SVI-Bench

A Dynamic Microworld for Strategic Video Intelligence

Yulu Pan[1,⋆] Han Yi[1,⋆] Seongsu Ha[1,⋆] Md Mohaiminul Islam[1,⋆]
Benjamin Zhang[1] Lorenzo Torresani[2] Gedas Bertasius[1]

[1]UNC Chapel Hill [2]Northeastern University

⋆Equal contribution

Abstract

True video intelligence demands more than recognizing what is visible: it requires reasoning about *why* events unfold, predicting *what would change* under different conditions, and deciding *what to do next*. We refer to this full progression—from perception through causal reasoning and simulation to strategic planning—as Strategic Video Intelligence (SVI). No existing benchmark evaluates this capability stack: in-the-wild videos lack verifiable ground truth for causal and strategic questions, while synthetic environments sacrifice the complexity of real multi-agent systems. To bridge this gap, we introduce SVI-Bench, a large-scale benchmark that leverages team sports as a *dynamic microworld*, a domain that uniquely combines the complexity of real-world multi-agent interaction (10–22 agents executing coordinated decisions under adversarial pressure) with the verifiability of explicit rules and definitive outcomes. SVI-Bench comprises ∼**35K hours** of broadcast video, ∼**15M** annotated actions, ∼**15K hours** of expert commentary, ∼**23K** game reports, and ∼**103K** structured statistical records across basketball, soccer, and hockey, all constructed via a data engine that transforms raw game data into a dense, cross-referenced corpus. We organize evaluation into **9 tasks** spanning a progressive four-pillar hierarchy: *Dynamic Scene Understanding*, *Causal Reasoning*, *Strategic Simulation*, and *Agentic Synthesis*. Evaluating strong multimodal and agentic baselines, we find a *capability cliff*: models perform competently on perceptual tasks (achieving ∼74% on fine-grained action QA) but degrade sharply at each successive cognitive level. Agentic tasks prove hardest of all: the strongest model achieves only ∼5% accuracy when required to autonomously gather and integrate evidence across a corpus of 1.8M clips. This consistent degradation across pillars reveals that current systems can see dynamic multi-agent worlds far better than they can reason, simulate, or plan within them. We release the full benchmark to catalyze progress toward the next generation of AI systems capable of strategic intelligence in complex, dynamic multi-agent environments.

Code github.com/texaser/svi-bench

Data huggingface.co/datasets/mvp-group/svi-bench

Website svi-bench.github.io

Contents

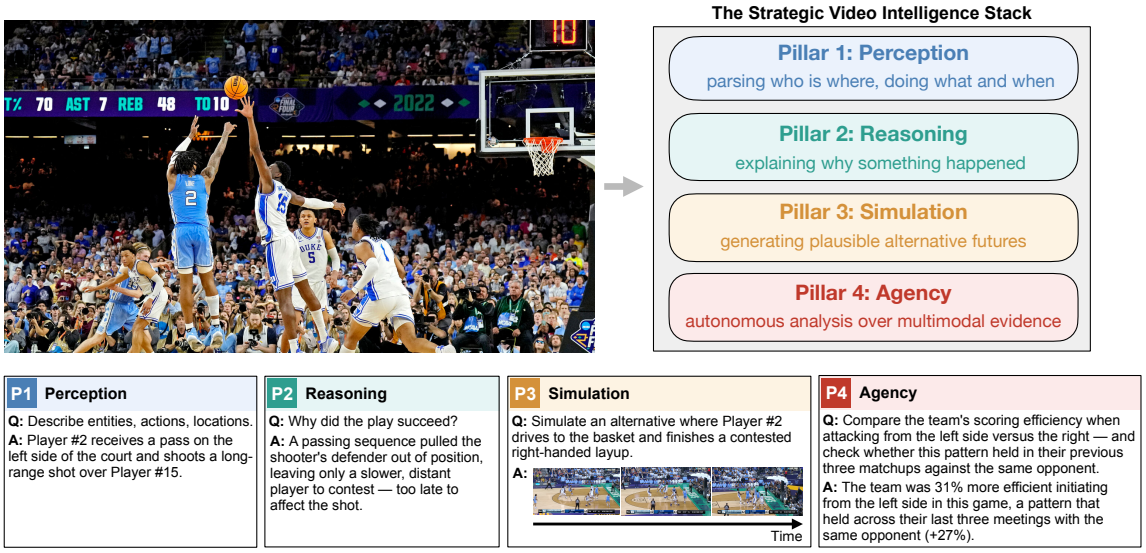


Figure 1. Overview of SVI-Bench, illustrated through a single play from the 2022 NCAA Final Four. SVI-Bench is the first large-scale video benchmark that evaluates the full SVI stack: Perception (describing what is happening), Reasoning (explaining why), Simulation (generating plausible alternatives), and Agency (autonomous analysis over multimodal evidence).

1 Introduction

With forty seconds remaining in Game 6 of the 1998 NBA Finals, Michael Jordan receives the ball trailing by one. Jordan perceives the defensive formation around him, infers that an aggressive first step will force his defender to overextend, simulates how that reaction will open a path that did not exist before, and selects the optimal action: a jump shot that wins the championship. Jordan does not merely react to what he sees; he reasons about its causes, anticipates its consequences, and synthesizes it all into strategic action.

This kind of intelligence—the ability to move from *seeing* to *reasoning* to *deciding*—remains out of reach for current AI systems. A state-of-the-art video-language model, given the same footage, can describe the scene: *a player receives the ball, drives to the basket, releases a shot*. But it cannot explain *why* the defense collapsed (a well-timed screen created a mismatch), predict *what would have happened* had the guard driven left instead of right (a help defender rotating too late to recover), or recommend the *optimal response* given the defensive configuration (attack the weak-side gap before the rotation). This gap extends well beyond sports: surgical teams, first responders, autonomous vehicles, and military units all require the same ability to reason about *why* events unfold, simulate *what-if* alternatives, and decide *what to do next*.

We argue that these abilities are not independent skills but facets of an integrated capability that we call **Strategic Video Intelligence (SVI)**: a progressive cognitive stack (Figure 1) spanning *perception* (parsing who is doing what), *causal reasoning* (explaining why actions lead to outcomes), *simulation* (generating futures and goal-directed strate-

gies), and *agentic synthesis* (integrating multimodal evidence into expert analysis).

Despite sitting at the intersection of three active frontiers (reasoning VLMs, video world models, and agentic intelligence), progress on SVI has been limited by the absence of suitable benchmarks. Existing video understanding benchmarks [21, 50, 85, 104] cover perception and temporal reasoning, but none evaluates the full stack to strategic agency (Table 1). Synthetic environments [4, 95] provide ground truth for causal questions but involve single objects in simplified worlds, while real-world benchmarks [21, 50] offer visual richness but no objective ground truth for causal or strategic reasoning.

Team sports offer a natural *dynamic microworld* that bridges this gap. Sports feature complex multi-agent dynamics (10 to 22 players executing coordinated decisions under adversarial pressure) while offering properties that make strategic reasoning measurable. First, *long-horizon causality*: early tactical setups (a screen, a formation shift) produce delayed outcomes (a scoring opportunity, a turnover) seconds or minutes later, requiring models to trace causal chains across extended temporal windows. Second, *unambiguous success signals*: scores, turnovers, and wins provide clear, discrete outcome labels. Third, *layered verifiability*: perceptual questions are verified against timestamped event logs; causal and explanatory questions are evaluated against expert judgment obtained via transcribed speech commentary from broadcast video; and strategic questions are grounded in outcome-conditioned evaluation, measuring whether a model's recommendations align with actions that led to favorable outcomes in similar situations.

With this motivation, we introduce SVI-Bench, the first large-scale benchmark designed to evaluate the full SVI stack, from perception through reasoning and simulation to agency, in real-world multi-agent video. SVI-Bench consists of ∼35K hours of broadcast video, ∼15M annotated actions, ∼15K hours of expert commentary, ∼23K game reports, and ∼103K statistical records across basketball, soccer, and hockey (Table 1). These five sources are integrated via a data engine that performs temporal alignment, cross-modal entity resolution, and LLM-powered instance generation with automated verification (§3). We organize evaluation into 9 tasks across a progressive four-pillar hierarchy (§4): (1) *Dynamic Scene Understanding*—parsing multi-agent scenes into structured spatiotemporal representations; (2) *Causal Reasoning*—explaining why events unfold and predicting outcomes; (3) *Strategic Simulation*—generating counterfactual futures and goal-directed strategies; and (4) *Agentic Synthesis*—autonomously gathering and integrating multimodal evidence to produce expert-level analysis.

Evaluating strong multimodal and agentic baselines, we find that models perform competently on perceptual tasks but decline sharply on causal reasoning and collapse on agentic tasks requiring autonomous evidence gathering. Our contributions are:

1. **The Strategic Video Intelligence framework**, formalizing video understanding as a progressive stack from perception through causal reasoning and simulation to agency.

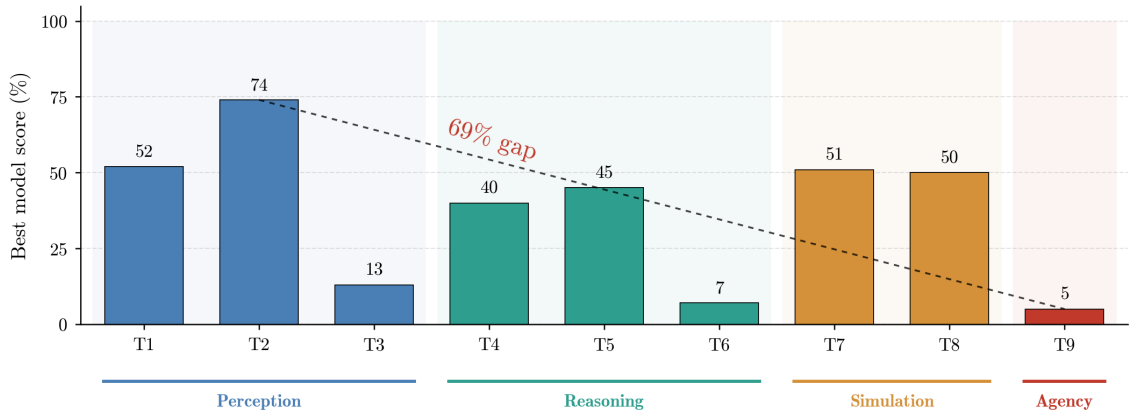


Figure 2. The performance cliff. Per-task best-model scores, normalized to a 0–100% range. From the strongest Perception result (T2, ∼74%) to Agentic Synthesis (T9, ∼5%), performance drops by 69 points, with Reasoning and Simulation falling in between.

2. **A data engine** that aligns five modalities via temporal alignment, cross-modal entity resolution, LLM-assisted instance generation, and multi-stage quality control.
3. **SVI-Bench**, a large-scale benchmark spanning the full perception-to-agency stack, with 9 tasks across the four pillars and training splits for 7 of them.
4. **Empirical findings** that localize where in the cognitive stack performance degrades, revealing a consistent capability cliff across pillars.

2 Related Work

Video Understanding Benchmarks. Early benchmarks focused on action recognition (UCF-101 [72], Kinetics [11, 34]) and temporal localization (ActivityNet [10], THUMOS [32]). Video QA benchmarks (MSRVTT-QA [89], ActivityNet-QA [97], NExT-QA [88], EgoSchema [50]) introduced language-grounded reasoning but focus on short clips with predominantly perceptual questions. Long-video benchmarks (Video-MME [21], MLVU [104], LongVideoBench [85]) extend temporal scope but lack verifiable causal ground truth. Egocentric planning benchmarks (Ego-Plan-Bench [16]) and temporal reasoning (TempCompass [46], Time-Blind [37]) benchmarks target single-agent or low-complexity settings. Reasoning-oriented benchmarks (MINERVA [54], Video-Holmes [17]) move toward higher-order reasoning but evaluate neither generative simulation nor corpus-scale agency. SVI-Bench is the first benchmark to combine real-world multi-agent complexity with verifiable evaluation spanning the full perception-to-agency stack.

Causal and Counterfactual Reasoning in Video. Synthetic environments provide verifiable ground truth for causal and physical reasoning. CLEVRER [95] evaluates coun-

Table 1. Comparison with existing video benchmarks. SVI-Bench is the first to combine large-scale, real-world multi-agent video with cross-referenced multimodal data and evaluation spanning perception through strategic agency. Counts are reported where available.

					Modalities			Evaluation Tasks			
	Benchmark	Video Hours	Duration Range	Annotated Actions	Expert Commentary	Long-Form Reports	Structured Metadata	Perception	Reasoning	Simulation	Agency
General	Kinetics-700	1.9K	10s	650K	✗	✗	✗	✓	✗	✗	✗
	ActivityNet	849	5–10 min	30K	✗	✗	✗	✓	✗	✗	✗
	Video-MME	254	11s–1h	–	✗	✗	✗	✓	∼	✗	✗
	Ego-Exo4D	1.4K	1–42 min	432K	6K hrs	✗	✗	✓	✗	✗	✗
	Ego4D-HCap	3.7K	5s–2h	3.8M	✗	8K	✗	✓	✗	✗	✗
Reasoning	Causal-VidQA	28	10s	–	✗	✗	✗	✓	∼	✗	✗
	MINERVA	∼150	2–100 min	–	✗	✗	✗	✓	✓	✗	∼
	Video-Holmes	∼14	1–5 min	–	✗	✗	✗	✓	✓	✗	∼
Synth.	CLEVRER	5	5s	–	✗	✗	✗	✓	✓	∼	✗
	PHYRE	–	5s	–	✗	✗	✗	∼	✓	∼	✗
Sports	SoccerNet	500	90 min	300K	✗	✗	500	✓	✗	✗	✗
	SportsMOT	14	14.4–33.8s	–	✗	✗	✗	✓	✗	✗	✗
	BASKET	4.5K	8–10 min	–	✗	✗	✗	✓	✗	✗	✗
	SVI-Bench (Ours)	**35K**	**10s–2.5h**	**15M**	**15K hrs**	**23K**	**103K**	✓	✓	✓	✓

terfactual reasoning under rigid-body collisions. PHYRE [4] and CoPhy [5] test physical intuition with simple block-and-ramp dynamics. Physion [6] probes intuitive physics across a broader range of object interactions. These benchmarks involve single objects in simplified worlds without multi-agent behavioral complexity. Real-video causal QA (Causal-VidQA [38], MECD [15]) attempts causal reasoning in natural settings but relies on subjective annotations without objectively verifiable outcomes. SVI-Bench brings causal evaluation to real-world multi-agent video where explicit rules, definitive outcomes, and convergent expert analysis provide layered verification.

Sports Video Analysis. Prior work in sports video targets action detection (SoccerNet [18, 20, 24]), tracking (SportsMOT [19]), trajectory prediction (Baller2Vec [1]), and fine-grained skill recognition (FineSports [90], BASKET [59]). SPORTU [87] and SportR [86] extend evaluation to rule comprehension but stop short of counterfactual reasoning, simulation, or agentic inference. NBA tracking datasets [12, 47] provide player trajectory data but lack video or language annotations. SVI-Bench is the first to unify multi-sport video with evaluation spanning the full cognitive stack from perception to agency.

