SPORTD: HOW DO VLMS PHYSICALLY STRATEGIZE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) can describe a scene, but can they *act* well within one? We study whether VLMs can make sound strategic decisions, using soccer as an objective testbed with quantifiably-valued actions. We introduce **SportD**, a dataset and evaluation consisting of 1421 decision scenarios across professional men's and women's soccer games, where a VLM must decide what action to take next. Models on average select the optimal action around 27% of the time, less often than the professional players, and capture markedly less of the value at stake. Furthermore, they exhibit a clear preference for safer actions, favoring lower-variance, lower-value choices that also make less physical progress toward goal. Frontier VLMs are better at estimating whether an action will succeed, placing the highest-success-probability action among their top choices in 72–85% of cases. Yet VLMs systematically conflate likelihood with value, assigning higher value to actions that are more likely to succeed ($\rho = +0.30$ to $+0.52$), despite no such relationship in the ground truth ($\rho = -0.08$). Modifying the deliberation instructions to encourage risk-taking brings the frontier models closer to the players' skill levels. **SportD** opens a new direction for rigorously evaluating physical strategic decision-making in VLMs, showing that careful decomposition of their choices can reveal the mechanisms underlying systematic biases such as risk aversion.

1 INTRODUCTION

Vision-language models (VLMs) have become strong at understanding visual scenes, but whether they can reason strategically about how to act in the real physical world remains unclear. This requires models to integrate fine-grained spatial perception with reasoning over possible actions and their consequences. Such reasoning is essential for agents to operate in the physical world, but it is largely absent from benchmarks centered on recognition, captioning, and visual question answering, which evaluate representations of a scene rather than decisions within it.

Team sports let us study a broader question: can VLMs turn visual understanding into good strategic actions? Existing sports benchmarks primarily test understanding (Xia et al., 2025; 2026) of what is happening, while benchmarks evaluating strategic-reasoning capabilities largely operate in abstract or simulated environments (Xu et al., 2026). We instead test strategic decision-making in the physical world, asking VLMs to select the next action given an observation of the environment, and evaluating their choices against actions that can be quantifiably valued.

We introduce **SportD**, a comprehensive evaluation and dataset of 1421 on-ball situations from the 2022 men's and 2023 women's World Cups (Fig. 1). Soccer provides a useful setting for studying strategic physical decision-making: each possession unfolds among many moving players, admitting many plausible actions. Soccer also offers rich event and tracking data, together with established models for valuing actions. For each example, a VLM observes the seconds leading up to an on-ball situation and selects the next action, either passing to a specific teammate or shooting. Because every candidate action can be assigned a quantitative value, **SportD** measures not only whether a model selects the best action, but whether its suboptimal choices follow systematic patterns.

Our findings shed light on the limitations of VLMs in strategically reasoning about the physical world. In **SportD**, VLMs select the optimal action on 27.4% of decisions on average, less often than the professional players (35.1%), and capture far less of the value at stake (mean skill 0.18 versus

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

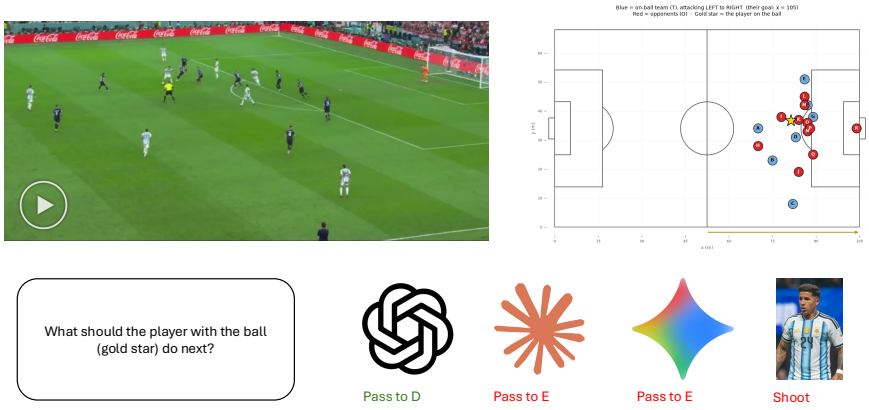


Figure 1: **The SportD task.** From the run-up clip, decision frame, and top-down map (lettered circles for players, a gold star for the player on the ball, an arrow for the attacking direction), the VLM chooses one action: a pass to a lettered teammate or a shot. *Bottom right:* the three VLMs' and the real player's choices, green if optimal and red otherwise. Here one VLM, GPT 5.6 Sol, finds the optimal pass; the others, and the player (who shot), do not.

0.34). More importantly, VLMs are risk averse, with models choosing lower-variance actions than the optimum. The models' internal beliefs reveal a mismatch between likelihood and value, identifying which actions are most likely to succeed with 72–85% recall, but the most valuable actions only 50–58% of the time. Moreover, while success probability and expected value are not correlated in the value model ($\rho = -0.08$), every VLM assigns them a positive relationship ($\rho = +0.30$ to $+0.52$). These results suggest that VLM risk aversion arises not merely from poor perception or imitation, but from systematically treating what is easier to execute as what is more valuable.

2 SPORTD

2.1 DATASET

SportD contains 1421 on-ball decisions sourced from two professional tournaments — 726 from the 2022 men's World Cup and 695 from the 2023 women's World Cup (StatsBomb, 2023). Each decision d is a single moment at which the player on the ball must act, and we present to the VLM (i) images from the previous ~ 5 seconds, sampled at 0.5 second intervals; (ii) an image of the moment immediately preceding the decision; and (iii) a bird's-eye view map showing the positions of all visible players. The main prompt is given in App. A.1.

The VLM selects exactly one action a from the available action set $\mathcal{A}_d$, where

$$\mathcal{A}_d = \{\texttt{SHOOT}\} \cup \{\text{pass to teammate } t : t \in T_d\}. \tag{1}$$

Here, T_d is the set of visible teammates in the image. The median event offers approximately six candidate actions.

