3): CD008785, 2014. [1](#)
- [13] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [1](#), [2](#)
- [14] Fritz Heider and Marianne Simmel. An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, 57(2):243–259, 1944. [2](#)
- [15] Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. Apathetic or empathetic? evaluating llms' emotional alignments with humans. In *Advances in Neural Information Processing Systems*, pages 97053–97087. Curran Associates, Inc., 2024. [2](#)
- [16] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering, 2024. [2](#)
- [17] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. [3](#)
- [18] Chris L. Kleinke. Gaze and eye contact: A research review. *Psychological Bulletin*, 100(1):78–100, 1986. [2](#)
- [19] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotic: Emotions in context dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2309–2317, 2017. [2](#)
- [20] Satyapriya Krishna, Andy Zou, Rahul Gupta, Eliot Krzysztow Jones, Nick Winter, Dan Hendrycks, J. Zico Kolter, Matt Fredrikson, and Spyros Matsoukas. D-rex: A benchmark for detecting deceptive reasoning in large language models, 2025. [2](#)
- [21] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topview: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 1786–1807, 2024. [4](#)
- [22] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Visual social relationship recognition, 2018. [2](#)
- [23] Yuxuan Li, Vijay Veerabadran, Michael L. Iuzzolino, Brett D. Roads, Asli Celikyilmaz, and Karl Ridgeway. Ego-tom: Benchmarking theory of mind reasoning from egocentric videos, 2025. [2](#)
- [24] Zhaochen Liu, Kaiwen Gao, Shuyi Liang, Bin Xiao, Limeng Qiao, Lin Ma, and Tingting Jiang. Beyond the visible: Benchmarking occlusion perception in multimodal large language models, 2025. [1](#), [2](#)
- [25] Xiaomeng Ma, Lingyu Gao, and Qihui Xu. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind, 2024. [2](#)
- [26] Matteo Marcuzzo, Alessandro Zangari, Andrea Albarelli, Jose Camacho-Collados, and Mohammad Taher Pilehvar. Morables: A benchmark for assessing abstract moral reasoning in llms with fables, 2025. [2](#)
- [27] Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. Advancing social intelligence in AI agents: Technical challenges and open questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20541–20560, 2024. [2](#)
- [28] Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models, 2025. [1](#), [2](#)
- [29] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. [2](#)

- [30] Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting, 2025. [2](#), [4](#)
- [31] Wenshuo Qin and Leyla Isik. Simple 3d pose features support human and machine social scene understanding. *arXiv preprint arXiv:2511.03988*, 2025. [8](#)
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. [3](#)
- [33] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialliqa: Commonsense reasoning about social interactions, 2019. [2](#)
- [34] Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. Evaluating the moral beliefs encoded in llms, 2023. [2](#)
- [35] Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multimodal multi-agent theory of mind, 2025. [2](#)
- [36] Beate Sodrian. Theory of Mind in Infancy. *Child Development Perspectives*, 5(1):39–43, 2011. [2](#)
- [37] Helen Tager-Flusberg. Evaluating the theory-of-mind hypothesis of autism. *Current Directions in Psychological Science*, 16(6):311–315, 2007. [1](#)
- [38] Alexander Todorov, Christopher Y. Olivola, Ron Dotsch, and Peter Mende-Siedlecki. Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, 66(1):519–545, 2015. [2](#)
- [39] Patrice D Tremoulet and Jacob Feldman. Perception of Animacy from the Motion of a Single Object. *Perception*, 29(8): 943–951, 2000. [2](#)