Multimodal LLMs and Agentic AI. Multimodal LLMs [23, 43, 44, 57, 100] achieve strong video description but exhibit brittle temporal reasoning and causal grounding [21, 50]. Agentic AI systems (ReAct [94], Toolformer [68], Voyager [79]) combine LLMs with tool

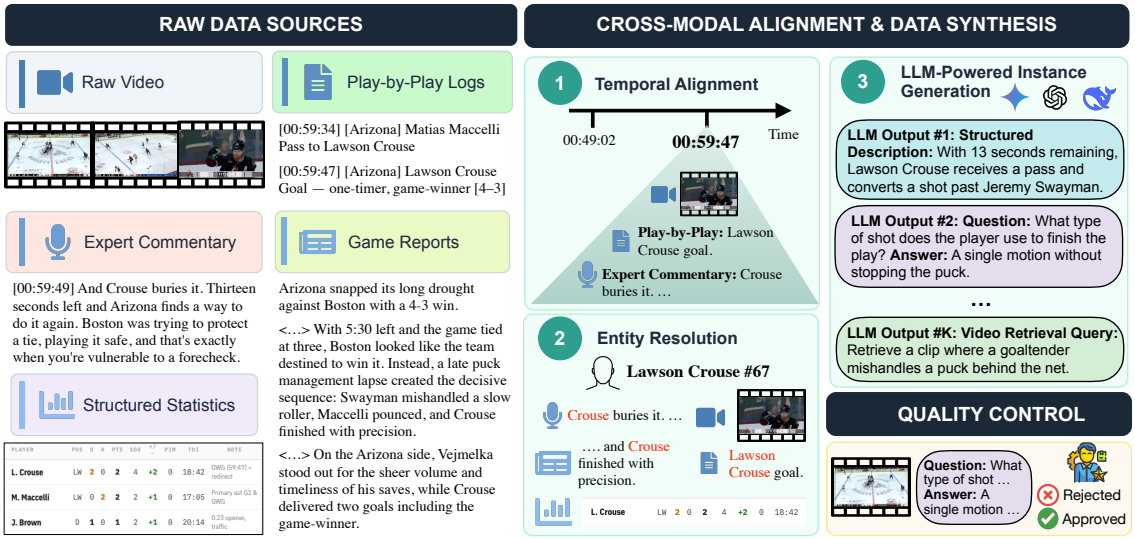


Figure 3. The SVI-Bench data engine. Five raw sources are transformed into a cross-referenced corpus via (1) temporal alignment, (2) cross-modal entity resolution, (3) LLM-assisted instance generation, and (4) automatic and human quality control.

use for complex problem solving. Recent work extends these capabilities to video via temporal search [58], tool-augmented reasoning [81, 93], and RL-based evidence gathering [63, 101]. SVI-Bench's Pillar 4 introduces the first such evaluation at corpus scale, requiring autonomous navigation of large-scale multimodal video data.

Strategic Game AI and World Models. Superhuman game AI [69, 70, 77] operates in fully observable, discrete-state environments, while SVI-Bench targets continuous, partially observable real-world multi-agent video. Closer to our setting, world models have advanced from model-based RL (Dreamer [27–29]) to diffusion-based (DIAMOND [2]), transformer-based (TWM [65], IRIS [52]), and foundation world models [9, 31, 56]. Trajectory forecasting in sports [1, 26, 67, 73] focuses on short-term prediction without strategic reasoning or causal structure. SVI-Bench provides the first benchmark for evaluating strategic planning capabilities grounded in real-world multi-agent video.

3 Data Engine

A core contribution of SVI-Bench is a data engine that transforms raw game data into a dense, cross-referenced corpus suitable for evaluating the full SVI stack. The engine is designed around two principles: (i) primary evidence is human- or league-derived (play-by-play logs, official statistics, journalist reports, broadcast commentary), and (ii) LLMs are used to *scale* task-instance generation from these grounded sources, with manual human verification on a representative subset of every task. The combination of human-

	Basketball	Hockey	Soccer	Total
Leagues	17	44	3	64
Games	25K	35K	4K	64K
Video hours	18.5K	12.5K	4.5K	35.5K
Annotated actions	8.9M	4.3M	1.8M	15M
Commentary (hrs)	9K	5.2K	0.8K	15K
Game reports	13K	8.5K	1.6K	23.1K
Stat. records	48.5K	35.4K	19K	102.9K

Table 2. Per-sport corpus breakdown across the modalities and sports. The corpus covers 64 leagues, 35.5K hours of video, 15M annotated actions, 15K hours of commentary, 23.1K game reports, and 102.9K statistical records across basketball, hockey, and soccer. The benchmark is organized around game-level alignments across videos, timestamped actions, commentaries, reports, and statistical records, with consistent player references across different modalities.

derived primary annotations and LLM-assisted instance generation produces supervision at a scale and density unmatched by prior sports video resources.

3.1 Data Sources and Scale

SVI-Bench spans three professional team sports selected for their complementary multi-agent properties in team size, pacing, spatial scale, camera dynamics, and strategic structure: basketball (10 players, compact court, frequent transitions), soccer (22 players, large pitch, continuous fluid dynamics), and hockey (rapid line changes, fast-panning camera). The corpus spans 64 leagues over seven years (2018–2025).

The corpus comprises five synergistic modalities, all temporally aligned and cross-referenced through shared game and player identifiers (Figure 3, left; Table 2): **broadcast video** ($\sim$35K hours of professional footage); **play-by-play logs** ($\sim$15M timestamped event records from official league data feeds, with player identities and spatial coordinates); **expert commentary** ($\sim$15K hours of broadcast commentary and analyst narration collected via Whisper ASR [62]); **game reports** ($\sim$23K post-game journalist recaps and editorial analyses); and **box-score statistics** ($\sim$103K records of player and team performance metrics). Together, these modalities give every game in the corpus a verifiable event timeline, an expert-language interpretation of those events, and structured numerical context. This combination is what enables tasks across all four pillars to be grounded in objective evidence.

3.2 Data Construction Pipeline

The engine (Figure 3) transforms these sources into a unified corpus through four stages.

(1) Temporal alignment. Play-by-play logs provide the primary temporal reference via game-clock timestamps; commentary transcripts and game reports are aligned using timestamp matching and textual cues, producing temporally grounded segments that link every video clip to its corresponding events, commentary, and statistical context.

(2) Cross-modal entity resolution. References to the same player, team, or event are linked across modalities and organized into identity graphs capturing relationships (teammate, opponent) and attributes (position, statistics, role). When a journalist describes a player's late-game three-pointer, the corresponding play-by-play event, commentary segment, statistical record, and video clip are all tied to a single canonical identity.

(3) LLM-assisted instance generation. Using the assembled multimodal context, LLMs synthesize task instances guided by task-aware prompt templates, producing question–answer pairs, plausible distractors, difficulty-calibrated instances, and free-form annotations such as dense captions and narrative summaries.

(4) Quality control. All instances pass through three filtering stages: automatic consistency checks against event logs, task-specific filters, and human expert review by domain-knowledgeable annotators on a stratified subset spanning all sports, pillars, and difficulty levels. Per-task filtering statistics and human-review protocols are in Appendix A.2.

4 Benchmark Tasks

The SVI-Bench comprises 9 tasks organized into a four-pillar hierarchy (Table 3), from perception through causal reasoning and simulation to agency. Construction details, prompts, per-task statistics, and complete results are in Appendices B–E. Below, we present each pillar and its tasks.

4.1 Pillar 1: Dynamic Scene Understanding (T1–T3)

Strategic reasoning begins with perception, parsing a dense, fast-moving multi-agent scene into spatiotemporal primitives: which agents are present, where they are, what they are doing, and how these compose into higher-order events. The three tasks in this pillar evaluate this foundation using short clips, establishing the perceptual floor upon which higher-level reasoning depends (Figure 4).

Table 3. Summary of SVI-Bench benchmark tasks. Nine tasks across four cognitive pillars (perception, reasoning, simulation, agency) covering basketball (B), hockey (H), and soccer (S). # is total task instances. Train indicates whether a training split is provided.

ID	Task	Pillar	Context	Sports	#	Train	Format	Primary Metric
T1	Structured Play Description	Perception	10s	B,H,S	1.5M	✓	Open	Avg. Score
T2	Fine-Grained Action QA	Perception	10s	B,H,S	1.5M	✓	MCQ	Accuracy
T3	Compositional Video Retrieval	Perception	10s	B,H,S	306K	✓	Retrieval	R@1
T4	Strategic Reasoning QA	Reasoning	55–150 min	B,H,S	1K	✗	Open	Avg. Score
T5	Outcome Forecasting	Reasoning	3–15 min	B,H,S	114K	✓	MCQ	Accuracy
T6	Long-form Narrative Synthesis	Reasoning	55–150 min	B,H,S	19K	✓	Open	Saliency
T7	Motion-Conditioned Generation	Simulation	5–10s	B,S	290K	✓	Generation	Video mIoU
T8	Goal-Conditioned Action Gen.	Simulation	5–10s	B	74K	✓	Generation	Goal Acc.
T9	Cross-Corpus Agentic Reasoning	Agency	Multi Source	B,H,S	1K	✗	Open	Accuracy

T1: Structured Play Description

Task formulation. Given a 10-second video clip, the model must generate a dense, structured caption that describes the actions, player identities, spatial positioning, and game context. Unlike standard captioning benchmarks [35, 82, 91, 105] describing a single salient action per clip, T1 requires simultaneous precision across 10+ coordinated agents, parallel sub-actions, and game-state context.

Data and construction. We extract 10-second video segments aligned to play-by-play events and synthesize a structured caption for each segment by composing the timestamped actions, player identities, and game state, then polishing the result with GPT-4o-mini for fluency. This yields 1.5M video–caption pairs across the three sports (500K per sport). For practical training and evaluation we use a 280K/3K/3K train/val/test subset. Full construction details, prompts, and statistics are in Appendix B.1.

Evaluation. We use an LLM-as-a-judge protocol that assigns Likert scores from 0–5 along six axes: action accuracy, identity accuracy, causality/outcome, spatial understanding, temporal understanding, and contextual details. The mean across the six axes is our primary metric, with GPT-5.2 as the primary judge. To verify judge reliability, three annotators independently scored 60 randomly sampled captions. The human–judge mean absolute error is 0.40 on the 0–5 scale (8% of the scoring range), within typical inter-annotator variation.

Baselines. We evaluate frontier proprietary models (GPT-5.2, Gemini-3-Flash) and open-source video–language models (LLaVA-Video-7B, Qwen3-VL-32B), all in a few-shot setting with two in-context examples. We also evaluate a fine-tuned LLaVA-Video-7B trained jointly on T1 and T2 (580K samples total). Full training details are in Appendix B.1.

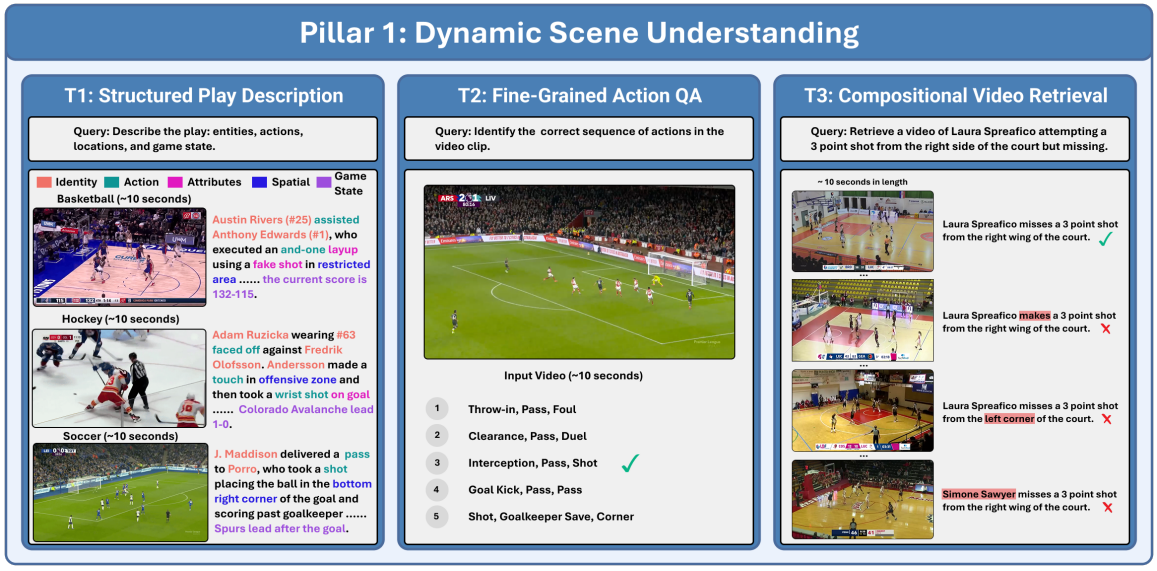


Figure 4. Overview of Pillar 1: Dynamic Scene Understanding. This pillar evaluates foundational perceptual capabilities through three tasks: structured play description (T1), fine-grained action QA (T2), and compositional video retrieval (T3).

Main results. Table 4 reports overall and per-sport average scores. No model exceeds 3/5. The frontier proprietary models reach 1.61 (GPT-5.2) and 1.67 (Gemini-3-Flash). A fine-tuned LLaVA-Video-7B reaches 2.17 overall, an improvement of +1.28 over its zero-shot counterpart and outperforming both proprietary baselines despite being an order of magnitude smaller. Performance is lowest on hockey (1.77) and soccer (1.81) and highest on basketball (2.92).

Per-axis breakdown. Table 5 presents the fine-tuned model's performance across different axes. The model performs best on action accuracy, spatial understanding, and temporal understanding, achieving average scores of 2.14, 2.74, and 2.72 across the three sports, respectively. In contrast, identity recognition remains the weakest axis across all sports (1.46 in basketball, 1.20 in soccer, and 0.66 in hockey). Causality and outcome reasoning is also challenging, particularly in soccer (1.58) and hockey (1.39). Overall, the model is more effective at understanding what is happening and where, but struggles to identify who is involved and why events occur.

Judge robustness. Re-scoring with four independent LLM judges (GPT-5.2, Gemini-3-Flash, Qwen3-235B, DeepSeek-V3.2) preserves the model ranking. Pairwise Spearman rank correlations between judges fall in 0.66–0.73 (Appendix Table 16), and the GPT-5.2 judge scores GPT-generated captions 0.21 points lower than the mean of the non-GPT judges, ruling out self-preference.

Qualitative example. Figure 5 shows a representative basketball clip. Action and spatial axes score 5/5 while identity is 1/5 as the fine-tuned model misattributes the play to the wrong player. Additional examples on soccer and hockey are in Appendix Fig. 25.

Model	Bball.	Soccer	Hockey	Overall
LLaVA-Video-7B (few-shot)	1.12	0.89	0.66	0.89
Qwen3-VL-32B	1.61	1.35	1.07	1.34
GPT-5.2	2.09	1.50	1.23	1.61
Gemini	2.15	1.66	1.21	1.67
LLaVA-Video-7B (FT)	**2.92**	**1.81**	**1.77**	**2.17**

Table 4. T1 structured play description results. Each model's caption is scored by a GPT-5.2 judge on a 0–5 scale, with the best per-sport and overall entries bolded.

Axis	Bball.	Soccer	Hockey	Avg.
Action	3.03	1.84	1.56	2.14
Identity	1.46	1.20	0.66	1.11
Causality	2.50	1.58	1.39	1.82
Spatial	3.55	2.00	2.67	2.74
Temporal	3.37	2.59	2.19	2.72
Contextual	3.62	1.67	2.15	2.48

Table 5. T1 per-axis breakdown. LLaVA-Video-7B-ft scores across the six evaluation axes per sport, plus the cross-sport average (GPT-5.2 judge, 0–5 scale).

T2: Fine-Grained Action QA

Task formulation. Given a 10-second clip, a question, and 5 candidate answers, the model must select the correct one. Unlike prior video QA benchmarks [50, 88, 89, 97] that feature single-agent scenarios with coarse-grained questions, T2 targets multi-agent interactions where correct answers depend on precise details (e.g., which player initiated a screen, the exact pass sequence, or spatial relationships). The task spans 31 question types across three sports, organized into six capability categories: action recognition, temporal ordering, play analysis, spatial reasoning, player identification, and OCR. The full question taxonomy is in Appendix B.2.