2.2 VALUE MODEL

We value actions with the VAEP (Valuing Actions by Estimating Probabilities) model (Decroos et al., 2019). Each match is first converted into SPADL (Soccer Player Action Description Language), a unified representation of on-ball actions in which each action is described by its type, start and end location, body part, and outcome. VAEP values an action by how it changes the acting team's chances of scoring and conceding. From each game state, separate models estimate the probabilities that the team scores or concedes within the next ten actions, giving the state value $V = P_{\text{score}} - P_{\text{concede}}$.

The value of an action is the change in V from immediately before to immediately after it. We fit separate value stacks for the men's and women's benchmarks. The men's stack is trained on

Euro 2020 and Euro 2024, with the 2022 men's World Cup reserved for evaluation. The women's stack uses the same specification, trained on the Euro 2022 and Euro 2025 women's matches, and evaluated on the 2023 women's World Cup.

2.3 METRICS

2.3.1 ACTION VALUATION

We assign a numerical value to every action with VAEP (Decroos et al., 2019), which scores an action by how much it changes the team's chance of scoring next. Higher-valued actions are deemed more optimal to carry out. An action succeeds with some probability and yields one value if it does and another if it does not, so its expected value is

$$V(a) \;=\; p_a\, v_a^+ + (1 - p_a)\, v_a^-, \tag{2}$$

where v_a^+ and v_a^- are the values of the game state that results if the action succeeds or fails, respectively, and p_a is its success probability. For a pass, p_a is the completion probability and v_a^+, v_a^- are the values of keeping possession at the target location versus losing it. For a shot, p_a is the expected-goal probability (Lucey et al., 2015) with a scored goal worth $v_a^+ = 1$.

2.3.2 SCORING A DECISION

Regret Regret is the value lost relative to the optimal action. For a chosen action with value V,

$$R = V^\star - V \geq 0, \tag{3}$$

where $V^\star$ is the value of the best available action. If the optimal action was selected, then R is 0. Better actions have lower regret R.

We group the events into three tiers of human decision quality. The `low` tier contains the 499 events where the player chose the optimal action ($a^\star$), matching the players' accuracy in Table 1. Among the events where the player chose a suboptimal action, the `medium` tier contains 482 with small regret ($R \leq 0.034$) and the `high` tier contains 440 with larger regret ($R > 0.034$).

Skill Values vary substantially across events, so mean regret can be dominated by a small number of high-value decisions. We report *skill* that rescales each event to its own range,

$$s \;=\; \frac{V(\text{chosen action}) - \bar{V}}{V^\star - \bar{V}}, \tag{4}$$

where $\bar{V}$ is the average value of the actions on offer. Skill is 1 for the best action, 0 for a choice no better than picking at random, and negative for worse.

Risk To measure whether a VLM prefers *safe* actions, we use the variance of each action's payoff under the same value model,

$$\text{Risk}(a) = p_a(1 - p_a)\big(v_a^+ - v_a^-\big)^2, \tag{5}$$

where p_a is the probability that action a succeeds, and v_a^+ and v_a^- are its values if it succeeds or fails.

3 RESULTS

3.1 VLMS PERFORM WORSE THAN PROFESSIONAL PLAYERS

We evaluate five frontier VLMs, comprising of GPT 5.6 Sol, Claude Opus 4.8, Gemini 3.5 Flash, Qwen3-VL-8B, and Llama 4 Maverick on **SportD**. As shown in Table 1, no VLM matches the players' accuracy of 35.1% or mean skill of 0.34. On average, proprietary VLMs attain an accuracy of 30.2% and mean skill of 0.24, and the open-source models, Qwen3-VL-8B and Llama-4-Maverick, attain the lowest accuracy and skill, at roughly 23% accuracy with skills 0.07 and 0.11, respectively.

Even if we consider an answer to be correct if it ranks among the best $X\%$ of the available actions, models are still worse than humans. At a top-20% threshold, the VLMs reach 33.4–46.3% versus

Table 1: Main results. Accuracy (absolute Top-$X\%$), skill, and risk percentile. Top-$X\%$ is the fraction of choices ranking among the best $X\%$ of available actions.

strategy	accuracy	top-20%	top-30%	mean skill	risk percentile
real player	35.1%	49.8%	57.4%	0.335	60
GPT 5.6 Sol	34.6%	46.3%	54.1%	0.298	50
Gemini 3.5 Flash	30.3%	43.3%	51.6%	0.263	53
Claude Opus 4.8	25.8%	36.7%	44.3%	0.157	46
Qwen3-VL-8B	23.1%	33.5%	40.3%	0.069	53
Llama-4-Maverick	23.2%	34.9%	41.3%	0.107	48

Table 2: Per-tournament breakdown of the headline metrics shown in Table 1.

	accuracy		mean skill		risk percentile	
strategy	men	women	men	women	men	women
real player	35.3%	35.0%	0.315	0.356	60	60
GPT 5.6 Sol	34.0%	35.1%	0.268	0.330	51	48
Gemini 3.5 Flash	29.1%	31.6%	0.224	0.304	56	51
Claude Opus 4.8	20.8%	31.1%	0.085	0.233	46	46
Qwen3-VL-8B	23.7%	22.4%	0.070	0.069	54	53
Llama-4-Maverick	23.3%	23.2%	0.091	0.123	49	47

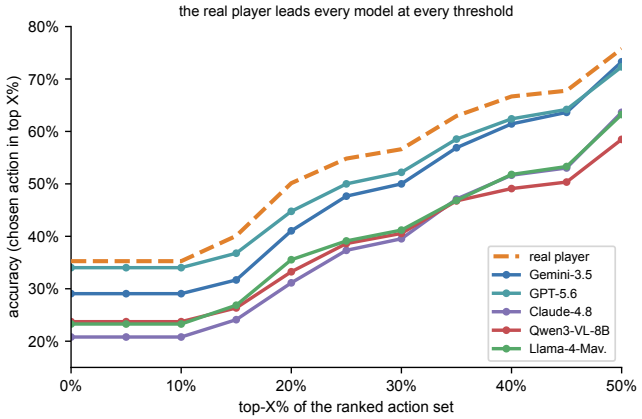


Figure 2: Rank-percentile accuracy: fraction of events on which the chosen action falls within the top $X\%$ of the ranked action set. The real player attains a higher adjusted accuracy than every VLM at every threshold. The comparison is purely ordinal.

