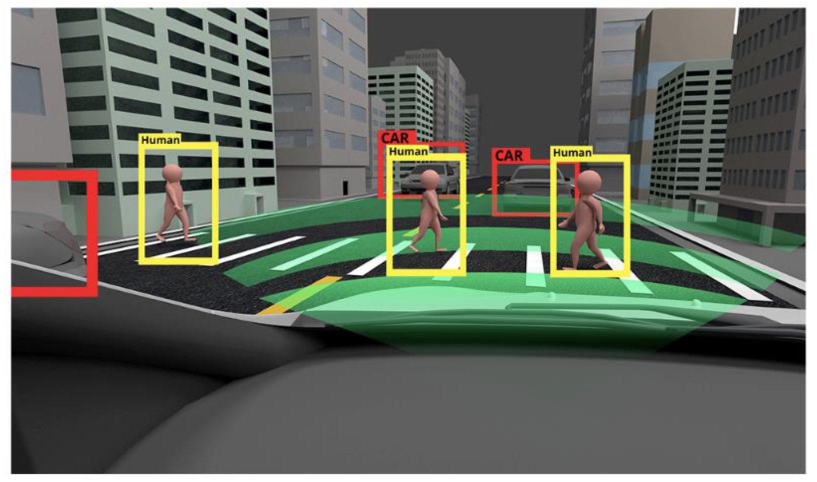
Spot The Ball: Inferring Hidden Information from Human Behavioral Cues

Neha Balamurugan, Sarah Wu, Cristóbal Eyzaguirre, Adam Chun, Gabe Gaw, Tobias Gerstenberg
Stanford University: Computer Science, Psychology

Motivation

How do humans and VLMs reason about occluded elements in a scene from social cues?

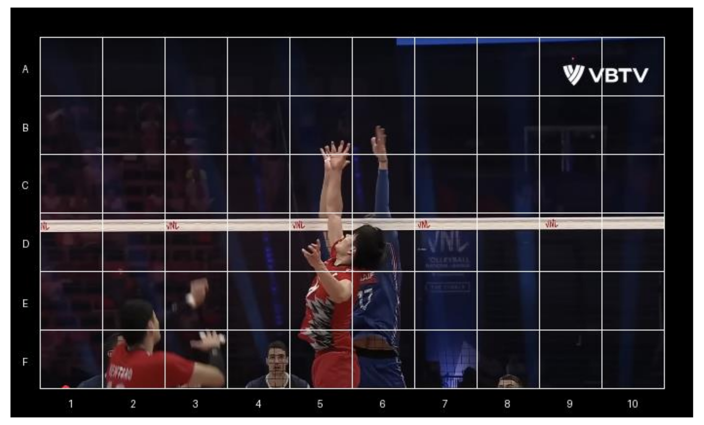


Where is the ball?

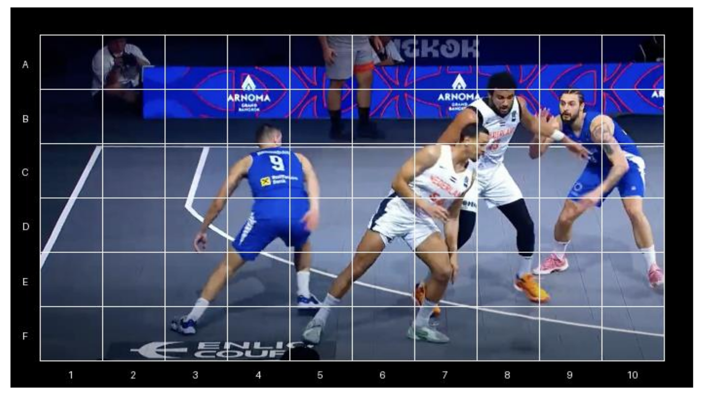
Key Contributions

1. **Multi-sport model benchmark**
2. **3000+ image dataset with metadata**
3. **Image generation pipeline**