Data and construction. Using the same 10-second clips as T1, we generate multiple-choice questions across 31 question types spanning the six capability categories. For each question we sample candidate answers from the play-by-play event vocabulary and balance answer distributions within each question type to reduce label bias. This yields 1.5M question–answer pairs across the three sports (500K per sport). For practical training and evaluation we use a 300K/30K/30K train/val/test subset. The full question taxonomy and construction details are in Appendix B.2.

Evaluation. We use multiple-choice accuracy as the primary metric.

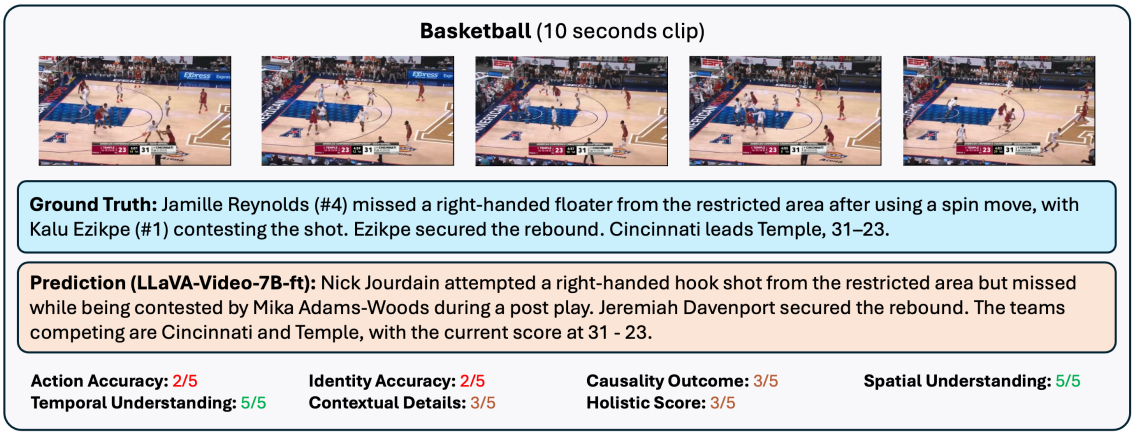


Figure 5. T1 qualitative example (basketball). Given a 10-second clip, the model must produce a dense structured caption covering actions, identities, spatial positions, and game context. Comparing the ground truth (blue) with the fine-tuned LLaVA-Video-7B-ft prediction (red), the model correctly captures the spatial layout (5/5) and temporal sequence (5/5), but misclassifies the action type (2/5) and misattributes both the shooter and the defender (identity 2/5). Per-axis LLM judge scores shown in gray.

Baselines. We evaluate proprietary models (GPT-5.2, Gemini-3-Flash) and open-source video–language models (LLaVA-Video-7B, Qwen3-VL-32B) in a zero-shot setting. We also evaluate a fine-tuned LLaVA-Video-7B trained jointly on T1 and T2.

Main results. Table 6 reports validation accuracy. The fine-tuned LLaVA-Video-7B reaches 73.91% overall—an improvement of +36.90 points over its zero-shot counterpart and +15.16 over the strongest zero-shot baseline (Gemini-3-Flash, 58.75%). The pattern is consistent across all three sports. Human evaluators with five-plus years of sport experience reach 78.33% on basketball, 74.00% on soccer, and 74.00% on hockey. The fine-tuned model lands within 5 points of human accuracy on basketball, matches it on soccer (75.48% vs. 74.00%) and on hockey (74.00% vs. 72.83%).

Per-category breakdown. Table 7 reports per-category accuracy of the fine-tuned model across all three sports. Action recognition, temporal ordering, and OCR are strongest (80–95% accuracy). Play analysis and player identification are weakest (51–67%).

Human comparison. Humans reach 91–100% on temporal ordering, action recognition, and OCR, while the fine-tuned model trails by 8–18 points. On player identification, humans reach only 56.7% because most participants struggle to recognize players from less familiar leagues. On spatial reasoning, the model reaches 73.2% versus 60.0% for humans, likely because spatial questions in T2 (e.g., court zones, attack flanks) use a standardized vocabulary that fine-tuning maps directly onto (Appendix B.2).

Qualitative example. Figure 6 shows a representative soccer failure in player identification, where all three evaluated models fail to identify the player taking the goal kick. The

Model	Bball.	Soccer	Hockey	Overall
LLaVA-Video-7B	35.76	38.72	36.56	37.01
Qwen3-VL-32B	41.88	48.18	41.05	43.70
GPT-5.2	52.93	56.70	49.10	52.91
Gemini-3-Flash	60.70	62.50	53.05	58.75
LLaVA-Video-7B-ft	**73.43**	**75.48**	**72.83**	**73.91**
Human	*78.33*	*74.00*	*74.00*	*75.78*

Table 6. T2 multiple-choice video QA accuracy (%). All baselines are evaluated zero-shot. LLaVA-Video-7B-ft shares the 580K-sample training set with T1. Human numbers come from a study with sport-experienced participants (Appendix B.2).

Category	Basketball	Soccer	Hockey
Action recognition	82.7	84.6	83.3
Temporal ordering	88.8	81.1	92.7
Play analysis	57.5	66.2	51.6
Spatial reasoning	73.2	67.4	87.8
Player identification	58.8	55.1	50.8
OCR	84.3	79.7	89.9
Overall	**73.4**	**75.5**	**72.8**

Table 7. T2 per-category accuracy (%). LLaVA-Video-7B-ft performance across the six question categories (action recognition, temporal ordering, play analysis, spatial reasoning, player identification, and OCR), reported per sport.

pattern is consistent with the low player-identification scores, where every model falls below 60% accuracy (see Appendix B.2).

T3: Compositional Video Retrieval

Task formulation. Given a natural-language query describing a specific composition of visual attributes (entity, dynamics, context, spatiotemporal structure) and a candidate pool of one positive and 5,000 negative videos, the model must rank the candidate videos by semantic similarity with the query, aiming to place the ground-truth video at the top. Queries are organized into three compositional tiers by the number of attributes they specify (1, 2, or 3+). Unlike standard video retrieval [3, 35, 66, 91] where visually diverse candidates allow coarse features to suffice, T3 candidates depict the same sport and differ only in their specific attribute composition.

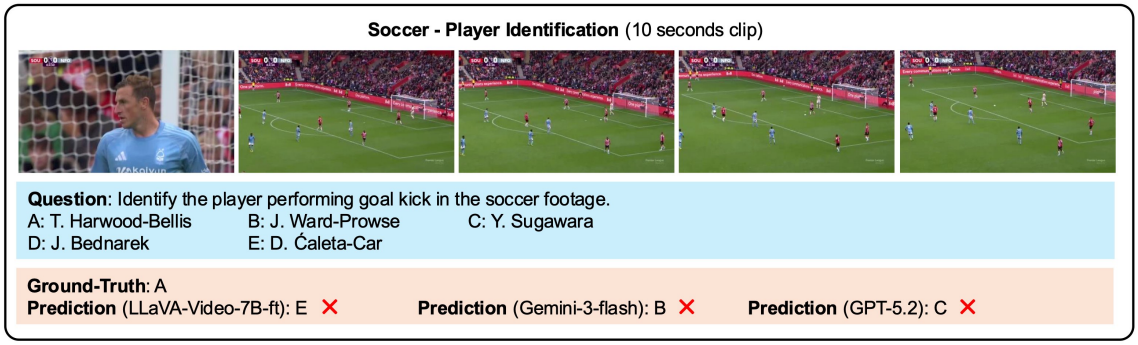


Figure 6. T2 qualitative example (soccer, player identification). Fine-tuned LLaVA-Video-7B, GPT-5.2, and Gemini-3-Flash all fail to identify the player executing the goal kick. Player identification is the weakest T2 category across all evaluated models.

Data and construction. Queries are generated from ground-truth attributes of each video and refined into natural language via LLM paraphrasing, with hard-negative mining to ensure challenging distractors. The benchmark contains approximately 291K training samples and 15K evaluation queries (1K validation and 4K test per sport). Each evaluation query is paired with one positive video and 5,000 negatives. Full construction details, attribute taxonomy, and per-sport statistics are in Appendix B.3.

Evaluation. We report Recall@K for $K \in \{1, 5, 10, 50, 100, \ldots, 500\}$, with Recall@1 as the primary metric. For category- and composition-level analyses, we report R@100 since R@1 values are generally too low to yield meaningful comparisons.

Baselines. We fine-tune InternVideo2-Stage2 1B [83] under two training regimes: *full-caption training*, where each video is paired with a single complete description, and *attribute dropout*, where caption variants at varying levels of specificity are constructed by progressively dropping attributes. Full training details are in Appendix B.3.

Main results. Figure 7 reports Recall@K across all three sports under full-caption training. Basketball, soccer, and hockey achieve R@1 of 2.2%, 3.4%, and 3.5% respectively, and R@100 of 24.7%, 33.9%, and 39.7%. Recall remains below 65% for all sports even at $K{=}500$, confirming that fine-grained compositional retrieval among visually similar clips remains challenging for current video-language models.

Effect of attribute dropout. Attribute dropout outperforms full-caption training on average (R@100: 32.7% $\rightarrow$ 36.5%), with the largest gains in basketball (24.7% $\rightarrow$ 31.8%) and soccer (33.9% $\rightarrow$ 41.4%). Evaluation queries typically specify only a subset of attributes, better matching the dropout training distribution.

Effect of hard negatives. Retrieval performance degrades substantially as the number of visually similar distractors in the candidate pool grows. The sharpest decline occurs at low hard-negative counts: moving from 0–100 to 100–200 hard negatives per query,

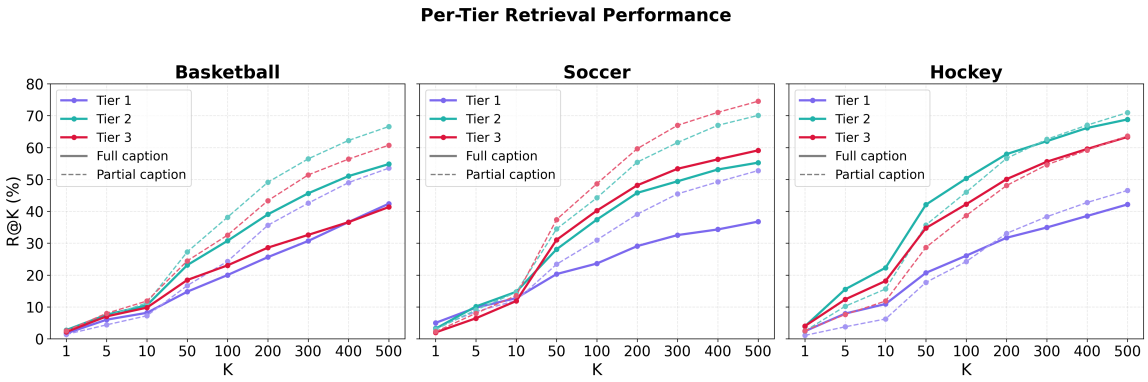


Figure 7. T3 Recall@K curves across tiers and sports on the test split. Solid lines: full-caption training. Dashed lines: attribute dropout training. Colors indicate difficulty tiers.

R@100 drops from 76.3% to 39.1% on hockey, with similar trends on basketball and soccer. Beyond 200 hard negatives, performance continues to degrade but plateaus at substantially lower levels. The full details are in Appendix B.3.

Qualitative example. Figure 8 shows a representative failure case in T3. The top-ranked clips capture the broad action and rink region described in the query, but fail to satisfy the finer visual constraint that identifies the correct clip. In this example, the retrieved clips show puck carriers entering the offensive zone, yet the jersey number does not match the query. This illustrates that to solve T3, the retrieval model must preserve fine-grained attribute consistency.

> **PILLAR 1 TAKEAWAY — DYNAMIC SCENE UNDERSTANDING**
>
> Perception is the strongest pillar, yet significant gaps remain. Fine-grained action QA reaches 73.91%, while structured captioning and compositional retrieval reveal persistent weaknesses in identity grounding and multi-attribute composition—capabilities that all higher pillars depend on.

4.2 Pillar 2: Causal Reasoning (T4–T6)

Perception tells us *what* happened; the reasoning pillar asks *why*. These tasks require reasoning about causal mechanisms linking actions to outcomes over 55–150 minutes of continuous play, spanning event explanation (T4), forward-looking prediction (T5), and extended narrative synthesis (T6).

Figure 8. T3 qualitative example (hockey). Given a natural-language query specifying multiple visual attributes, the model must retrieve the matching clip from 5,001 candidates (1 positive and 5000 negatives). Attributes in the query are color-coded by category: entity (jersey number) and spatial (rink location). The correct video is retrieved at the rank-792. The top-2 retrieved clips match the location but show the wrong player number (17 and 89 instead of the queried 5).

T4: Strategic Reasoning QA

Task formulation. Given a full-game video (∼55–150 min) and a question, the model must produce a free-form response explaining strategic reasoning behind game events. Unlike T2, which tests localized perception over short clips, T4 requires reasoning over extended game portions: identifying strategic errors, evaluating tactical execution, and interpreting latent dynamics such as momentum shifts. Compared to existing long-form video QA benchmarks [21, 74, 104] whose questions remain predominantly perceptual, T4 targets strategic causal reasoning where evidence may be spread across minutes of footage interleaved with irrelevant events.

Data and construction. We curate 1,000 questions from professional commentaries and game reports across all three sports, totaling 825 unique games. A multi-stage pipeline generates open-ended question–answer pairs followed by bias-mitigation filtering and human validation of temporal alignment, factual validity, and question quality. Reference answers are drawn from expert commentary and game reports and often encode multiple complementary causal explanations rather than a single objective ground truth. The detailed question construction pipeline is in Appendix C.1.

Evaluation. We use an LLM-as-a-judge protocol that scores each response on a 0–5 scale, assessing strategic depth, factual consistency with the reference answer, and reasoning coherence. The judge prioritizes coverage of key concepts and reasoning traces over surface-level phrasing. To accommodate the fact that strategic questions often admit multiple valid explanations, each model produces $k{=}5$ candidate answers, and we report the best-scoring response. The detailed judge prompt is in Appendix C.1.

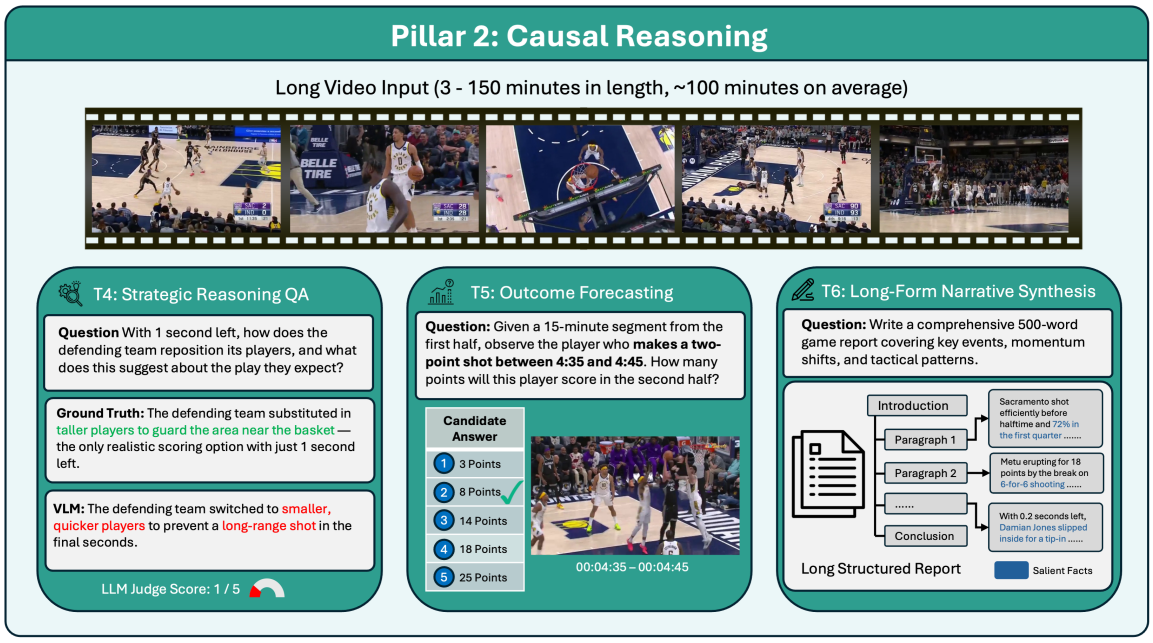


Figure 9. Overview of Pillar 2: Causal Reasoning. This pillar evaluates the ability to reason about game-level context through three tasks: strategic reasoning QA (T4), outcome forecasting (T5), and long-form narrative synthesis (T6).