49.8% for the real player, and the ordering holds at every threshold of the full rank-percentile curve (Fig. 2). Relaxing the metric does not close the performance gap between the VLMs and the player. Even when we make the visual format easier for the VLM to parse by labeling the players on the actual gameplay scene directly, there is no meaningful difference in accuracy (Appendix E).

3.2 VLMs CARRY OUT RISK-AVERSE ACTIONS

The models' conservatism is visible directly in the geometry of their passes. For each pass selected by the VLM, we measure *goal-ward gain*: the number of meters the ball moves toward the attacking goal (Fig. 4). The optimal pass is decisively forward, advancing the ball 9.0 m on average and progressing toward goal on 86% of events. The VLMs' passes barely move the ball forward — Gemini $+1.2$ m, GPT -1.2 m, Claude -1.7 m (i.e. near zero or backward) — going forward on



Figure 3: Example of an image with labels for all players and the on-ball player (yellow oval) used for the easier visual control experiment (Appendix E)

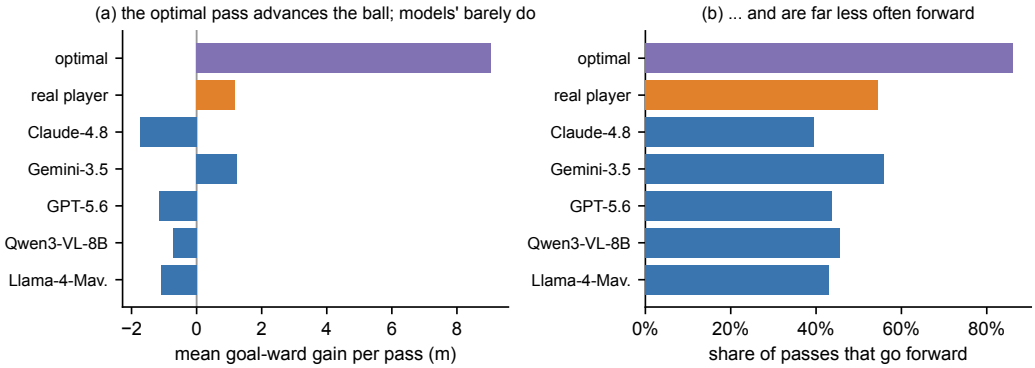


Figure 4: Progression of the chosen pass. (a) Mean goal-ward gain (meters the ball moves toward the attacking goal): the optimal pass advances it $\sim$9 m, the real players $\sim$2 m, and the VLMs near zero or backward. (b) Share of passes that go forward at all.

only 40–56% of events, comparable to the real players (+1.2 m, 55%) yet far short of the optimum, even though the magnitudes of the pass distances between VLMs and the optimal baseline are nearly equal (14–15 m vs. 15 m for $a^\star$).

We see similar trends when we look at the risk metric defined in 2.3. Across all possible decision points, the optimal action lies at the 75th risk percentile and the real player's choice at the 60th. VLMs select actions much closer to the middle of the distribution (46th–53rd percentile; Table 1). This conservatism carries a clear value cost: mean skill ranges from 0.067–0.298 across VLMs, compared with 0.335 for the real player. The models therefore systematically favor safer actions while leaving value unrealized.

Risk tolerance converges across model generations. We run the same evaluation on older model generations: GPT-4o, Gemini 2.5 Flash, and Claude Haiku 4.5. The older models are more divergent in their risk profiles, with their mean risk percentile of the chosen action scattering across 10.8 points. Gemini 2.5 Flash is the most aggressive at 56.2, GPT-4o the most conservative at 45.4, and Claude Haiku 4.5 between them at 48.1. The current frontier roster of models collapses into a 7.9-point band (45.5–53.4; Fig. 5). The families converge from both sides: GPT rises, Gemini falls, and Claude edges down. Increasing capability did not diversify risk tolerance across the frontier, instead *standardizing* it.

3.3 The models estimate likelihood of success well, but value poorly

Why are VLMs risk-averse? One possibility is a distortion in their internal consistency in which they may misjudge the consequences and returns of actions that are more likely to succeed. But do they even know which actions are more likely to succeed in the first place? In human decision-making, probability and value are separate inputs to choice, yet people conflate them: a single impression

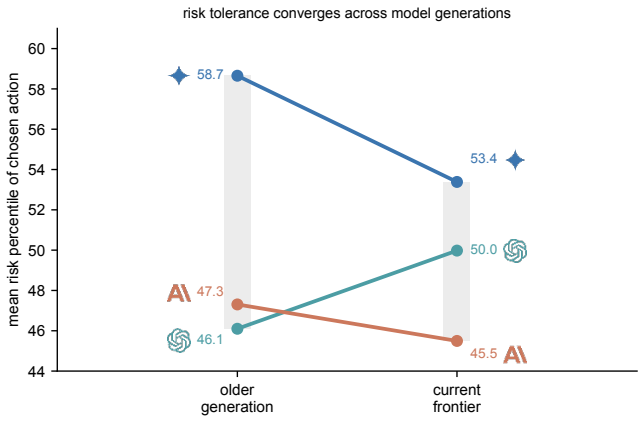


Figure 5: Risk tolerance converges across model generations. The y-axis is the mean within-event risk percentile of the chosen action. Each line is one model family, from its older generation to the current frontier. The spread shrinks from 10.8 to 7.9 points. The families converge from both sides: GPT rises ($45.4 \rightarrow 50.0$), Gemini falls ($56.2 \rightarrow 53.4$), and Claude edges down ($48.1 \rightarrow 45.5$).