- [40] Zihan Wang, Isaac Davis, and Julian Jara-Ettinger. Modeling Other Minds: A Computational Account of Social Cognition and Its Development. *Annual Review of Developmental Psychology*, 2025. [2](#)
- [41] Henry M. Wellman. *Making minds: How theory of mind develops*. Oxford University Press, New York, NY, US, 2014. [1](#), [2](#)
- [42] Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssef Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. 2023. [2](#)
- [43] Te-Lin Wu, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Nischal Reddy Chandra, Marjorie Freedman, Ralph M. Weischedel, and Nanyun Peng. Acquired: A dataset for answering counterfactual questions in real-life videos, 2023. [2](#)
- [44] Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore, 2023. Association for Computational Linguistics. [2](#)
- [45] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. [2](#)
- [46] Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, Sixian Li, Wang Pan, Yue Zhao, Jieyu Zhao, Xiangliang Zhang, and Xiuying Chen. Socialmaze: A benchmark for evaluating social reasoning in large language models, 2025. [1](#), [2](#)
- [47] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8799–8809, 2019. [2](#)
- [48] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning, 2019. [2](#)

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.>
```

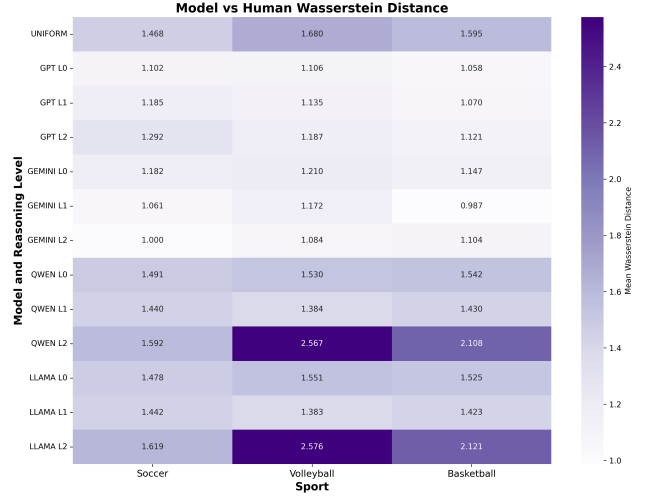


Figure 9. Wasserstein Distances with all the levels and models. L0 refers to Base prompt, L1 to cue-directed prompt and L2 to chain-of-thought prompt.

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.

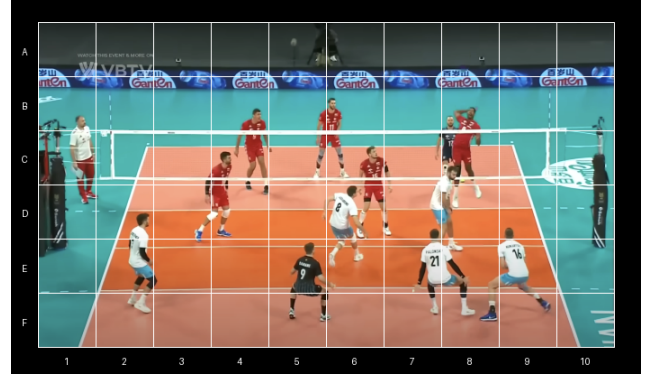


Figure 11. An example of the volleyball image shown to the models. The ball is masked, and a grid is overlaid; the responses in the following tables are generated based on this image.

10. Code and Dataset

All code required to reproduce the experiments, analyses, and figures is available at [REDACTED]. The corresponding dataset, including evaluation splits and annotation meta-data, can be accessed at [REDACTED].

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).

Where are the players located?	The players are located at the following grid positions: - 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
Where are the players looking?	The players are looking in the following directions: - 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. Overall, most players are looking upwards, likely tracking the volleyball in play.
How are the players positioned?	The players are positioned as follows: <i>Team in red (top side of the net):</i> - One player at B4 - One player at B6 - One player at B8 - One player at C3 - One player at C5 - One player at C7 <i>Team in white and blue (bottom side of the net):</i> - One player at E2 - One player at E6 - One player at E7 - One player at E8 - One player at F5 - One player at F6
Reasoning	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.
Grid cell guess	E5

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

Where are the players located?	Here are the locations of the players in the image: * **1, C**: * **2, D**: * **3, D**: * **4, C**: * **5, E**: * **5, B**: * **6, D**: * **6, C**: * **7, E**: * **8, C**: * **8, D**: * **9, E**: * **9, B**: * **10, C**:
Where are the players looking?	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: **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. **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.
How are the players positioned?	Here is the approximate positioning of the players based on the grid: **Team 1 (Left Side):** * Player 1: A-1/2 * Player 2: D-2/3 * Player 3: C-4 * Player 4: D-6 **Team 2 (Right Side):** * Player 1: B-9 * Player 2: D-8 * Player 3: E-7 * Player 4: F-9 **Mid-Court Players:** * Player 5: E-5 * Player 6: B-5 * Player 7: C-6 * Player 8: B-8 * Player 9: C-4
Reasoning	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.
Grid cell guess	Cell: E4