Baselines. We evaluate frontier proprietary models (GPT-5.2, Gemini 3.1 Pro) and open-source video–language models (Qwen3-VL-32B, Molmo 2-8B). All models receive the full-game video and question. We additionally evaluate an *oracle baseline* in which GPT-5.2 receives the detailed play-by-play game log—a textual record of every action, its timestamp, the players involved, and the resulting game state—instead of the video. This isolates the reasoning component from the visual perception bottleneck.

Main results. Table 8 presents T4 scores by sport. All four models score near 2/5 on average, with Gemini-3.1-Pro achieving the highest overall average (2.17) and GPT-5.2 second (2.06). Gemini-3.1-Pro's lead is driven primarily by soccer (2.49), while GPT-5.2 leads on basketball (1.99). Basketball is the most difficult sport across all models (1.85–1.99). These low scores reflect the genuine difficulty of T4: answering these questions requires not just recognizing individual events but reasoning about causal relationships between tactical decisions, evaluating execution quality, and understanding how spatial structures and player interactions evolve over the course of a full game.

Oracle baseline. We evaluate GPT-5.2 with structured play-by-play logs replacing video input. The model reaches 2.46/5 on average, only ∼0.4 higher than the video-based setting (2.06, Table 8). Even with a precise description of every game event, the model must interpret the subtleties of the play to reach high judge scores. The pattern confirms that perception alone is not the primary bottleneck.

Judge reliability. We validate the LLM judge through two checks. First, two annotators

Sport	Molmo2-8B	Qwen3-VL-32B	GPT-5.2	Gemini-3.1-Pro
Hockey	1.89	2.08	2.07	**2.09**
Soccer	1.71	1.99	2.13	**2.49**
Basketball	1.85	1.96	**1.99**	1.93
Average	1.82	2.01	2.06	**2.17**

Table 8. T4 strategic reasoning QA results. Each model produces $k{=}5$ candidate answers per question. The best-scoring response is rated by a DeepSeek-V3 judge on a 0–5 scale and reported per sport. Temporal sampling details: Molmo2-8B (300 frames), Qwen3-VL-32B (768 frames), GPT-5.2 (500 frames), Gemini-3.1-Pro (1 FPS, compressed to 1 hr).

independently scored 60 T4 question–answer pairs, yielding a Mean Absolute Error of 0.12 against the LLM judge on the 0–5 scale and a Spearman's $\rho{=}0.85$ between human and LLM-judge rankings. Agreement is strong on both absolute scoring and model ranking. Second, because T4 questions are open-ended and may admit multiple valid answers, we check whether the judge credits alternative reasoning. In cases where two annotators identified different but equally valid causes for the same event—for example, attributing a missed defensive stop to two distinct preceding plays—both received maximum scores. This demonstrates that the rubric rewards valid alternative reasoning rather than enforcing a single formulaic answer.

Qualitative example. Figure 10 illustrates a recurring T4 failure mode. When faced with an unusual play, models default to plausible-but-generic actions instead of describing what actually happened in the video. Here, the inbounder deliberately throws the pass off the defender's back to score before the shot clock expires. All four models hallucinate canonical half-court actions instead (screen-the-screener, back-screen alley-oop).

T5: Outcome Forecasting

Task formulation. Given a video segment capturing a sequence of play (3–15 minutes) and a question about a future event, the model must predict the outcome by selecting the correct answer from a candidate set. The target event occurs beyond the input window, requiring the model to infer the most probable course of game development. Unlike trajectory forecasting [26, 41, 49, 67] that predicts short-horizon spatial positions, T5 targets semantically rich outcomes (who will score, which strategy will be employed, how a game state will evolve) requiring understanding of complex causal mechanisms. Questions span *performance forecasting* (predicting player or team statistical accomplishments), *game state evolution* (anticipating scores, possessions, and game development), and *strategic intention* (identifying the most probable tactical shifts).

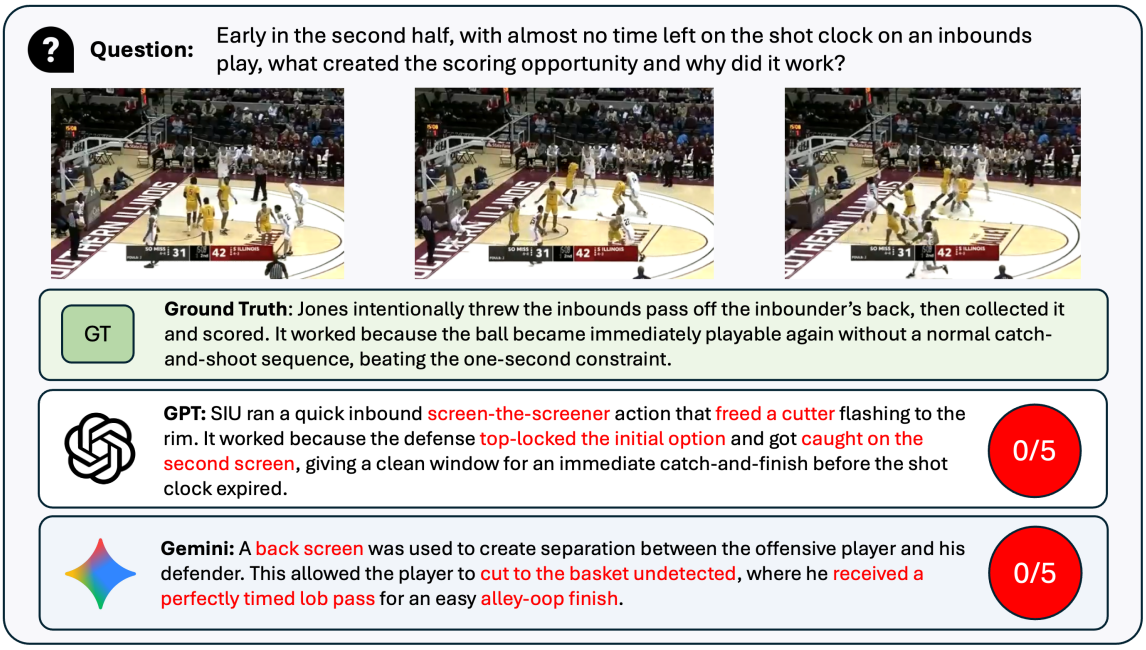


Figure 10. T4 qualitative example (basketball). The question asks how a late-shot-clock scoring opportunity was created on an inbounds play. The correct answer is that the inbounder intentionally threw the pass off the defender's back to bypass the catch-and-shoot sequence. Both models hallucinate common half-court actions (screen-the-screener, back-screen lob) instead. Each model's best response (out of k=5) is shown with its LLM judge score (0–5).

Data and construction. We curate 114K multiple-choice questions spanning 15 question types organized into the three forecasting categories above, using full-game videos with dense play-by-play event annotations and observation windows of 3–15 minutes. All questions reference future events relative to the observation window, and target players are identified by indirect descriptions of their actions rather than by name to prevent shortcut solutions. Full construction details and per-sport statistics are in Appendix C.2.

Evaluation. We report top-1 accuracy as the primary metric. We also report calibration error (CE) following [42], which measures alignment between predicted confidence and empirical correctness. CE $= 0$ indicates perfect calibration.

Baselines. We evaluate five models in a zero-shot setting: GPT-5.2, Gemini 3.0 Pro, Qwen3-VL 8B, BIMBA [33], and Molmo 2 8B. We additionally fine-tune Qwen3-VL 8B and BIMBA on the combined three-sport T5 training set. To isolate the perception bottleneck, we evaluate an oracle baseline in which GPT-5.2 receives the play-by-play log—a textual record of every event in the observation window, including timestamps, player actions, and game state—instead of video. Full training details are in Appendix C.2.

Main results. Table 9 reports accuracy and calibration error per sport and overall. Random guessing yields 29.1% overall. Among zero-shot baselines, Gemini 3.0 Pro is strongest with 43.2% overall accuracy. Fine-tuning provides large gains on both accuracy and cal-

Model	Bball.	Soccer	Hockey	Overall	CE ↓
Random baseline	27.3	32.1	29.8	29.1	–
Molmo 2 8B (ZS)	36.0	36.5	33.7	35.1	0.22
BIMBA (ZS)	34.9	34.1	33.1	34.0	0.16
BIMBA (FT)	41.9	41.6	37.3	39.9	0.06
Qwen3-VL 8B (ZS)	37.4	39.6	35.3	36.9	0.23
Qwen3-VL 8B (FT)	**46.1**	45.1	**43.5**	**44.8**	**0.01**
GPT-5.2 (ZS)	38.1	44.3	33.1	38.2	0.28
Gemini 3.0 Pro (ZS)	45.9	**49.2**	38.4	43.2	–
Oracle (GPT-5.2)	42.3	42.7	40.9	41.9	0.28
Human	*55.0*	*73.3*	*45.0*	*58.9*	–

Table 9. T5 outcome forecasting results. Accuracy (%) and calibration error (CE, lower is better, with 0 = perfect calibration) per sport and overall. Best zero-shot or fine-tuned value in each column is bolded. Full ablations across temporal sampling rates are in Appendix C.2. The oracle baseline replaces video with play-by-play logs.

ibration: fine-tuned Qwen3-VL 8B reaches 44.8% overall (+7.9 points over its zero-shot baseline), and its calibration error drops from 0.23 to 0.01. The oracle baseline on basketball reaches 42.3%, only 4.2 points above the non-oracle video-based variant (38.1%). This small gain has two possible explanations: either accurate perception alone is insufficient for strong forecasting, or play-by-play logs do not fully capture all visually salient information. Hockey is the most difficult sport across all models, with the best fine-tuned model reaching only 43.5%.

Prediction horizon and calibration. Figure 11 analyzes how the prediction horizon—the temporal distance between the end of the observation window and the target event—affects accuracy and confidence. Using zero-shot Qwen3-VL 8B on basketball game-outcome forecasting, we construct four variants of the same question per game, each with a different observation window separated by at least 15 minutes, while keeping the question and answer options fixed. This isolates temporal distance as the primary varying factor. As the horizon lengthens, accuracy drops steadily from 91% to 45%. However, the model's confidence barely decreases (98% to 94%), creating an overconfidence gap that widens with horizon length. The model does not adjust its confidence to reflect the increased uncertainty of longer-horizon predictions.

Per-category breakdown. Accuracy varies substantially across the three forecasting categories. For GPT-5.2 zero-shot, *game state evolution* questions are easiest (62.95%), benefiting from directly observable signals such as current score, time remaining, and lead margin that can be read via OCR. *Performance forecasting* and *strategic intention* are sub-

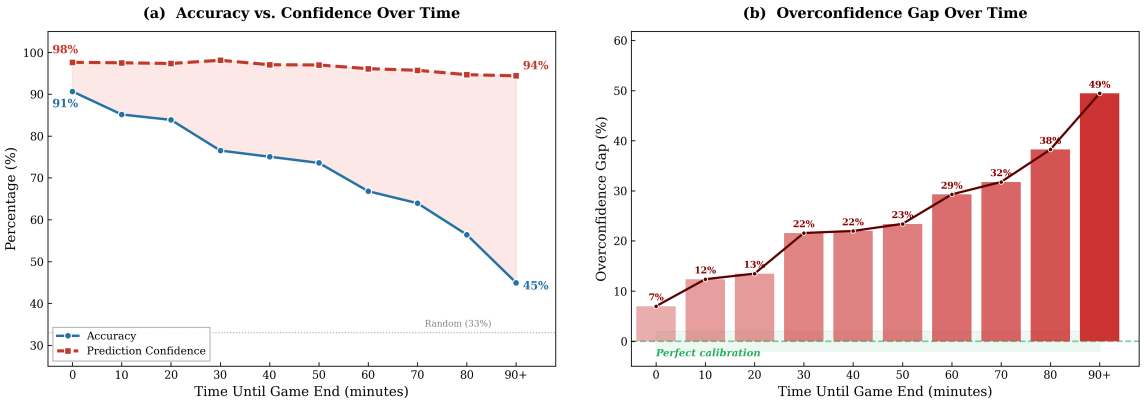


Figure 11. T5 prediction horizon analysis (Qwen3-VL 8B zero-shot game-outcome forecasting). (a) Accuracy drops steadily as the temporal distance to the target event increases. (b) Model confidence remains high, creating an overconfidence gap that widens with horizon length.

stantially harder, achieving 33.03% and 37.60% respectively. Both demand finer-grained evidence integration: tracking individual player performance trajectories, recognizing tactical patterns, and projecting how observed tendencies will continue.

Human comparison. We conducted a human study with sport-experienced participants, each answering 10 questions and self-reporting how confident they were in each answer on a 1–3 scale (1: guessing, 2: somewhat confident, 3: confident). Humans achieved 55.0% on basketball (4 participants), 73.3% on soccer (3 participants), and 45.0% on hockey (2 participants). The gap to the strongest models is largest on soccer (+24.1 points over Gemini 3.0 Pro), with smaller gaps on basketball (+8.9 over FT Qwen3-VL) and hockey (+1.5 over FT Qwen3-VL). Unlike models, human accuracy rises monotonically with self-rated confidence, ranging from 50% on low-confidence responses to 100% on high-confidence ones, indicating well-calibrated uncertainty about predictions.

Qualitative example. Figure 12 shows a representative T5 failure. GPT-5.2 assigns 92% confidence to the wrong answer. Qwen3-VL exhibits the opposite failure mode, spreading confidence near-uniformly across options without committing. Neither model integrates the visual evidence with calibrated uncertainty about the future.

T6: Long-Form Narrative Synthesis

Task formulation. Given a full game video (∼55–150 min) and a writing prompt, the model must synthesize a narrative report (∼500 words) covering key events, standout performances, and strategic developments. Unlike video summarization methods [14, 30, 60, 103] that rely on dialogue or narration, T6 requires narratives grounded entirely in visual evidence from hours of multi-agent interaction, demanding saliency and factual precision across extreme temporal scales.