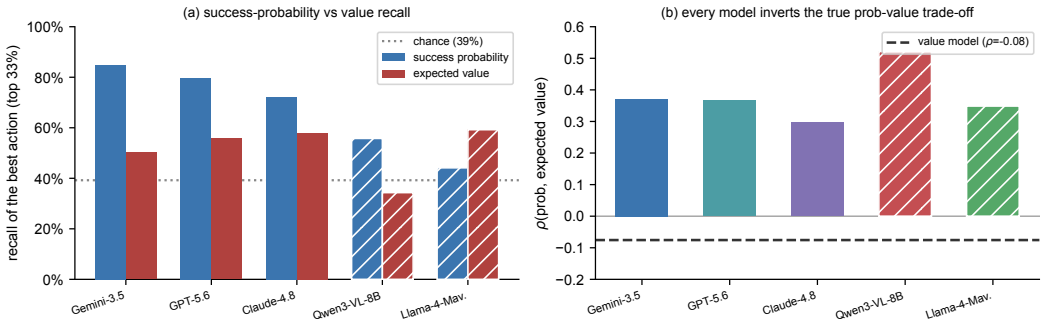


Figure 6: VLMs estimate likelihood better than value. Most models identify high-success-probability actions far more reliably than high-value actions (left), and systematically assign higher value to actions they believe are more likely to succeed (right), reversing the ground-truth probability–value relationship.

can drive both risk and benefit judgments (Finucane et al., 2000), so that safer-seeming actions are also judged more valuable. We test for the same conflation by eliciting each model's beliefs on the same events as two separate rankings (App. A.3): one orders every legal action by its probability of success, the other by its expected value. Each is prompted in a fresh context so that neither response can condition on the other.

We evaluate each model's ranking with recall at the top third of the ranked actions: whether the truly best action appears among the top 33%. Figure 7 sweeps every cutoff. Models identify which actions are most likely to succeed well, reaching success-probability recall of 72–85%. They identify which actions are most valuable far less well, with expected-value recall of only 50–58%. The gap holds at every cutoff (Fig. 7): the models rank safety far more reliably than value. Table 4 gives the per-model recall.

These value errors follow a clear pattern. Models systematically overvalue actions that are more likely to succeed. Across the available actions an action's success probability and its expected value are exhibit almost no correlation (Spearman $\rho = -0.08$). Instead, every VLM asserts a *positive* correlation between probability and expected value ($\rho = +0.30$ to $+0.52$; Fig. 6b, per-model values in Table 5). The models behave as though the easier action were the better one. The conservatism therefore reflects a distorted notion of value, with models treating success probability as a proxy for how worthwhile an action is.

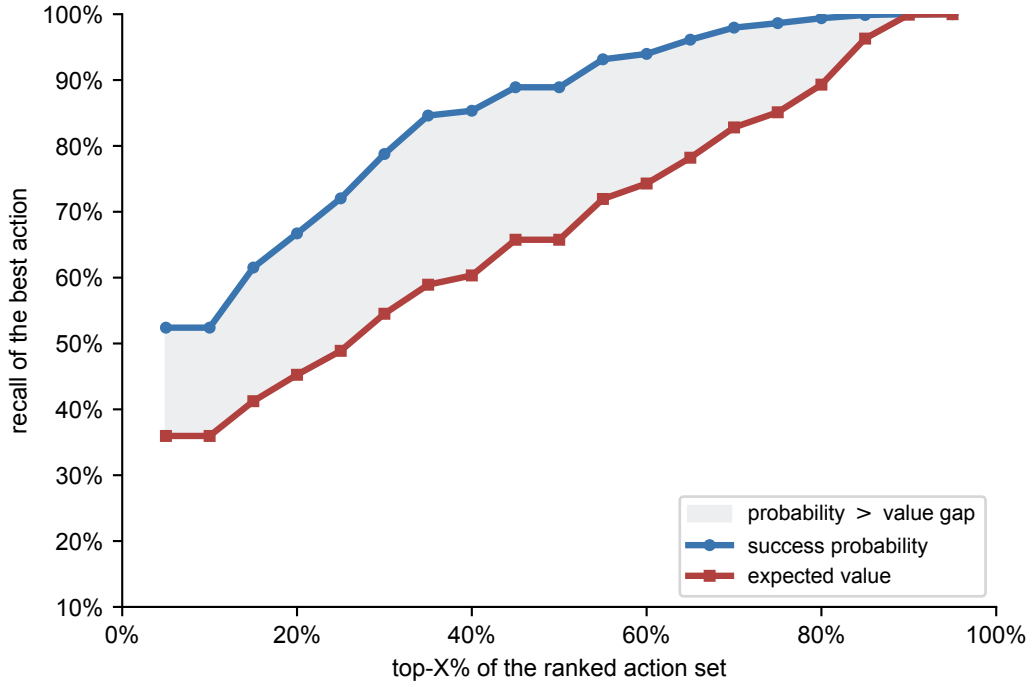


Figure 7: Recall of the best action as the top-X% cutoff sweeps. Success-probability recall (blue) exceeds expected-value recall (red) at every cutoff; the gap is shaded. The models rank likelihood far better than value.

3.4 VLMs RECOVER VALUE WHEN STEERED TOWARD RISK, AND FORFEIT IT WHEN STEERED AWAY

Motivated by the risk aversion and valuation miscalibration observed in Sec. 3.2 and Sec. 3.3, we test whether encouraging risk-taking improves VLM decisions. We modify only the deliberation instructions to encourage riskier attacking actions (SEEK) or prioritize retaining possession (AVOID). App. A.4 provides the full prompts. We evaluate each event on the men's set, using the unmodified benchmark run from Sec. 3.1 as the NEUTRAL baseline.