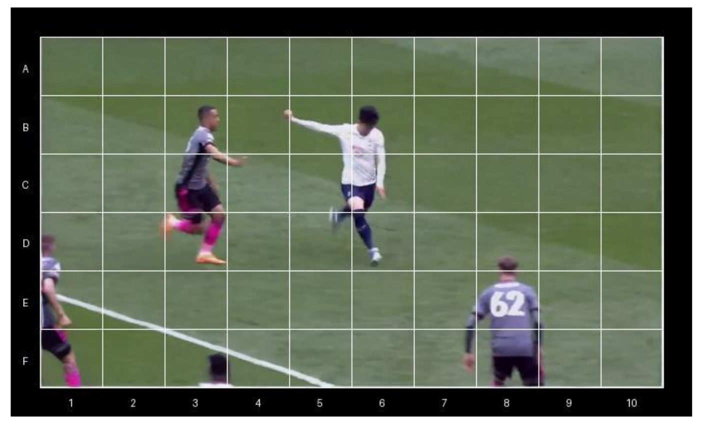
Spot the Ball Task



Volleyball Task



Basketball Task



Soccer Task

Goal: Identify the correct grid cell containing the masked ball.

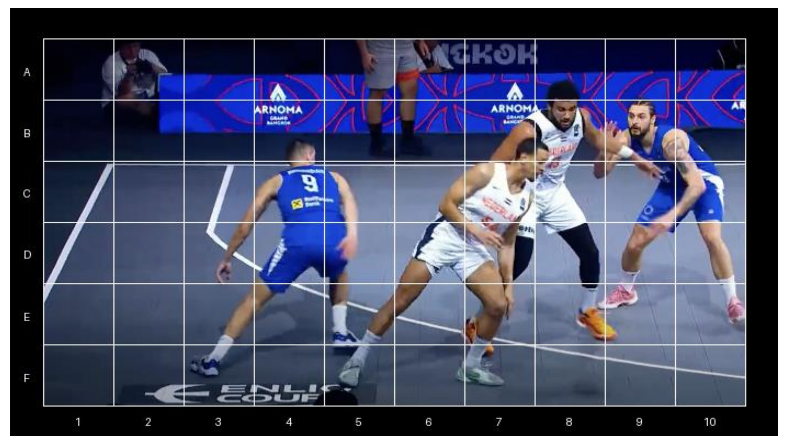
- The ball is removed from real sports images using inpainting.
- Humans and models predict the ball's original location by selecting one of 60 overlaid grid cells.
- Reframes the task as a 60-way classification problem under occlusion.
- The task tests **spatial reasoning without direct visual evidence**.
- We evaluate model performance across **three prompting levels** that add increasing amounts of reasoning support.

Level	Prompting Condition
0	Image + Basic instruction: "Which grid cell contains the ball?"
1	Image + Instruction encourages attention to pose/gaze from image
2	Image + Chain-of-thought prompt over location/pose/gaze of players

Image Generation Pipeline

Original Image

Ball Removal



Grid Overlay

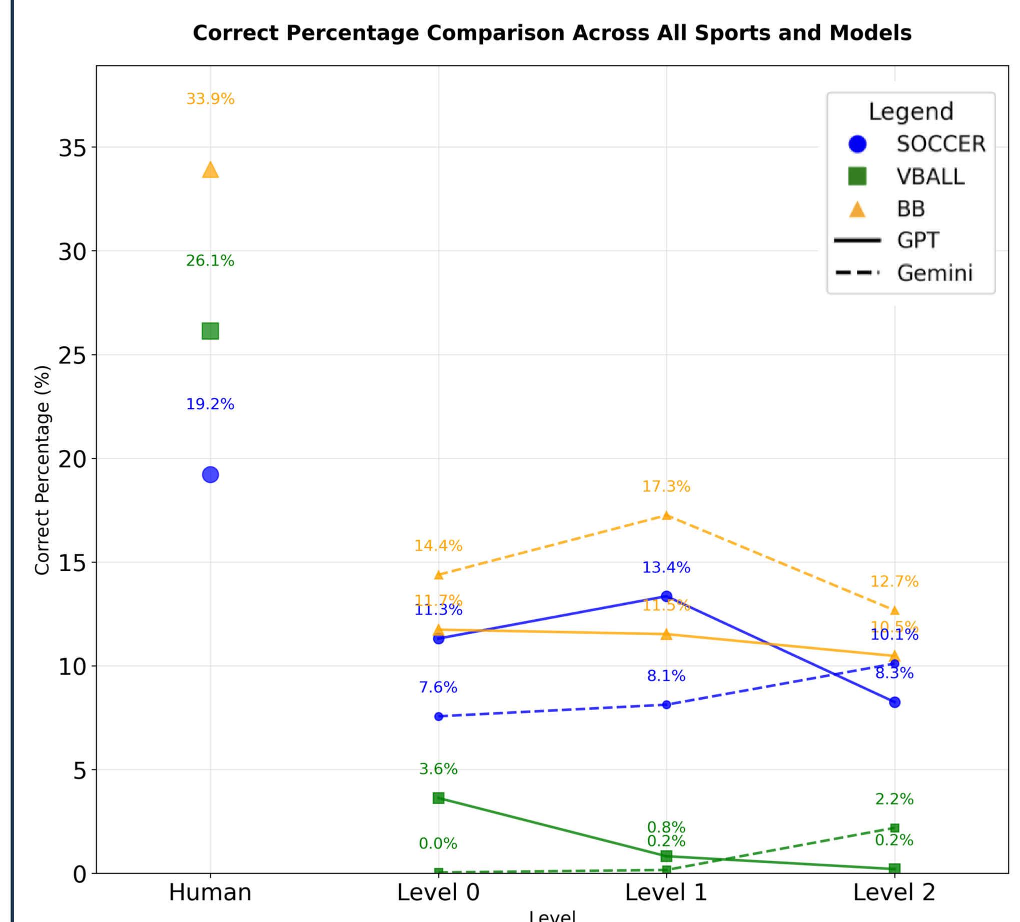
The image generation pipeline enables scalable, reproducible generation of high-quality evaluation stimuli. By decoupling image creation from specific datasets, we support extension to other sports and downstream tasks involving occluded object inference.