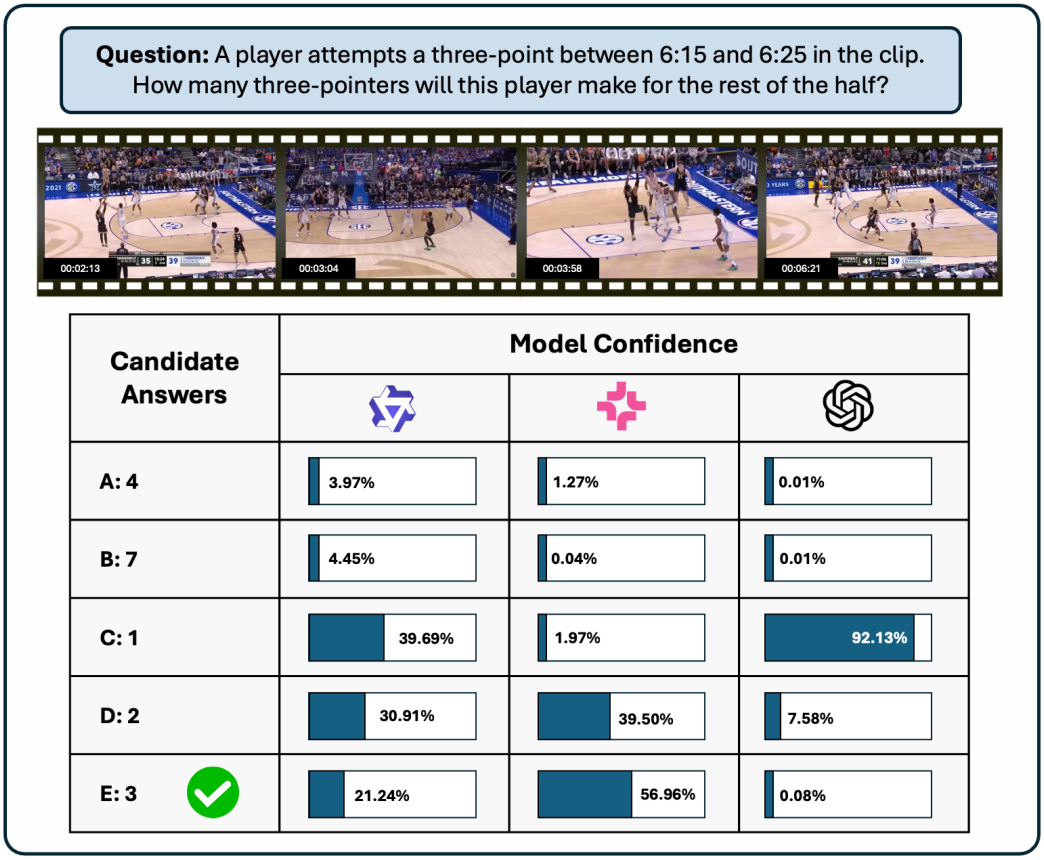


Figure 12. T5 qualitative example. The target player is identified via an indirect reference ("the player who makes a three-point attempt between 6:15 and 6:25") rather than by name, requiring the model to first ground the description in the observation window before forecasting how many three-pointers that player will make for the remainder of the half. The correct answer is E: 3. Confidence distributions for each evaluated model are shown.

Data and construction. We define 10 report templates per sport, five targeting single-game analysis (e.g., overall game summary, player performance, decisive moments) and five targeting multi-game synthesis across 2–10 games linked by a coherent theme. Across the three sports we curate 17,715 training samples and 1,000 evaluation samples spanning roughly 11,500 hours of full-game video. Reference reports are generated by GPT-5 conditioned on aligned multimodal resources (video, play-by-play logs, structured metadata, and journalist reports) together with the writing template, ensuring consistent structure and verifiable factual content grounded in game-observable evidence. Full per-sport statistics and template details are in Appendix C.3.

Evaluation. We evaluate along three dimensions using an LLM-as-a-judge framework with Qwen3-235B-A22B Thinking as the judge: *factual accuracy* via atomic fact decomposition against ground-truth report [53], *saliency* measuring coverage of key events and performances as identified by a state-of-the-art LLM given oracle game information (play-by-play logs, box scores, and original journalist reports), and *writing style* rated on a 1–5

scale for coherence, topic adherence, and length compliance.

Baselines. We evaluate four model configurations. *Qwen3-VL 8B* processes frames sampled at 1 FPS provided alongside the writing instruction for end-to-end report generation. *LLoVi* [99] is a two-stage pipeline that first uses a LLaVA-Video model fine-tuned on Pillar 1 perception tasks to generate dense 10-second captions for each input video, then feeds these captions with the writing instructions to GPT-5 for final report generation. *GPT-5* processes 500 uniformly sampled frames (budget split evenly across games for multi-game settings). *Gemini 3.1 Pro* processes the full video downsampled to fit the 1-hour API input limit. We additionally evaluate an *oracle baseline* in which GPT-5 receives the detailed play-by-play log—a textual record of every event in the game—instead of video, isolating the perception bottleneck. Full training details are in Appendix C.3.

Main results. Table 10 reports factual accuracy, saliency, and writing style. GPT-5 and Gemini 3.1 Pro both reach roughly 72–73% factual accuracy and 4.7–4.8 writing style. However, both score only ∼7% on saliency. The open-source Qwen3-VL 8B trails substantially across all three metrics (35.66% factual, 2.13% saliency, 3.50 style). LLoVi—which combines a Pillar-1-fine-tuned captioner with GPT-5 for synthesis—reaches 25.20% factual accuracy, comparable to or below standalone Qwen3-VL, indicating that stage-wise pipelines do not straightforwardly compose into stronger long-form synthesis. The gap between factual accuracy and saliency is the central T6 finding: models can produce reports that are largely factually correct but cover only a small fraction of the content that an expert journalist considers most important. The oracle baseline, which replaces video with the complete play-by-play log, raises factual accuracy to 87.19% but lifts saliency only to 20.60%. This indicates that even with perfect access to game events, the model cannot identify which events are worth reporting.

Failure mode analysis. Figure 13 shows a generated report against the ground truth with three types of recurring failures highlighted. *Factual errors* span three difficulty levels: game-state errors involve incorrect OCR-derived facts, single-play errors involve misattributed events, and aggregated-statistics errors require correct player identification across multiple plays and arithmetic over a time window. *Saliency failures* follow a consistent pattern: models fill reports with generic, low-information statements rather than the trajectory-changing moments and concrete statistics that distinguish expert reports.

Cross-source validation. Our saliency metric measures coverage against a single reference report, raising the question of whether the target is achievable or specific to our reference choice. To check this, we scraped reports from a second source covering the same games (10 randomly selected test samples). These alternative professional reports achieve 45.8% saliency against our ground-truth reports—25 points above the oracle baseline. This demonstrates that different professional journalists emphasize simi-

Sport	Model	Factual ↑	Saliency ↑	Style ↑
	Qwen3-VL 8B	26.57	3.48	3.49
	LLoVi	30.15	3.69	4.67
Basketball	GPT-5	63.11	4.08	**4.76**
	Gemini 3.1 Pro	62.41	4.59	4.71
	Oracle	**85.09**	**15.49**	4.52
	Qwen3-VL 8B	37.96	1.22	3.66
	LLoVi	18.17	3.09	4.51
Hockey	GPT-5	75.73	9.64	**4.76**
	Gemini 3.1 Pro	75.33	8.91	4.51
	Oracle	**87.08**	**24.30**	4.62
	Qwen3-VL 8B	43.57	1.62	3.33
	LLoVi	27.62	3.45	4.61
Soccer	GPT-5	77.98	7.54	**4.93**
	Gemini 3.1 Pro	82.66	8.68	4.90
	Oracle	**89.76**	**22.23**	4.86
	Qwen3-VL 8B	35.66	2.13	3.50
	LLoVi	25.20	3.41	4.60
Overall	GPT-5	71.99	7.06	**4.81**
	Gemini 3.1 Pro	73.01	7.33	4.70
	Oracle	**87.19**	**20.60**	4.66

Table 10. T6 long-form narrative synthesis results. Factual accuracy and saliency coverage are reported as percentages (%). Writing style is scored on a 1–5 scale. *Oracle* replaces video with ground truth play-by-play logs fed to GPT-5. *LLoVi* uses a fine-tuned perception model to generate dense captions, then GPT-5 for report synthesis.

lar core events, confirming the validity of the saliency evaluation and indicating that the gap between current models (7%) and the attainable upper bound (45%) is meaningful.

Judge reliability. We validate the T6 evaluation protocol via a human agreement study over 15 generated reports containing 383 atomic facts, each annotated by two human raters. The LLM judge reaches 99.3% agreement with human annotations on saliency. Model rankings remain stable across independent judges, with Spearman ρ=0.98.

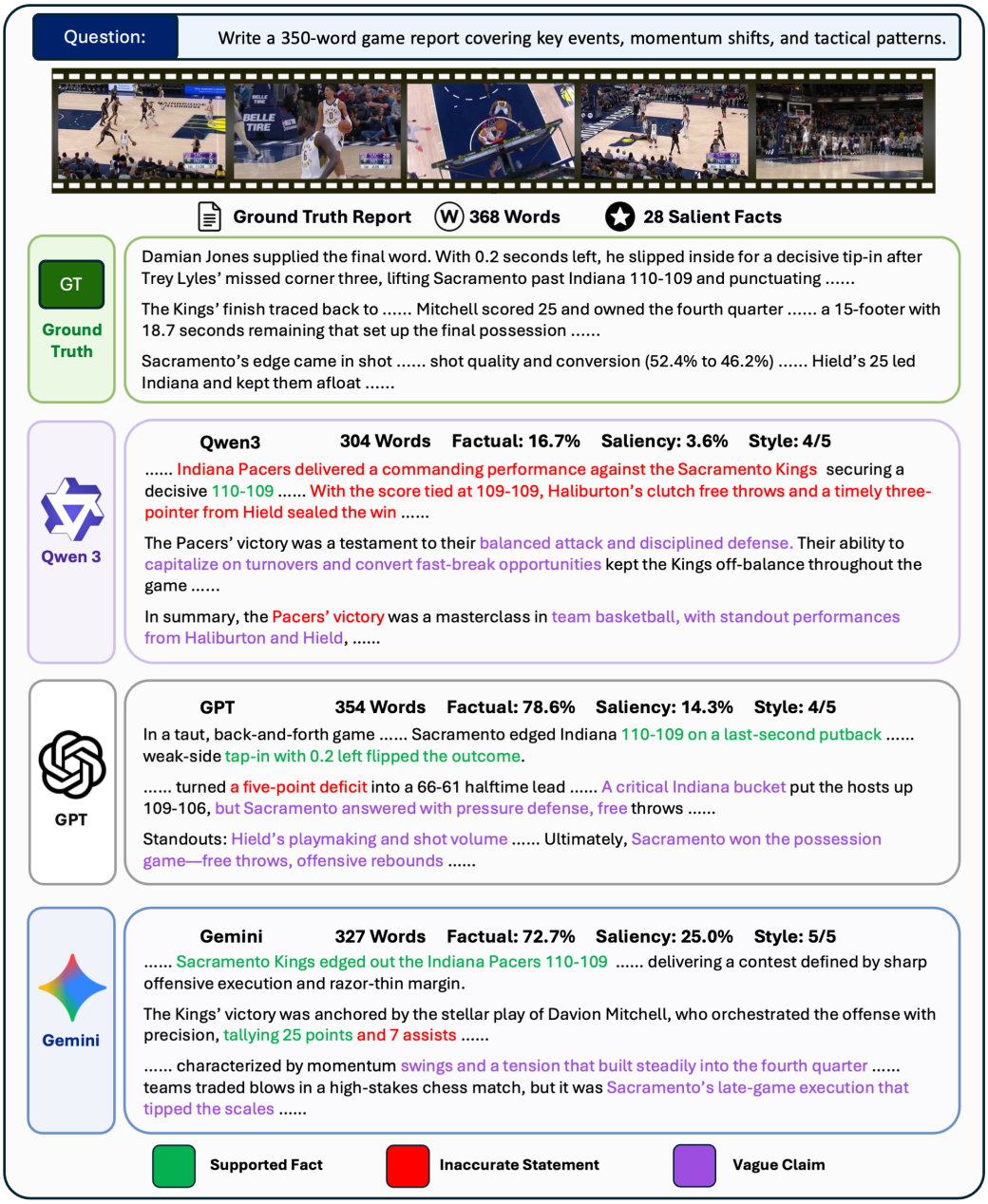


Figure 13. T6 qualitative comparison. Green: factually correct statements. Red: factual errors (misattributions or wrong statistics). Purple: vague, low-information claims.

> **PILLAR 2 TAKEAWAY — CAUSAL REASONING**
>
> Models that perform competently on perception degrade substantially on reasoning tasks. Oracle experiments providing perfect perceptual inputs improve factual recall but leave large gaps in strategic reasoning and saliency, suggesting that causal and strategic reasoning is the major bottleneck across this pillar.

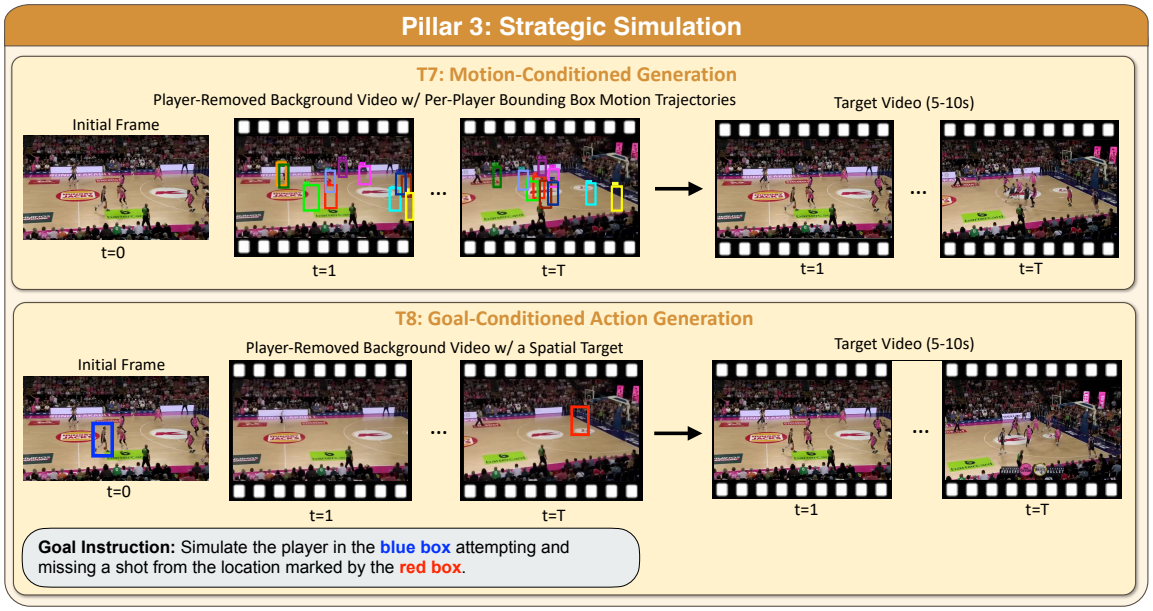


Figure 14. Overview of Pillar 3: Strategic Simulation. This pillar tests the ability to simulate alternative futures through two video generation tasks: motion-conditioned generation (T7), where players follow prescribed trajectories, and goal-conditioned action generation (T8), where the model plans actions toward a specified goal.

4.3 Pillar 3: Strategic Simulation (T7–T8)

This pillar evaluates whether models can simulate alternative futures through video generation. Given a short game clip (5–10s), both tasks require producing a realistic video of how the scene evolves if players follow specified trajectories (T7) or execute a specified action to achieve a goal (T8). Both settings require generating physically plausible multi-agent dynamics across 10 or more interacting agents (Figure 14).