The frontier models are steerable, and risk-seeking prompts recover value. The SEEK instructions move GPT, Claude, and Gemini toward riskier actions and increase their skill (Fig. 8). GPT rises from 0.27 to 0.37, surpassing the real players (0.32 on the men's set), while Claude more than doubles from 0.09 to 0.21. GPT is the only model to reach the players' level; under SEEK the other closed models improve but still trail the humans (Gemini 0.28, Claude 0.21), and the open Llama does not benefit (0.04). AVOID moves all models toward safer actions and lowers skill, confirming that their default policies lie on the conservative side of optimal.

3.5 VLMs MATCH THE HUMAN ACTIONS LESS AS THE HUMANS PERFORM WORSE

Do VLMs just copy the human's choice? As the humans' actions grow more suboptimal (larger regret R), the frontier models copy it less. Across the three regret strata as established in Section 2.3, the human-match rate falls by 23, 20, and 15 percentage points for GPT, Gemini, and Claude per 0.1 increase in R, whereas open models already have a low accuracy to begin with so the decline is not as visible there. The models therefore do not simply reproduce human mistakes; they increasingly diverge from the human choice as it becomes more suboptimal.

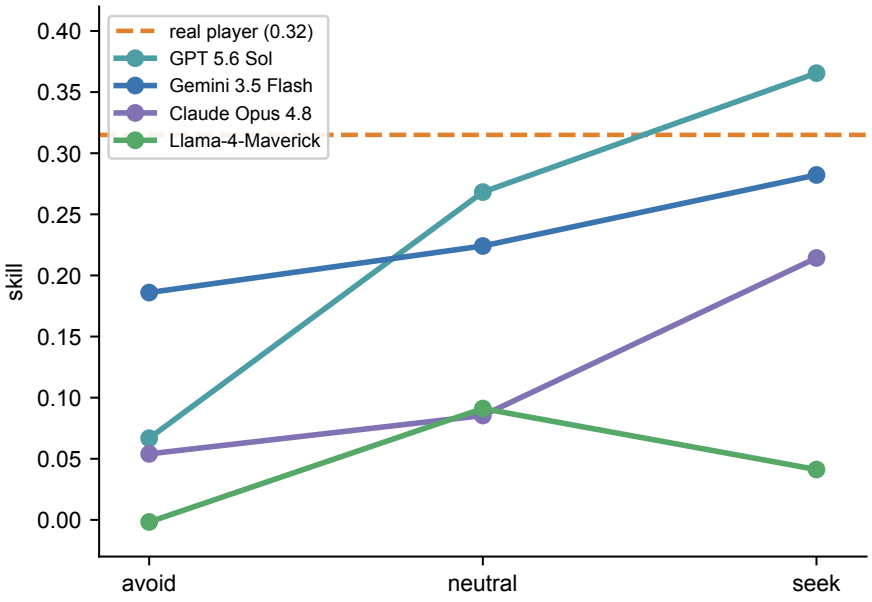


Figure 8: Risk-steering the decision prompt on the men's set. Each line is one model; the dashed line is the real players' skill. Steering from AVOID to NEUTRAL to SEEK raises skill for the closed-source models and lifts GPT to the players' level. AVOID lowers every model, and the weakest model, Llama-4-Maverick, moves the wrong way.

Table 3: As the human's action grows more suboptimal (`low`→`medium`→`high`) , the frontier models match the human action less as well.

	`low` $= a^h$	`medium` $= a^h$	`high` $= a^h$
GPT 5.6 Sol	58.7%	37.8%	31.1%
Gemini 3.5 Flash	47.1%	30.3%	23.2%
Claude Opus 4.8	35.0%	23.2%	16.9%
Qwen3-VL-8B	27.5%	17.4%	19.6%
Llama-4-Maverick	26.7%	28.2%	20.2%

4 Related Work

Vision–language models and visual reasoning. A large body of work evaluates whether VLMs can understand and reason about visual observations, including compositional and relational reasoning (Johnson et al., 2017; Antol et al., 2015; Hudson & Manning, 2019), spatial reasoning (Liu et al., 2023; Chen et al., 2024), expert multimodal knowledge (Yue et al., 2024), challenging perceptual failures (Fu et al., 2024; Tong et al., 2024), and temporal understanding in video and egocentric settings (Xiao et al., 2021; Mangalam et al., 2023; Li et al., 2024; Grauman et al., 2022). These benchmarks largely evaluate understanding or reasoning *about* visual observations. **SportD** instead asks whether a VLM can use those observations to choose a strategically valuable action, evaluating not what the model can describe about a scene but what it decides to do within one.

Physical and strategic reasoning. Physical-reasoning benchmarks study whether models can predict how scenes will evolve (Battaglia et al., 2013; Bakhtin et al., 2019; Yi et al., 2020; Bear et al., 2021; Piloto et al., 2022), while world models learn latent dynamics that can support prediction and planning (Ha & Schmidhuber, 2018; Hafner et al., 2025; Bruce et al., 2024; LeCun, 2022). Separately, strategic agents have reached strong performance in games such as Go, chess, poker,

Stratego, and real-time strategy (Silver et al., 2016; 2018; Brown & Sandholm, 2019; Perolat et al., 2022; Vinyals et al., 2019; Berner et al., 2019), and language-conditioned agents can negotiate and cooperate (Bakhtin et al., 2022; Park et al., 2023). These systems generally act from symbolic or simulated state; similarly, VS-Bench (Xu et al., 2026) evaluates strategic VLM agents in abstract multi-agent environments. **SportD** connects physical perception and strategic choice by requiring models to select among alternative actions directly from real-world visual observations.

Sports understanding and action valuation. Sports analytics provides quantitative models for evaluating decisions, including possession-value and space-control approaches (Decroos et al., 2019; Fernández et al., 2021; Cervone et al., 2016; Fernández & Bornn, 2018; Spearman, 2018; Lucey et al., 2015; Decroos & Davis, 2019). In parallel, multimodal sports benchmarks study action spotting, broadcast understanding, perception, and comprehension (Giancola et al., 2018; Deliège et al., 2021; Xia et al., 2025; 2026). To our knowledge, these lines have not been combined to evaluate whether a VLM can make a strategically valuable decision given the current state of the environment it is in, while considering alternatives holistically. Related work on behavioral cloning learns to reproduce expert behavior (Pomerleau, 1991; Ross et al., 2011; Ho & Ermon, 2016), while vision–language–action and generalist agents extend learned action selection to embodied settings (Reed et al., 2022; Driess et al., 2023; Brohan et al., 2023). **SportD** instead connects visual understanding with value-based action selection, enabling us to study not only whether VLMs make good decisions, but also the systematic biases underlying the actions they choose.