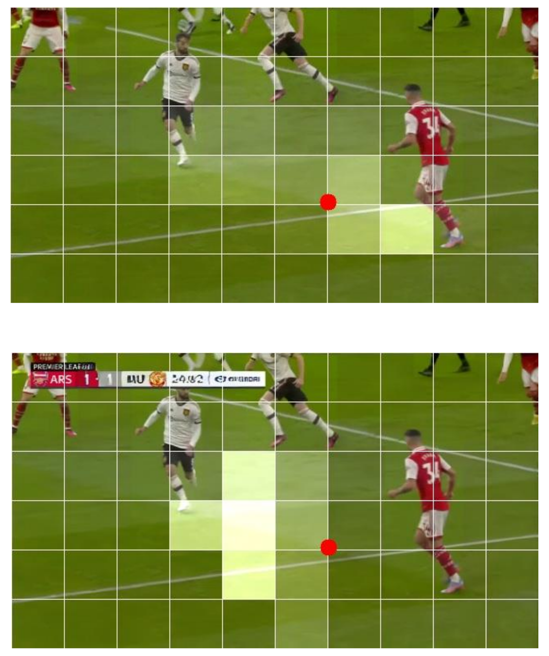
Step 1: Frame Retrieval
- Download YouTube sports videos using keyword-based search.
- Use CLIP to rank and extract frames that match a semantic caption.
- Apply YOLOv3 to detect players and the sports ball.
- Ensure valid player–ball interaction via bounding box overlap checks.

Step 2: Ball Masking and Inpainting
- Create a binary mask over the detected ball.
- Use Stable Diffusion inpainting to generate a realistic, ball-free image while preserving context.

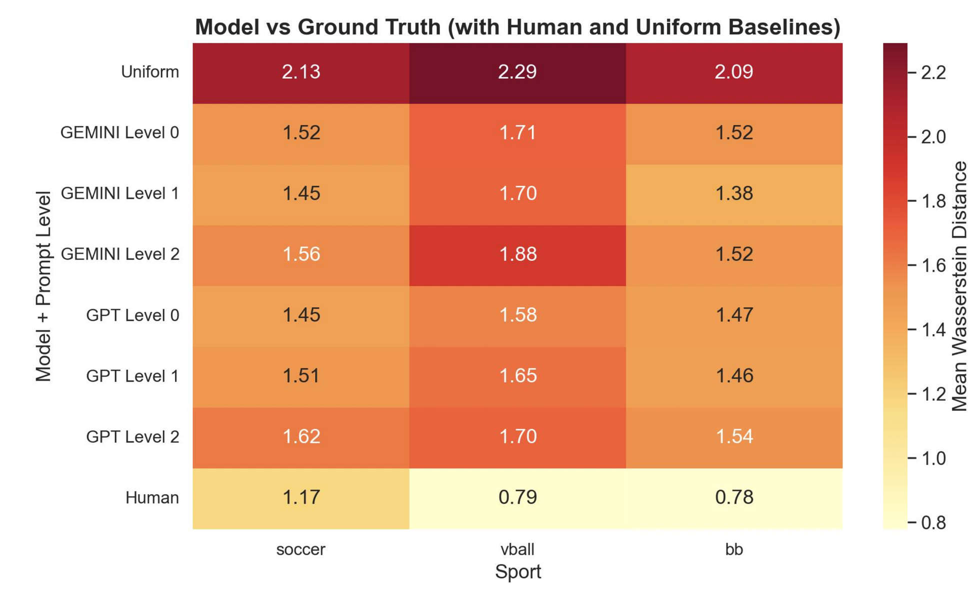
Step 3: Label Generation
- Divide each image into a 6×10 grid.
- Store ball's original coordinates and associated cell label for training and evaluation.