T7: Motion-Conditioned Generation

Task formulation. Given an initial frame showing all players in their starting positions, a *player-removed background video* (the original video with all players erased via video inpainting, leaving only the court or pitch and static scene elements), and a set of per-player motion trajectories specified as time-aligned bounding-box sequences, the model must generate a video in which players follow the prescribed trajectories while remaining visually, physically, and temporally coherent. The first frame serves as an appearance reference for all initially visible players, while players entering the scene later are additionally conditioned on a crop from their first visible bounding box. Unlike prior trajectory-conditioned generation methods [48, 55, 80, 96], which primarily focus on one or two objects in relatively simple scenes, T7 evaluates large-scale multi-agent coordination, where 10+ players move simultaneously, interact physically, and frequently

		Video mIoU ↑	Feature Sim. ↑
Soccer	ATI [78]	0.402	0.507
	MagicMotion [40]	0.544	0.708
	Ours (100 clips)	**0.611**	**0.804**
	Ours (1000 clips)	0.619	0.794
Basketball	ATI [78]	0.397	0.617
	MagicMotion [40]	0.466	0.725
	Ours (100 clips)	**0.513**	**0.787**
	Ours (1000 clips)	0.509	0.782

Table 11. T7 motion-conditioned generation results. Video mIoU measures trajectory fidelity. Feature similarity measures visual consistency over time. ATI and MagicMotion are evaluated on a shared 100-clip subset due to computational cost. Our method is evaluated on both the same 100-clip subset (for direct comparison, bolded) and the full 1,000-clip validation set.

occlude one another. The model must therefore synthesize realistic player appearances into the empty background while preserving identity consistency, trajectory fidelity, and temporal continuity throughout the sequence. Formally, the trajectory for player i is represented as a sequence of frame-aligned bounding boxes $\{(x^i_{\min,t}, y^i_{\min,t}, x^i_{\max,t}, y^i_{\max,t})\}_{t=1}^{T}$, where t indexes the frame and T is the clip length. The full motion condition is the set of all player trajectories $\{\mathbf{b}^i\}_{i=1}^{N}$ for N tracked players.

Data and construction. Each instance consists of (1) an initial frame, (2) per-player motion trajectories as bounding-box sequences, and (3) a player-removed background video generated via video inpainting [36]. We apply explicit quality filtering to remove instances with unstable tracking, severe occlusion, or visible inpainting artifacts (e.g., residual player silhouettes, texture bleeding). After filtering, T7 contains 123K soccer clips and 166K basketball clips totaling roughly 484 hours of video. The two sports differ substantially in scene density: soccer clips contain 20.6 visible players on average versus 10.0 for basketball. We evaluate on a 1,000-clip subset of the test split per sport. Full dataset statistics are in Appendix D.1.

Evaluation. We evaluate with two metrics: *Video mIoU* [7], measuring spatiotemporal alignment between player trajectories in generated and reference videos, and *temporal feature similarity*, comparing SigLIP [76] features from corresponding player regions across frames to assess visual consistency over time. Higher is better for both.

Baselines. Our reference method fine-tunes Wan 2.1 [75] on SVI-Bench data, extending it to accept structured input conditions (initial frame, per-player bounding-box trajecto-

	Video mIoU ↑			Feature Similarity ↑		
#Players	ATI	MagicM.	Ours	ATI	MagicM.	Ours
11–15	0.429	0.573	**0.663**	0.533	0.746	**0.831**
16–20	0.401	0.557	**0.611**	0.528	0.719	**0.817**
21+	0.394	0.516	**0.595**	0.468	0.681	**0.775**
Overall	0.402	0.544	**0.611**	0.507	0.708	**0.804**

Table 12. T7 generation quality by player count (soccer). Performance binned by the number of visible players per clip. MagicM. = MagicMotion.

ries, and player-removed background video). We additionally evaluate two off-the-shelf motion-conditioned generation baselines: *ATI* [78] operates on point-trajectory inputs, which we derive from our bounding-box centers, and *MagicMotion* [40] operates directly on bounding-box trajectories. Both baselines are applied zero-shot without fine-tuning on SVI-Bench data. ATI and MagicMotion are evaluated on a shared 100-clip subset due to computational cost. Our method is evaluated on both this 100-clip subset (for direct comparison) and the full 1,000-clip validation set (to confirm stability at scale). Full training and inference details are in Appendix D.1.

Main results. Table 11 reports trajectory fidelity (Video mIoU) and visual consistency (temporal feature similarity) on both sports. Our method achieves Video mIoU of 0.611 on soccer and 0.513 on basketball, outperforming MagicMotion (0.544 / 0.466) and ATI (0.402 / 0.397). Temporal feature similarity follows the same trend (soccer: 0.804 vs. 0.708 / 0.507. Basketball: 0.787 vs. 0.725 / 0.617). Scores remain stable when evaluated on the full 1,000-clip set ($\pm$0.01 from the 100-clip subset). Soccer outperforms basketball across all methods, likely because soccer scenes contain smaller players with fewer severe inter-player occlusions. Even with task-specific fine-tuning, generated player positions deviate from the prescribed trajectories in roughly half of cases, leaving substantial headroom for future work.

Effect of player count. Soccer clips span a wide range of player counts (11 to 21+), enabling analysis of how generation quality degrades in denser scenes (basketball clips typically contain $\sim$10 players, so a similar analysis is less informative). Table 12 reports performance across three player-count bins. Both metrics degrade monotonically as player count increases. For our method, Video mIoU drops from 0.663 (11–15 players) to 0.595 (21+), and feature similarity drops from 0.831 to 0.775. Similar degradation appears for both baselines. Our method consistently outperforms ATI and MagicMotion across all player-count ranges. The degradation reflects the increased difficulty of gen-

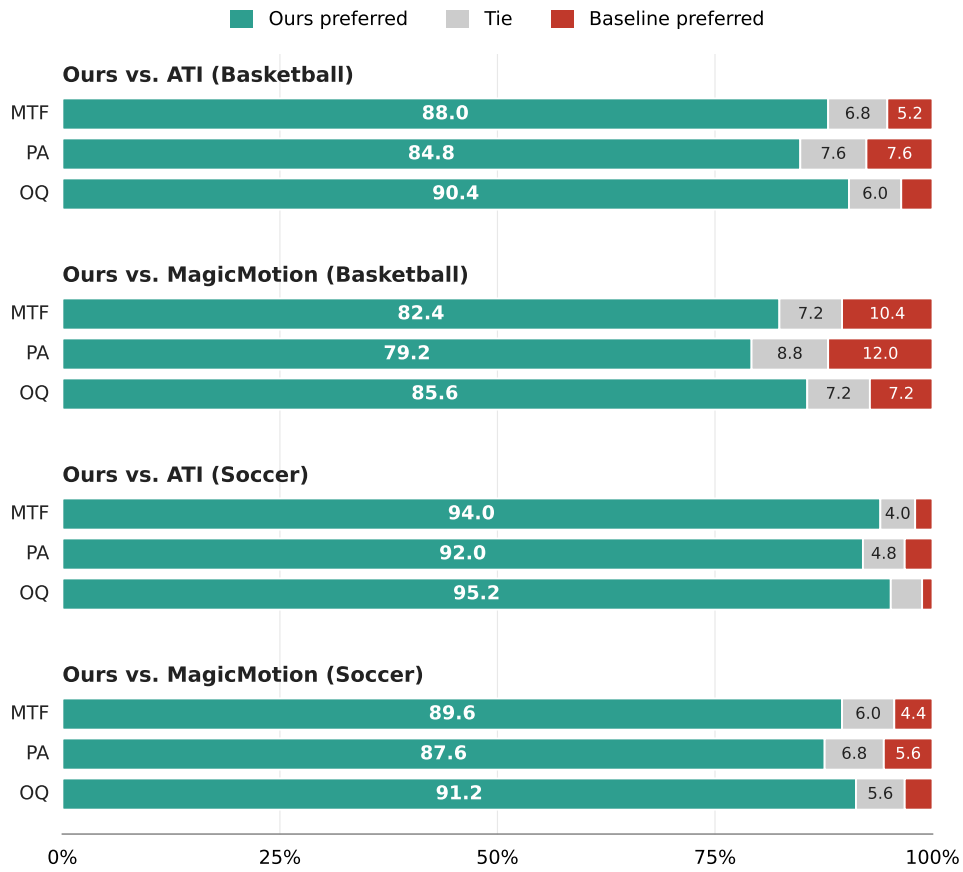


Figure 15. T7 human evaluation preference rates. Ten participants each judged 25 video pairs per sport-baseline combination across three criteria: Motion Trajectory Following (MTF), Player Appearance (PA), Overall Quality (OQ). Our method is preferred in 79.2–95.2% of comparisons across both baselines and both sports. Full numerical breakdown in Appendix D.1.

erating coherent multi-agent motion in dense environments, where more players introduce additional occlusions, interactions, and spatial constraints.

Human evaluation. We complement automatic metrics with a human study. Ten participants each judged 25 video pairs per sport-baseline combination across three criteria: Motion Trajectory Following (MTF), Player Appearance (PA), and Overall Quality (OQ). Figure 15 reports per-criterion preference rates. Participants consistently prefer our method, with win rates of 79.2–90.4% on basketball and 87.6–95.2% on soccer across both baselines and all three criteria. The soccer-vs-basketball gap mirrors the pattern observed in the automatic metrics.

Qualitative example. Figure 16 compares generated outputs from ATI, and our method on a representative soccer clip against the ground-truth video. Our method follows the prescribed trajectories most faithfully. ATI produces trajectory violations and spurious players. However, even our best outputs exhibit identity inconsistencies, with jersey colors blurring across frames and appearance details degrading over time, indicating

Figure 16. T7 motion-conditioned generation (soccer). Top row: prescribed trajectories (each color = one player). Second row: ground truth. Remaining rows: generated outputs. **Yellow** circles highlight players deviating from the prescribed trajectories. **Blue** circles mark players that follow the correct motion but exhibit incorrect visual appearance (e.g., wrong jersey color). The zoomed crop (right) compares a generated player against the ground truth at higher resolution.

remaining gaps between generated and real footage.

Method	mIoU ↑	Feature Sim. ↑	Goal Acc. (%) ↑
ATI [78]	0.047	0.067	40.5
MagicMotion [40]	0.129	0.169	31.4
Ours (100 clips)	**0.344**	**0.468**	**50.2**
Ours (1000 clips)	0.309	0.415	48.3

Table 13. T8 goal-conditioned action generation results. Last-frame mIoU (spatial accuracy), last-frame feature similarity (SigLIP-v2 embeddings, visual fidelity), and goal accuracy from a fine-tuned QA evaluator. ATI and MagicMotion are evaluated on a shared 100-clip subset. Our method is evaluated on both the same subset (bolded) and the full 1,000-clip set.

T8: Goal-Conditioned Action Generation

Task formulation. Given an initial frame, a player-removed background video, and a textual instruction specifying target player(s), spatial constraints (start and end bounding boxes), and a desired action outcome (e.g., a rebound, a contested layup), the model must generate a video in which the specified players execute a coherent action sequence that achieves the described objective. An example instruction is: *"Simulate Player #23 executing a 3 Pt Made from the Wing with a step-back jump shot while contested by Defender #5."* Unlike T7, which prescribes exact motion trajectories, T8 requires the model to *plan* intermediate actions toward a high-level goal under explicit spatial constraints. This demands implicit understanding of environment dynamics and goal-directed reasoning. This distinguishes T8 from standard open-ended text-conditioned video generation settings [8, 25, 31, 71].

Data and construction. T8 contains 74,003 basketball clips (64,003 train / 5,000 validation / 5,000 test) at 81 frames, 15 fps, 832×480 resolution. Each clip specifies one or more target players, each with an associated action outcome and spatial constraints (start and end bounding boxes), for 79,448 target players in total. The actions span five categories: shooting (47.2%), playmaking (14.2%), offensive set plays (14.5%), possession changes (14.4%), and fouls (7.8%). We restrict T8 to basketball because reliable annotations for goal-conditioned generation are currently available only for this sport. Full taxonomy and per-action statistics are in Appendix D.2.

Evaluation. We evaluate with three complementary metrics. *mIoU* on the final frame measures bounding-box overlap between generated and target player positions. *Feature similarity* on the final frame assesses visual fidelity of the realized outcome. *Goal accuracy* is computed via a fine-tuned video-language model that evaluates whether the generated video achieves the specified objective. Higher is better for all three.

Category	n	mIoU ↑	Feature Sim. ↑	Goal Acc. (%) ↑
Scoring	364	0.294	0.403	**47.0**
Ball Handling	152	0.306	0.408	45.4
Set Plays	232	**0.335**	**0.436**	16.7
Def./Foul	252	0.309	0.420	30.3
Overall	**1000**	0.309	0.415	48.3

Table 14. T8 per-category breakdown for our method (1,000-clip set). Best entry per metric is bolded. Full per-action table is in Appendix D.2.

Baselines. Our reference method adapts the T7 Wan 2.1 [75] framework to the goal-conditioned setting, replacing trajectory inputs with textual goal specifications and spatial endpoint constraints. We also evaluate *ATI* [78] and *MagicMotion* [40] off-the-shelf, providing them with the start and end bounding boxes. ATI and MagicMotion are evaluated on a shared 100-clip subset due to computational cost. Our method is evaluated on both the 100-clip subset (for direct comparison) and the full 1,000-clip validation set (to confirm stability at scale). Full training and inference details are in Appendix D.2.

Main results. Table 13 reports the three evaluation metrics. Our method substantially outperforms both baselines across all three: mIoU 0.344 vs. 0.129 (MagicMotion) and 0.047 (ATI), feature similarity 0.468 vs. 0.169 and 0.067, and goal accuracy 50.2% vs. 31.4% and 40.5%. ATI and MagicMotion fail at goal realization because neither has a mechanism to translate a high-level goal description into the intermediate actions required to satisfy the spatial constraints. Even our fine-tuned method reaches only 50% goal accuracy, indicating that goal-directed video generation remains largely unsolved.

Per-category breakdown. Table 14 reports per-category performance. Set plays achieve the highest mIoU (0.335) but the lowest goal accuracy (16.7%)—the model places players at prescribed endpoints but fails to generate the intermediate motion that defines set plays such as pick-and-rolls and screens. Scoring actions show the inverse pattern (mIoU 0.294, goal accuracy 47.0%) because shot outcomes produce distinctive visual signatures that the evaluator can recognize even when player positioning is imprecise.

Human evaluation. We complement automatic metrics with a human study. Ten participants each judged 25 video pairs per baseline across three criteria: Target Player Positioning (TPP), Action Correctness (AC), and Overall Quality (OQ). Figure 17 reports per-criterion preference rates. Participants consistently prefer our method, with win rates of 94.4–96.0% against ATI and 92.0–93.6% against MagicMotion.

Qualitative example. Figure 18 shows a multi-player scenario that requires coordinated

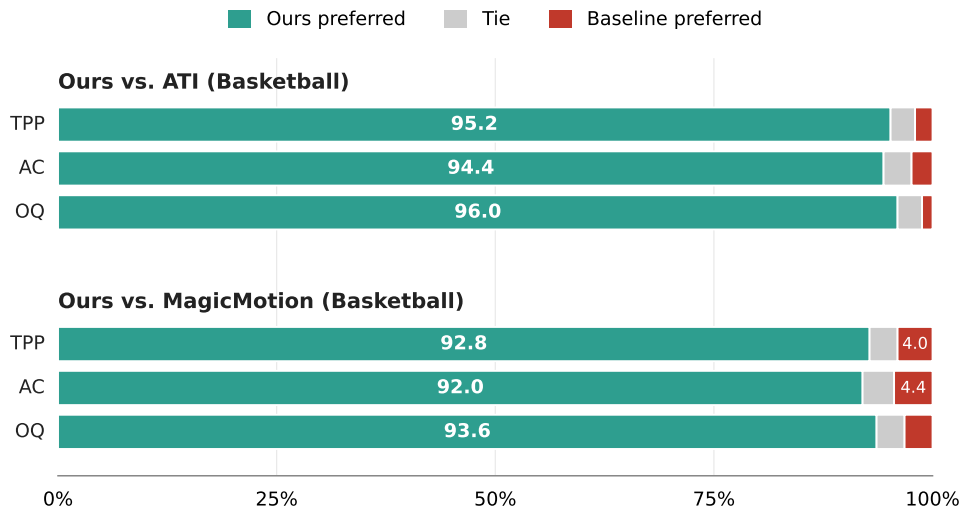


Figure 17. T8 human evaluation preference rates. Ten participants each judged 25 video pairs per baseline across three criteria: Target Player Positioning (TPP), Action Correctness (AC), and Overall Quality (OQ). Our method is preferred in 92.0–96.0% of comparisons.

motion between two players. ATI and MagicMotion fail to satisfy the spatial constraints and produce inconsistent player positions, with ATI additionally exhibiting background collapse in later frames. Our model partially succeeds, correctly placing one player at its target but failing for the second.