5 Conclusion

We introduce **SportD**, a benchmark for evaluating whether VLMs can turn visual understanding into strategically valuable actions in the physical world. Rather than asking models to describe what is happening in a scene, **SportD** asks what a player should do next and evaluates that decision against the estimated value of every available alternative. Across 1421 decisions from the 2022 men's and 2023 women's FIFA World Cups, every VLM we evaluate falls short of the professional players who faced the same situations. More importantly, their errors exhibit a consistent structure.

VLMs systematically favor safer actions. They choose lower-risk actions than the value-optimal alternative on most events and make substantially less progressive passes, sacrificing expected value alongside risk. The models' elicited beliefs point to a miscalibration of valuation. Frontier VLMs are substantially better at identifying which actions are likely to succeed than which actions are most valuable. Yet across all models, estimated success probability is positively associated with estimated value, even though the corresponding relationship under the value model is slightly negative. VLMs therefore behave as though actions that are easier to execute are also more worthwhile. The evidence is therefore most consistent with a systematic distortion in how VLMs translate beliefs about possible outcomes into strategic value, rather than with a deliberate risk–reward tradeoff. Steering models towards favoring more riskier actions increases their average skill, but still lagging behind professional players.

Soccer provides a controlled setting for exposing this distinction because it combines complex real-world visual observations, multiple plausible actions, and quantitative estimates of their consequences. Future work can extend **SportD** to other sports and physical domains, richer spatiotemporal value models, and broader action spaces including other movements, like dribbling. More generally, our results highlight a capability that descriptive vision benchmarks can miss: a model may understand what is likely to happen without understanding which possible future is worth pursuing. Evaluating this gap will be important as multimodal models are increasingly expected not only to perceive the physical world, but to make decisions within it.

AI use statement

In this work, we used generative AI tools to propose and refine hypotheses, to design and give feedback on our research methodology and experiments, to implement methods (writing and debugging the evaluation pipeline, the analysis code, and the figure-generation scripts), to clean and reformat the dataset (joining and normalizing the answer keys and assembling the released data). We did not use generative AI tools to develop the paper's conceptual framework or to interpret results. Generating synthetic datasets, formulating mathematical claims, providing critical ingredients

for or writing proofs, translation, and qualitative or thematic data analysis are not applicable to this work. We additionally used generative AI tools to assist with drafting and editing the manuscript and to help identify related literature. We have reviewed all AI-assisted work: every AI-written or AI-modified piece of code was read and tested by the authors, every reported number was recomputed and checked against the paper, and all AI-assisted text was verified for accuracy. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A new benchmark for physical reasoning. In *NeurIPS*, 2019.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

Daniel Bear, Elias Wang, Damian Mrowca, Felix Jedidja Binder, Hsiao-Yu Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin A. Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel LK Yamins, and Judith E Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *NeurIPS*, 2021.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d. O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023.

Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *ICML*, 2024.

Daniel Cervone, Alex D'Amour, Luke Bornn, and Kirk Goldsberry. A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, 2016.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024.

Tom Decroos and Jesse Davis. Player vectors: Characterizing soccer players' playing style from match event streams. In *ECML-PKDD*, 2019.

Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *SIGKDD*, 2019.

Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *CVsports Workshop @ CVPR*, 2021.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *ICML*, 2023.

Javier Fernández and Luke Bornn. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *MIT Sloan Sports Analytics Conference*, 2018.

Javier Fernández, Luke Bornn, and Daniel Cervone. A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, 110(6):1389–1427, 2021.

Melissa L. Finucane, Ali Alhakami, Paul Slovic, and Stephen M. Johnson. The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, 13(1):1–17, 2000.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *Computer Vision in Sports Workshop @ CVPR*, 2018.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *NeurIPS*, 2016.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Yann LeCun. A path towards autonomous machine intelligence. 2022.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *TACL*, 2023.

Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In *Proceedings of the 9th Annual MIT Sloan Sports Analytics Conference*, 2015.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *UIST*, 2023.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6(9): 1257–1267, 2022.

Dean A Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *TMLR*, 2022.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Karun Singh. Introducing expected threat (xT). `https://karun.in/blog/expected-threat.html`, 2019.

William Spearman. Beyond expected goals. In *MIT Sloan Sports Analytics Conference*, 2018.

StatsBomb. StatsBomb open data. https://github.com/statsbomb/open-data, 2023.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *CVPR*, 2024.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, Yuan fang Wang, Weining Shen, and Hanjie Chen. SportU: A comprehensive sports understanding benchmark for multimodal large language models. In *ICLR*, 2025.

Haotian Xia, Haonan Ge, Junbo Zou, Hyun Woo Choi, Xuebin Zhang, Danny Suradja, Botao Rui, Ethan Tran, Wendy Jin, Zhen Ye, Xiyang Lin, Christopher Lai, Shengjie Zhang, Junwen Miao, Shichao Chen, Rhys Tracy, Vicente Ordonez, Weining Shen, and Hanjie Chen. SportR: A benchmark for multimodal large language model reasoning in sports. In *ICLR*, 2026.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021.

Zelai Xu, Zhexuan Xu, Xiangmin Yi, Huining Yuan, Mo Guang, Kaiwen Long, Xinlei Chen, Yi Wu, Chao Yu, and Yu Wang. VS-Bench: Evaluating VLMs for strategic abilities in multi-agent environments. In *CVPR*, 2026.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: CoLlision Events for Video REpresentation and Reasoning. In *ICLR*, 2020.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

A PROMPTS

A.1 DECISION PROMPT

The main benchmark task (Sec. 2). The run-up is presented as a sequence of still frames sampled from the seconds before the decision, one every 0.5 s; `{n}` is the number of run-up frames, `{attack}` the attacking direction, and `{teammates}` the letters that can legally be passed to.