Results and Analysis





Human (top) and Gemini Level 0 (bottom) Prediction distribution in heatmap with ground truth (red)

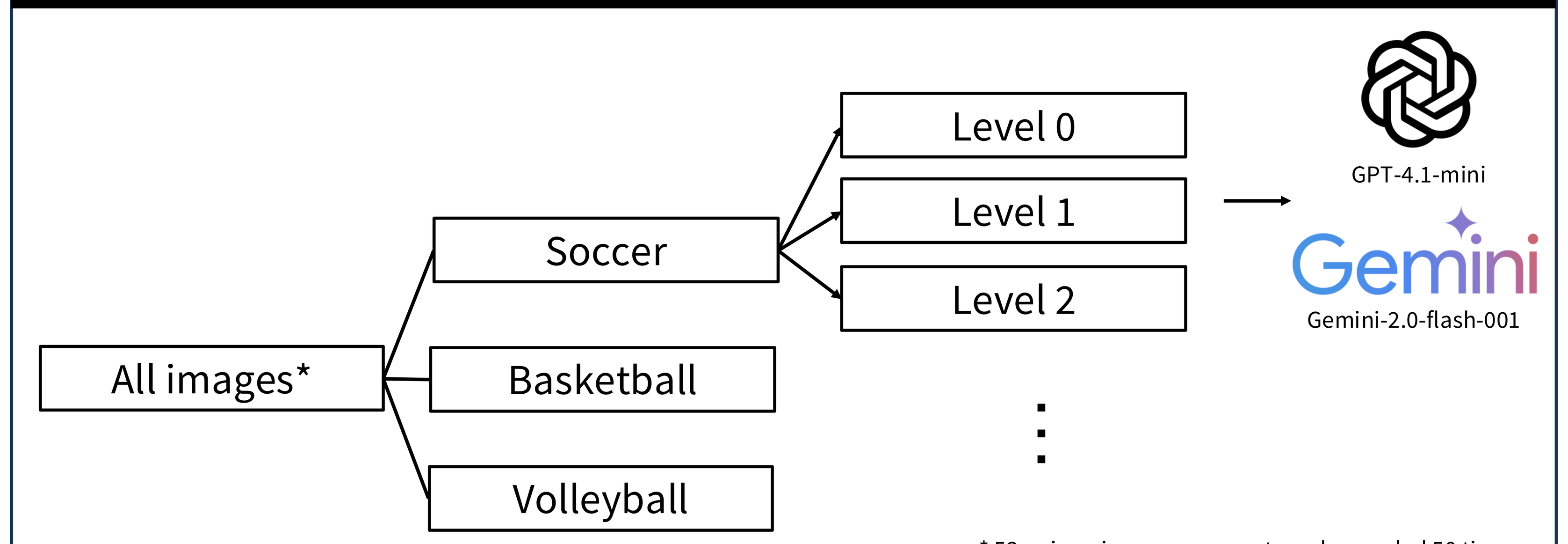


- Humans are more accurate than GPT and Gemini in all three sports tasks
- Models improve with structured prompting (Level 1), but regress with verbose CoT (Level 2)
- GPT and Gemini are more humanlike in soccer and basketball, but struggle in volleyball.

These results indicate:

1. **Language models struggle with occluded object localization without explicit reasoning support**
2. **Prompt engineering can help with performance, but varies by sport and model**

Experiments



* 52 unique images per sport, each sampled 50 times at Level 0 and 1 and 20 times at Level 2 prompting

Future

- Assess how explicit structured inputs (e.g., pose and gaze) affect model predictions.
- Explore improvements from temporal input: how do results change when using videos instead of single frames?
- Incorporate 3D scene information (e.g., from Unity environments) to test reasoning in richer spatial contexts.
- Advance visual models' ability to perform causal inference under occlusion

Pose and Gaze as structured cues