> **PILLAR 3 TAKEAWAY — STRATEGIC SIMULATION**
>
> Even fine-tuned video generation models cannot reliably coordinate multi-agent motion, with half of generated players deviating from prescribed trajectories. Performance degrades further when models must plan actions toward a goal, revealing that goal-directed video generation remains an open challenge.

4.4 Pillar 4: Agentic Synthesis (T9)

T9: Cross-Corpus Agentic Reasoning

The final pillar evaluates whether models can operate as autonomous analysts over a large cross-referenced multimodal corpus (∼1.8M clips, ∼33K documents) (Figure 19).

Task formulation. Given a complex natural-language query about a game, the model must plan a multi-step retrieval strategy, gather evidence from heterogeneous sources (video clips, game reports, statistical records), and reason over the collected evidence to produce a final answer. The agent is equipped with search and QA tools over a document database (post-game reports, game-level and season-level statistics) and a video database (footage segmented into 10–15s clips). While tool-augmented reasoning has

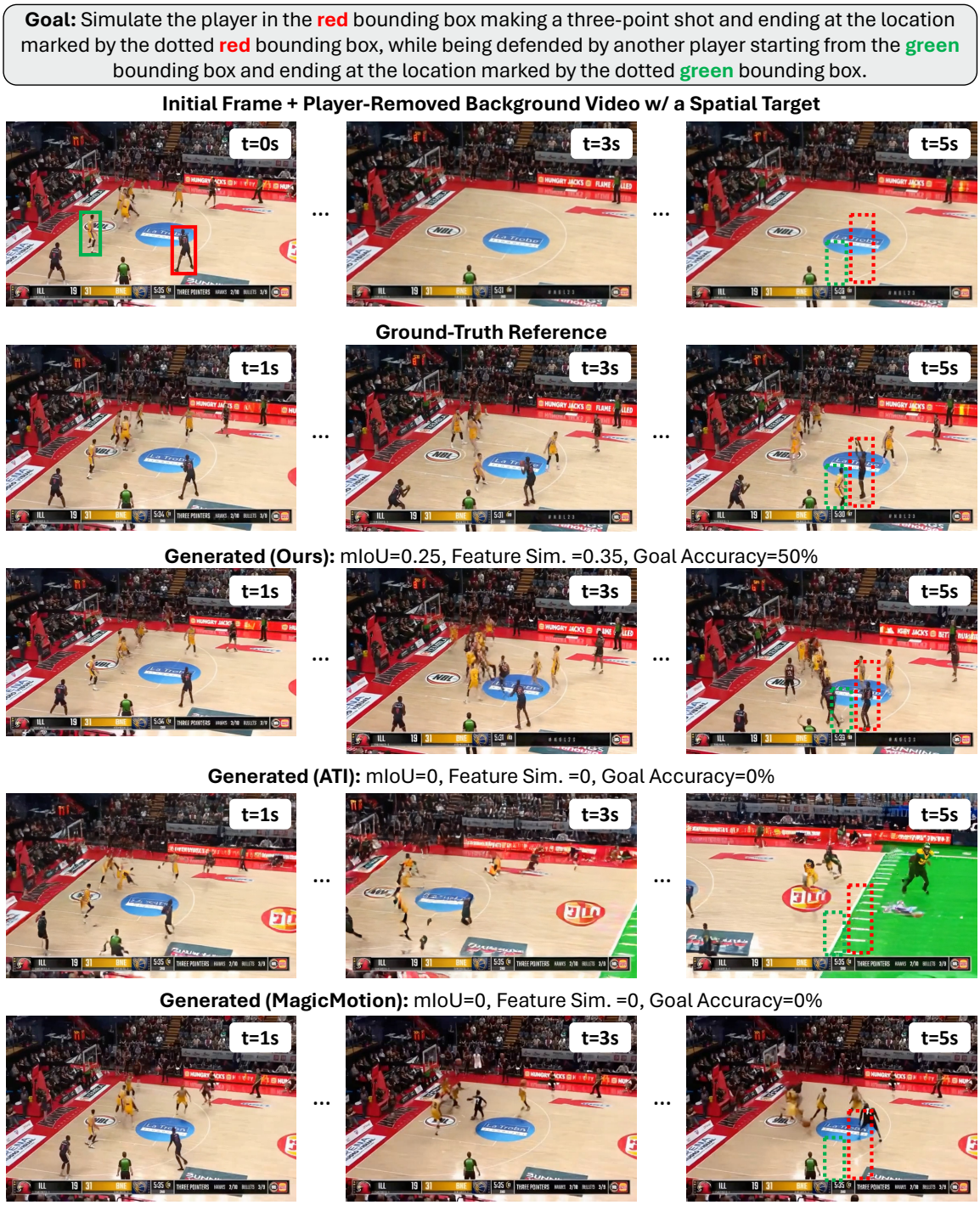


Figure 18. T8 multi-player goal-conditioned generation. The instruction requires coordinated motion between two players. The **red** box marks the first player's start position and the **dotted red** box marks its target. The **green** box marks the second player's start and the **dotted green** box marks its target.

been explored in text [39, 51, 61, 106] and video settings [58, 93, 101], T9 extends this to *multimodal evidence at corpus scale*: the agent must integrate evidence across modalities through complex reasoning patterns (looping, backtracking, conditional branching,

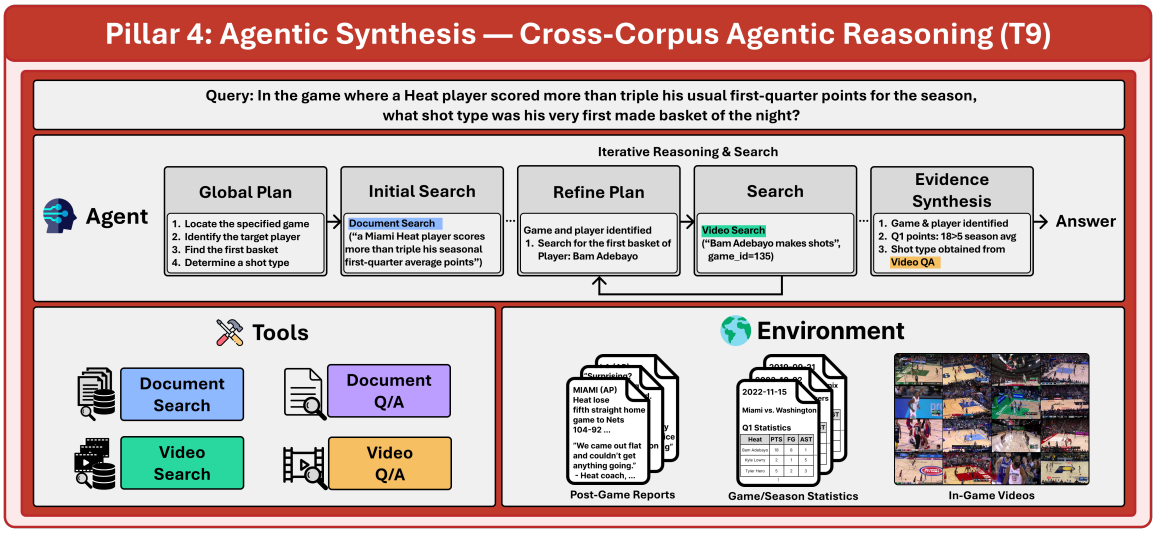


Figure 19. Overview of Pillar 4: Agentic Synthesis. This pillar evaluates the ability to autonomously gather and integrate multimodal evidence through a single task: cross-corpus agentic reasoning (T9), where the agent plans and executes tool-assisted search across large-scale heterogeneous sources to answer complex strategic queries.

numerical aggregation) over $\sim$1.8M clips and $\sim$33K documents across three sports.

T9 adopts the hard-to-find but easy-to-verify principle from prior agentic search work [84]. Each question begins with seed facts, such as a post-game news, specific play, score, or event attribute, and adds multi-hop narrative constraints that uniquely identify the relevant game event in the corpus. The answer is a short factual item, such as a player number, shot placement, or score. This makes brute-force lookup impractical across 7,430 games to encourage the agent to explore over a large-scale multimodal corpus, while the short-answer format supports reliable correctness judgments.

Data and construction. The corpus covers 7,430 basketball, hockey, and soccer games, with 26,448 statistical documents, 6,859 game reports, and $\sim$1.8M video clips ($\sim$5,670 broadcast hours). Questions require evidence from multiple sources. The final evaluation set contains 1,000 questions, balanced across sports.

Evaluation. We use an LLM judge (GPT-5.2) to assess accuracy by comparing the agent's response to the ground-truth answer. In addition to the default setting where the agent operates over raw video, we introduce an *oracle mode* in which the agent operates on ground-truth textual descriptions of each clip's content, effectively providing perfect visual information without requiring the model to actually watch video. This isolates reasoning and planning ability from visual perception.

Baselines. We evaluate five agent orchestrators spanning one closed-source model (GPT-5.2) and four open-source models (Qwen3-235B, Qwen3-Omni-30B, Qwen3-32B, and MiniMax-M2.5). All models use the same system prompt and tool access, including search

Model	Default ↑			Oracle ↑			Δ (avg)
	B	H	S	B	H	S	
GPT-5.2	4.5	5.7	3.6	**48.8**	**55.0**	**58.3**	+49.4
MiniMax M2.5	3.9	7.8	3.9	38.3	41.4	39.9	+34.7
Qwen3-235B	2.7	4.5	3.3	15.9	19.8	24.9	+16.7
Qwen3-32B	2.7	2.7	3.6	6.6	12.0	17.7	+9.1
Qwen3-Omni-30B	1.8	1.5	3.0	4.8	8.7	14.4	+7.2

Table 15. T9 cross-corpus agentic reasoning results. Accuracy (%) on the 1,000-question test set, broken down per sport (B = basketball, H = hockey, S = soccer; 334 / 333 / 333 questions). *Default*: full visual pipeline. *Oracle*: ground-truth event captions replace video frames. Δ (avg): oracle gain averaged across the three sports.

and question-answering tools over both documents and videos. We report each model in both default and oracle mode to separate visual perception errors from failures in search, planning, and multi-step reasoning.

Main results. Table 15 reports overall and per-sport accuracies. In the default setting where the agent analyzes raw video, even the strongest model (GPT-5.2) achieves only 4.6% accuracy averaged across the three sports, and the smaller Qwen models fare worse (Qwen3-Omni-30B: 2.1%). Under oracle mode, where ground-truth textual descriptions replace raw video, GPT-5.2 reaches 54.0% and MiniMax M2.5 reaches 39.9%, while Qwen 3-Omni-30B achieves only 9.3%. The frontier-vs.-smaller-model gap is consistent with patterns reported on web-search agentic benchmarks [45]. The improvement from 4.6% to 54.0% confirms that visual perception is a major bottleneck. However, the 54.0% oracle ceiling also shows that multi-step planning, cross-modal reasoning, and evidence integration remain unsolved even with perfect visual descriptions, with GPT-5.2 still failing on roughly half the questions.

Tool usage patterns. Tool-use behavior varies across models and correlates with accuracy. GPT-5.2 and MiniMax M2.5 each make approximately 21 tool calls per question in default mode—roughly 3–5× more than the smaller Qwen models (3–8 calls). The frontier models perform deeper search iterations and verification before committing to an answer, while the smaller models terminate early with insufficient evidence. Under oracle mode, GPT-5.2's total drops to ∼14 calls, with most of the reduction coming from `video_qa` calls. Accurate captions remove the need for repeated visual verification. Conversation length follows the same pattern: for GPT-5.2, successful answers require ∼13 turns on average, while failures require ∼21 turns. Longer trajectories reflect unproductive search rather than deeper reasoning. Full breakdowns are in Appendix E.1.

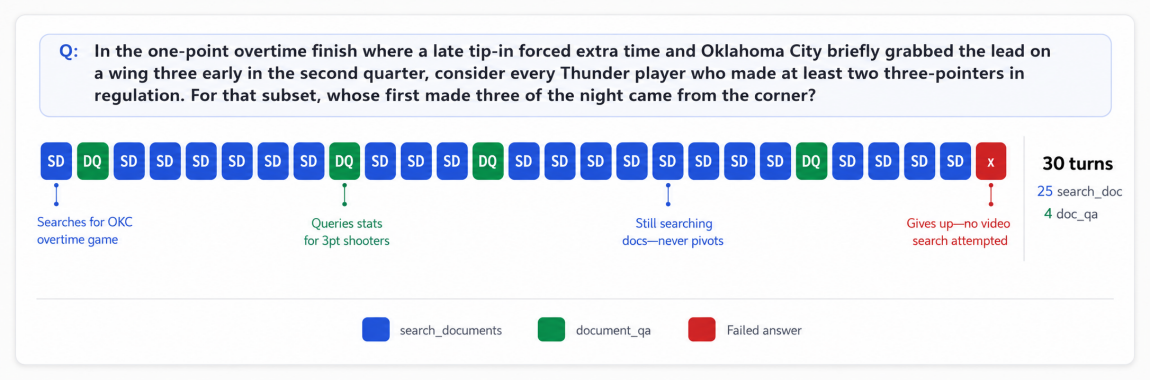


Figure 20. T9 modality-switch failure. The agent (GPT-5.2) must identify a game from narrative clues, then determine a player's shot type from video. The timeline shows 25 of 30 tool calls as document searches (one dominant color), with zero video searches attempted. After 17 turns of refined document queries, the agent gives up. The relevant clip is in the corpus.

Judge reliability. We validate the GPT-5.2 judge through three checks. First, two annotators independently re-judge 150 model answers stratified by sport and model. The LLM judge matches human verdicts on 97.3% of items on average, with inter-annotator agreement κ=0.97. Second, we re-judge a 300-response sample with three alternative LLM judges and find that model rankings remain highly stable (mean pairwise Spearman ρ=0.93). Third, we compare GPT-5.2 judgments against the mean of the non-GPT judges on GPT-5.2 outputs, finding no evidence that the primary judge favors GPT-generated answers. Full cross-judge and self-preference breakdowns are reported in Tables 42 and 43 in the appendix. Together, these checks support the reliability of our automatic evaluation. The high agreement is consistent with T9's hard-to-find but easy-to-verify design, where short factual answers admit clear binary judgments.

Qualitative example. Figure 20 shows a representative failure mode in which the agent (GPT-5.2) reasons within the wrong modality. The agent correctly identifies the target game from narrative clues using document search, but never pivots to the video modality where the answer (a player's shot type) actually resides. Across 17 turns it issues 30 tool calls—25 document searches, 5 document QA, and zero video searches—then gives up despite the relevant clip being in the corpus. This pattern, which we observe in multiple failure trajectories, suggests that even frontier agents lack a robust mechanism for recognizing when document evidence is exhausted and visual evidence is required.

> **PILLAR 4 TAKEAWAY — AGENTIC SYNTHESIS**
>
> Models reach only 5% accuracy in default mode. Under oracle mode, where ground-truth descriptions replace raw video, accuracy reaches just 54%, showing that perception and multi-step reasoning are each independently limiting.