Decision prompt (full stimulus: run-up frames + freeze-frame + map)

```
You are an elite football (soccer) analyst.  Watch this moment and
decide what the player on the ball should do next.  Judge it on what
is visible at the moment the ball is played, not on hindsight.

You are shown one moment from a match, three ways:

1.  RUN-UP - {n} still frames from the seconds leading up to the
moment, one every {interval} seconds of real time, in chronological
order.  Read them as a sequence:  they show who is moving where, and
how fast.
2.  FINAL FRAME - the broadcast frame at the exact instant the
decision is made.  This is the decision point.
3.  PLAYER MAP - a top-down view of that same instant.  Every player
on camera is a lettered circle.  Blue = the on-ball team (T). Red =
the opponents (O). The gold star is the player on the ball.  The arrow
shows which way the on-ball team is attacking.

THE DECISION
The player on the ball is the gold star.  {attack}

What should he do RIGHT NOW? Your options are:
- PASS to one of his lettered teammates (the blue circles on the map).
The teammates you can pass to are:  {teammates}.
- SHOOT.

Reason it through first:  read the shape of the defence, who is
actually free, who is covered or offside, where the space is.
Ask whether a shot from here beats the best available pass.  Say
concretely why your choice beats the obvious alternatives.

Then end your reply with this line and nothing after it:

ANSWER: <letter>

where <letter> is one teammate letter (e.g.  ANSWER: E) or the word
SHOOT.
```

A.2 LABELED BROADCAST-FRAME PROMPT

The ablation of Appendix E. The top-down player map is removed; instead the model sees the run-up frames and the real broadcast frame with every player tagged by a capital letter and the ball-carrier marked. `{n}` is the number of run-up frames and `{teammates}` the letters that can legally be passed to.

Labeled broadcast-frame prompt (run-up frames + labeled final frame, no map)

```
You are an elite football (soccer) analyst.  Watch this moment and
decide what the player on the ball should do next.  Judge it on what
is visible at the moment the ball is played, not on hindsight.

You are shown one moment from a match, two ways:

1.  RUN-UP - {n} still frames from the seconds leading up to the
moment, one every {interval} seconds of real time, in chronological
order.  Read them as a sequence:  they show who is moving where, and
how fast.
2.  LABELED FRAME - the broadcast frame at the exact instant the
```

```
decision is made.  Every player is tagged with a capital letter,
and the player on the ball is marked with a gold dot.  This is the
decision point.

THE DECISION
The player on the ball is the one with the gold dot.

What should he do RIGHT NOW? Your options are:
- PASS to one of his lettered TEAMMATES. The teammates you can pass to
are:  {teammates}.  Only these letters are legal answers.
- SHOOT.

Reason it through first:  read the shape of the defence, who is
actually free, who is covered or offside, where the space is.
Ask whether a shot from here beats the best available pass.  Say
concretely why your choice beats the obvious alternatives.

Then end your reply with this line and nothing after it:

ANSWER: <letter>

where <letter> is one teammate letter (e.g.  ANSWER: E) or the word
SHOOT.
```

A.3 BELIEF-ELICITATION PROMPTS

These are the prompts used to query the VLM for their estimates on the probabilities and expected values of each available action in the event (Sec. 3.3). Each of the two components are elicited independently, share the system message and context block below, and the model returns a ranked JSON list each time.

System message (decision and belief elicitation)

```
You are an elite football (soccer) tactical analyst.  You judge
decisions the way a possession-value model does:  the best action
is the one that most increases the chance your team scores in the
near future while not increasing the chance you concede.  Judge the
decision on what is visible at the moment the ball is played, not on
hindsight.
```

Shared context for the two below belief queries

```
You are shown one moment from a football match:  a FREEZE-FRAME at the
moment the on-ball player controls the ball, and a top-down PITCH MAP
of that same instant (gold star = the player on the ball, light blue =
his team T, navy = opponents O).

Here is the map as a table.  Coordinates are in metres on a 105 x 68
pitch.

{roster}

The on-ball player is marked by the gold star, at ({actor_x},
{actor_y}).  {attack_sentence}

The player's legal actions are exactly:  {actions}.
(Each letter = a pass to that teammate; SHOOT = shoot at goal.)
```

Query 1 — probability of success

```
For each legal action, estimate the PROBABILITY IT SUCCEEDS - a pass
"succeeds" if it reaches the intended teammate; a shot "succeeds"
if it scores.  Rank the actions from most likely to least likely to
succeed, and give the probability (0-1) for each.
Output ONLY JSON, best first:  [{"action":"A","prob":0.90}, ...]
```

Query 2 — expected value

```
For each legal action, estimate its EXPECTED VALUE - how much it
raises your team's chance of scoring, accounting for BOTH how likely
it is to work AND how good the result would be.  Rank the actions from
best to worst expected value, and give the expected value (0-1) for
each.
Output ONLY JSON, best first:  [{"action":"A","ev":0.30}, ...]
```

A.4 RISK-STEER PROMPTS

The risk-steering experiment (Sec. 3.4) replaces only the deliberation paragraph of the decision prompt (App. A.1, beginning "Reason it through first") with one of the two alternatives below. The rest of the stimulus, task, and answer format are unchanged, and NEUTRAL is the original wording.

SEEK deliberation

```
Think it through first:  read the shape of the defence, who is free,
who is offside or covered, and above all where the greatest attacking
upside is.  A safe ball is not more valuable for being secure.
Football is won by the actions that break the game open, so favour
the ambitious ball -- the penetrating pass, the runner in behind, the
shot -- even when it is less likely to come off, as long as its payoff
when it succeeds justifies the chance of losing possession.  Do not
settle for a safe, low-value option just because it is secure.  Be
concrete about why the option you pick is worth more than the obvious
safe alternatives.
```

AVOID deliberation

```
Think it through first:  read the shape of the defence, who is
free, who is offside or covered, and above all which option most
reliably keeps the ball.  A risky ball is not more valuable for being
ambitious.  Football is lost by the cheap giveaway, so favour the
secure ball -- the high-percentage pass, the simple lay-off, the
retention of possession -- even when its upside is modest, and only
take on a risky action when its payoff justifies the chance of losing
possession.  Do not force a low-percentage pass or a speculative shot
just because it is ambitious.  Be concrete about why the option you
pick is worth more than the obvious ambitious alternatives.
```

B MODEL-WISE BELIEF BREAKDOWN

Table 4 gives the per-model recall behind Fig. 7 at the top-33% cutoff. Table 5 gives the per-model probability–value correlation behind Fig. 6b.

Table 4: Per-model recall of the best action at the top-33% cutoff. Success-probability recall exceeds expected-value recall for most models.

model	success-prob. recall	expected-value recall
GPT 5.6 Sol	79.7%	55.8%
Gemini 3.5 Flash	84.9%	50.2%
Claude Opus 4.8	72.1%	58.0%
Qwen3-VL-8B	55.8%	34.3%
Llama-4-Maverick	44.1%	59.2%

Table 5: Per-model Spearman correlation between the model's success-probability ranking and its expected-value ranking. Every model is positive; the value model is negative.

	ρ(prob, expected value)
value model (ground truth)	-0.076
GPT 5.6 Sol	$+0.366$
Gemini 3.5 Flash	$+0.371$
Claude Opus 4.8	$+0.296$
Qwen3-VL-8B	$+0.521$
Llama-4-Maverick	$+0.349$

Table 6: The belief results recomputed against Expected Threat (xT) in place of VAEP, pooled across both tournaments. Entries are recall@k (top 33%): how often the model's expected-value ranking contains the value-optimal action, under each value model. The values are similar (and if anything lower) under xT, and the negative probability–value correlation reported in the text holds under both value models.

	EV recall@k	
model	VAEP	xT
GPT 5.6 Sol	56%	38%
Gemini 3.5 Flash	50%	32%
Claude Opus 4.8	58%	34%

C Value Model Robustness

The belief analysis in Sec. 3.3 scores each model's estimates against VAEP. To check that these results do not depend on that particular value model, we repeat the analysis using Expected Threat (xT) (Singh, 2019), another possession-value model that assigns each location on the pitch a scoring probability from a move-and-shoot transition matrix. We fit xT on the same training tournaments used for VAEP and re-value every candidate action under it. A pass is worth the change in xT from the ball's location to the receiver's, a shot keeps its expected-goals value, and each model's elicited beliefs are used unchanged. With the xT value model, probability of success and expected value are also nearly uncorrelated ($\rho = -0.15$), just as with VAEP. The value-recall gap persists under xT (Table 6): the models' expected-value ranking recovers the xT-optimal action on only 32–38% of events, if anything lower than under VAEP. The finding in Section 3.3 is therefore robust across different value models.

D The risk-aversion result is robust to value-model noise

The risk-aversion analysis (Sec. 3.2) locates each chosen action within its event's variance distribution (its risk percentile). A potential concern is that this ordering could be sensitive to noise in the estimated action values, or unstable on events where $a^\star$ only marginally exceeds the next-best action.

We alleviate this concern in two ways. First, we restrict the analysis to events in which $a^\star$ exceeds the second-best action by a clear value margin (we use a margin of 0.019, the median standard deviation among all possible actions across all events). Removing these near-tie events only sharpens the pattern: the optimal action's risk percentile rises from 76 to 84 and the real player's from 60 to 70, while every VLM stays at or below the player (between the 46th and 64th percentile). VLMs continue to select markedly lower-risk actions than both the optimum and the human.

Second, we perturb every action's outcome standard deviation with Gaussian noise scaled by the median action standard deviation and recompute the risk percentiles (Table 7). The ordering is essentially invariant: the optimum stays near the 71st–76th percentile and every VLM near the middle (46th–53rd), at or below the random baseline of 50. Noise in the value model does not affect

the conclusion that VLMs take on middling risk while the optimum and the professional players accept more variance.

Table 7: Noise perturbation of the value model. Each action's outcome standard deviation is perturbed by $\mathcal{N}(0, \text{scale} \cdot \sigma_0)$ with σ_0 the median action standard deviation, and the risk percentiles are recomputed. Entries are each strategy's mean within-event risk (variance) percentile (0=safest, 100=riskiest); the optimum and players stay well above the VLMs, which sit at the random baseline, at every noise level.

noise	$a^\star$	real player	GPT	Gemini	Claude	Qwen	Llama	random
none	76	60	50	53	46	53	48	50
1×sd	72	59	51	52	47	52	48	50
2×sd	71	59	53	52	47	52	48	50

E LABELED-FRAME CONTROL

To test whether performance is limited by difficulty parsing the visual input, we repeat the evaluation on the 2022 men's World Cup subset with additional perceptual scaffolding. Models receive the same run-up video frames, but the final broadcast frame has every player tagged with a capital letter and the ball carrier explicitly marked (Figure 3). We omit the player map used in the default condition, allowing the labeled broadcast frame itself to provide a simpler representation of the decision context (prompt in Appendix A.2). Despite this additional visual guidance, accuracy changes only modestly across models, suggesting that the observed decision-making failures are not primarily explained by difficulty identifying or localizing players.

Table 8: Decision accuracy with player-labeled versus default visual inputs.

model	labeled accuracy	default accuracy	Δ
GPT 5.6 Sol	33.5%	32.3%	+1.1
Gemini 3.5 Flash	26.7%	27.4%	−0.8
Claude Opus 4.8	20.4%	19.7%	+0.8
Qwen3-VL-8B	23.3%	23.1%	+0.2
Llama-4-Maverick	21.9%	24.6%	−2.6