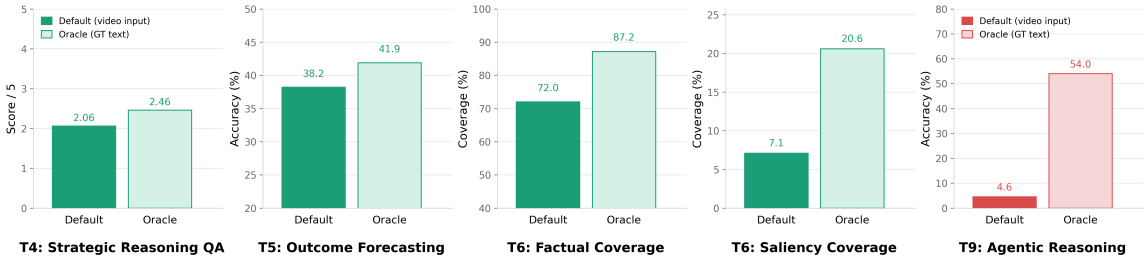


Figure 21. Oracle performance on reasoning and agentic tasks (T4, T5, T6, T9). The oracle variant replaces video with ground-truth textual descriptions of game events. All tasks use GPT-5.2. Gains are small on T4 and T5, moderate on T6, and largest on T9, indicating that strategic reasoning, forecasting, saliency judgment, and multi-step planning remain distinct bottlenecks.

5 Cross-Task Analysis

The four pillars use different metrics and evaluation protocols. This section analyzes per-task results to characterize the overall trends in performance from perception (T1–T3) through agentic synthesis (T9). Section 5.1 quantifies the performance cliff across the four pillars. Section 5.2 analyzes the effect of perception on higher-level reasoning capabilities. Section 5.3 compares model accuracy to human baselines on three tasks.

5.1 The Performance Cliff

Figure 2 plots the best-model result per pillar, with each pillar's primary metric normalized to a 0–100% range. The dashed line marks the drop from the strongest perception result (T2, fine-tuned LLaVA-Video-7B) to agentic synthesis (T9, GPT-5.2), with reasoning and simulation tasks in between. The performance drop is consistent across models within each pillar, and gains from task-specific fine-tuning at the perception level do not carry to higher pillars. This suggests that current systems can see dynamic multi-agent worlds far better than they can reason about, simulate, or plan within them.

5.2 Oracle Experiments

The per-task oracle baselines (§4.2–§4.4) replace video input with ground-truth textual descriptions of game events, derived from play-by-play logs. Here we compare them across tasks to isolate the contribution of perception. For each task we report the same model in default mode (video inputs) and oracle mode (ground-truth visual descriptions), summarized in Figure 21.

On T4, oracle access yields a 0.40-point gain (2.06 to 2.46 on the 0–5 score). On T5, oracle access yields a 3.7-point gain (38.2% to 41.9%). On T6, oracle factual accuracy rises by 15 points (71.99% to 87.19%), while oracle saliency rises to only 20.60%. On

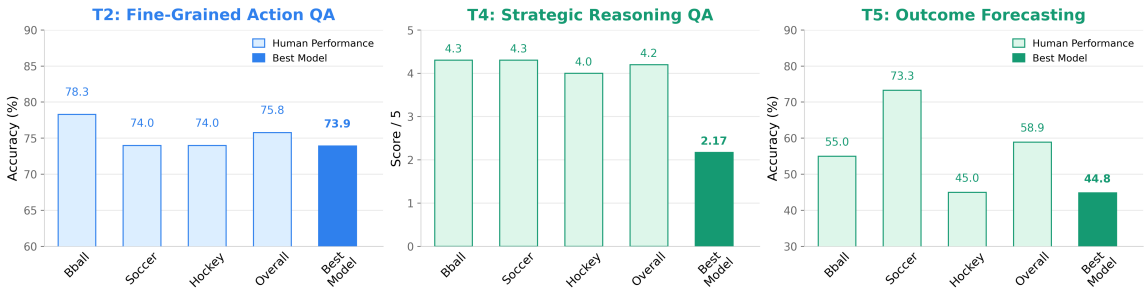


Figure 22. Human–model comparison on T2, T4, and T5. Bars show per-sport and overall human performance alongside the best model on each task. T2 and T5 use multiple-choice accuracy (%). T4 uses the open-ended 0–5 score. Best model: T2 = fine-tuned LLaVA-Video-7B, T4 = Gemini 3.1 Pro, T5 = fine-tuned Qwen3-VL-8B. Models nearly match humans on perception but trail them substantially on strategic reasoning and forecasting.

T9, oracle access raises accuracy from 4.6% to 54.0%.

The contribution of perception varies across tasks. Within T6, oracle access closes most of the factual gap but only a small fraction of the saliency gap. Oracle access yields substantial gains only on T9 and moderate gains on T6 factual recall, with minimal improvement elsewhere. No single capability accounts for the performance gap. Strategic reasoning (T4), forecasting (T5), saliency judgment (T6), and multi-step planning (T9) each limit performance independently.

5.3 Human Studies

We complement the per-task human evaluations with a cross-task comparison on T2 (perception), T4 (strategic reasoning), and T5 (forecasting). Participants have 5–10 or more years of experience in their sport and use the same inputs and response format as models. Figure 22 reports per-sport human results alongside the best model on each task.

Humans not only outperform models but also know when they are uncertain. On T2, human accuracy rises from 30% at low confidence to 90% at high confidence. On T5, it rises from 50% to 100%. Models do not show this pattern. On T5, GPT-5.2 reports similar confidence on incorrect and correct answers, producing a 28-point gap between average confidence and average accuracy.

On accuracy alone, models nearly match humans on perception (T2: 75.8% vs. 73.9%) but trail substantially on strategic reasoning (T4: 4.2/5 vs. 2.17/5) and forecasting (T5: 58.9% vs. 44.8%). The human-model gap mirrors the performance cliff: smallest on perception, widening on strategic reasoning and forecasting. Even under these small samples, experts consistently and substantially exceed the best models on strategic reasoning and forecasting, indicating these tasks are answerable by domain experts and that the gap reflects current model limitations rather than ill-posed questions.

6 Conclusion

We introduced SVI-Bench, the first large-scale benchmark for strategic video intelligence in real-world multi-agent video environments. Spanning 9 tasks across four pillars, our evaluation reveals a consistent degradation pattern: models reach 73.91% on fine-grained perception (T2) but fall to 4.6% on agentic synthesis (T9), with reasoning and simulation in between. Even given perfect visual information through oracle access, the strongest model reaches only 54% on the agentic task, indicating that the challenge extends beyond perception to reasoning, planning, and evidence integration.

Scope and domain generality. SVI-Bench is framed as a team-sports microworld—a controlled proxy for real multi-agent video—rather than a claim of cross-domain generalization. Team sports is the natural setting for this microworld. Verifiable causal ground truth is substantially harder to obtain in domains such as traffic, surgery, or robotics, while sports preserves the multi-agent complexity that defines those target domains, including dense interaction, occlusion, and long temporal horizons. The microworld nonetheless has sports-specific properties that other domains may not share, such as broadcast camera conventions, fixed rules, known team and player roles, and standardized commentary. Testing which of our findings transfer beyond sports is future work.

Limitations. Our task instances are LLM-generated but grounded in human- and league-derived primary sources—play-by-play logs, official statistics, broadcast commentary, and journalist-written reports—with manual verification on a representative subset of every task. Several tasks rely on LLM judges at evaluation time. We mitigate this through multi-judge robustness checks (Spearman ρ from 0.70 to 0.93 across four independent judges on the LLM-judged tasks) and human-agreement studies on the factual and saliency protocols, though judge bias remains a potential confound.

Future directions. The performance gaps revealed by SVI-Bench point to three research areas of broad current interest. The first is video models with stronger reasoning capabilities, able to ground long-form analysis in observed visual evidence (T4–T6). The second is generative video models with explicit notions of multi-agent dynamics, capable of producing goal-directed action sequences (T7–T8). The third is multimodal agents that can plan, retrieve, and reason across video and document corpora at scale (T9). Progress along any of these directions would expand the strategic video intelligence capabilities that intelligent systems will need in complex multi-agent environments.

Acknowledgements

This work was supported by Laboratory for Analytic Sciences via NC State University, ONR Award N00014-23-1-2356, Sony Focused Research award, Northeastern University startup funds, and the President Joseph E. Aoun Chair.

References

[1] Michael A. Alcorn and Anh Nguyen. baller2vec: A multi-entity transformer for multi-agent spatiotemporal modeling. *arXiv preprint arXiv:2102.03291*, 2021. URL https://arxiv.org/abs/2102.03291.

[2] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812, 2017.

[4] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[5] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations (ICLR)*, 2020.

[6] Daniel Bear, Elias Wang, Damian Mrowca, Felix Brock, Hsiao-Yu Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, and Fan-Yun Sun. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[7] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023.

[9] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, and Matthew Lai. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

[10] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 961–970, 2015.

[11] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. In *arXiv preprint arXiv:1907.06987*, 2019.

[12] 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*, 111(514):585–599, 2016. doi: 10.1080/01621459.2016.1141685.

[13] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2025. URL https://arxiv.org/abs/2402.03216.

[14] Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. Summscreen: A dataset for abstractive screenplay summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8602–8615, 2022.

[15] Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, and Weiyao Lin. Mecd: unlocking multi-event causal discovery in video reasoning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[16] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*, 2023.

[17] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM think like Holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

[18] Anthony Cioppa, Silvio Giancola, Adrien Deliège, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3491–3502, 2022.

[19] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9921–9931, 2023.

[20] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam Jamali Seber, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4508–4519, 2021.

[21] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

[22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[23] Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

[24] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1711–1721, 2018.

[25] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations (ICLR)*, 2024.

[26] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, 2018.

[27] David Ha and Jürgen Schmidhuber. World models. In *arXiv preprint arXiv:1803.10122*, 2018.

[28] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations (ICLR)*, 2021.

[29] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[30] Junhao He et al. Autolecture: Automatic lecture summarization with large language models. *arXiv preprint arXiv:2311.08414*, 2023.

[31] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[32] Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, February 2017. ISSN 1077-3142. doi: 10.1016/j.cviu.2016.10.018. URL http://dx.doi.org/10.1016/j.cviu.2016.10.018.

[33] Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. Bimba: Selective-scan compression for long-range video question answering, 2025. URL https://arxiv.org/abs/2503.09590.

[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[35] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, 2017.

[36] Yao-Chih Lee, Erika Lu, Sarah Rumbley, Michal Geyer, Jia-Bin Huang, Tali Dekel, and Forrester Cole. Generative omnimatte: Learning to decompose video into layers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[37] Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. TimeBlind: A spatio-temporal compositionality benchmark for video LLMs. *arXiv preprint arXiv:2602.00288*, 2026.

[38] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[39] Minghao Li, Feifan Song, Bowen Yu, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A comprehensive benchmark for tool-augmented llms. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[40] Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12112–12123, 2025.

[41] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10508–10518, 2020.

[42] Zhengyang Liang, Yan Shu, Xiangrui Liu, Minghao Qin, Kaixin Liang, Nicu Sebe, Zheng Liu, and Lizi Liao. Video-browser: Towards agentic open-web video browsing, 2026. URL https://arxiv.org/abs/2512.23044.

[43] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2023.

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023. URL https://arxiv.org/abs/2304.08485.

[45] Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, Jiayuan Song, Zhengmao Zhu, Wenhu Chen, Pengyu Zhao, and Junxian He. Webexplorer: Explore and evolve for training long-horizon web agents, 2025. URL https://arxiv.org/abs/2509.06501.

[46] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024. URL https://arxiv.org/abs/2403.00476.

[47] Nuo Luo, Zichen Cao, Hiroshi Mamitsuka, and Shanfeng Zhu. From tracking data to play prediction: A deep learning approach for basketball possession outcomes. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1866–1876. ACM, 2020. doi: 10.1145/3394486.3403055.

[48] Wan-Duo Kurt Ma, J.P. Huang, Nikolai Kolkin, et al. Trailblazer: Trajectory control for diffusion-based video generation. *arXiv preprint arXiv:2401.00896*, 2024.

[49] Karttikeya Mangalam, Harshayu Girase, Sally Aber, Jitendra Lee, and Jitendra Malik. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 759–776, 2020.

[50] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[51] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: A benchmark for general ai assistants. *International Conference on Learning Representations (ICLR)*, 2024.

[52] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023.

[53] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[54] Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, Cordelia Schmid, and Tobias Weyand. MINERVA: Evaluating complex video reasoning. *arXiv preprint arXiv:2505.00681*, 2025.

[55] Koichi Namekata et al. Sg-i2v: Self-guided trajectory control in image-to-video generation. In *arXiv preprint arXiv:2411.04989*, 2024.

[56] NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[57] OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

[58] Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. *arXiv preprint arXiv:2511.05489*, 2025.

[59] Yulu Pan, Ce Zhang, and Gedas Bertasius. Basket: A large-scale video dataset for fine-grained skill estimation. In *CVPR*, 2025.

[60] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay summarization using latent narrative structure. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1920–1933, 2020.

[61] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to

master 16000+ real-world apis. *International Conference on Learning Representations (ICLR)*, 2024.

[62] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[63] Hanoona Rasheed, Mohammed Zumri, Muhammad Maaz, Ming-Hsuan Yang, Fahad Shahbaz Khan, and Salman Khan. Video-com: Interactive video reasoning via chain of manipulations. *arXiv preprint arXiv:2511.23477*, 2025.

[64] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.

[65] Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. *arXiv preprint arXiv:2303.07109*, 2023. doi: 10.48550/arXiv.2303.07109. URL https://arxiv.org/abs/2303.07109.

[66] Anna Rohrbach, Marcus Rohrbach, Niket Tanber, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3212, 2015.

[67] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision (ECCV)*, pages 683–700, 2020.

[68] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

[69] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, and Thore Graepel. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[70] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, and Adrian Bolton. Mastering the game of go without human knowledge. *Nature*, 550(7676): 354–359, 2017.

[71] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations (ICLR)*, 2023.

[72] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. URL https://arxiv.org/abs/1212.0402.

[73] Chen Sun, Per Karlsson, Jiajun Wu, Joshua B Tenenbaum, and Kevin Murphy. Stochastic prediction of multi-agent interactions from partial observations. In *International Conference on Learning Representations (ICLR)*, 2019.

[74] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[75] Team Wan. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

[76] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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

[78] Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation. *arXiv preprint arXiv:2505.22944*, 2025.

[79] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. 2023. URL https://arxiv.org/abs/2305.16291.

[80] Jiawei Wang et al. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024.

[81] Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025.

[82] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.

[83] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.

[84] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents, 2025. URL https://arxiv.org/abs/2504.12516.

[85] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

[86] 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. *arXiv preprint arXiv:2511.06499*, 2025.

[87] Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, et al. SPORTU: A comprehensive sports understanding benchmark for multimodal large language models. In *International Conference on Learning Representations*, 2025.

[88] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. URL https://arxiv.org/abs/2105.08276.

[89] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.

[90] Jinglin Xu, Guohao Zhao, Sibo Yin, Wenhao Zhou, and Yuxin Peng. Finesports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In *CVPR*, 2024.

[91] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016.

[92] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *eprint arXiv: 2407.10671*, 2024.

[93] Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. Longvt: Incentivizing "thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*, 2025.

[94] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

[95] 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 *International Conference on Learning Representations (ICLR)*, 2020.

[96] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

[97] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering, 2019. URL https://arxiv.org/abs/1906.02467.

[98] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023.

[99] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering, 2024. URL https://arxiv.org/abs/2312.17235.

[100] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 543–560, 2023.

[101] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning, 2025. URL https://arxiv.org/abs/2508.04416.

[102] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[103] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 5905–5921, 2021.

[104] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding, 2025. URL https://arxiv.org/abs/2406.04264.

[105] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[106] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. *International Conference on Learning Representations (ICLR)*, 2024.

Appendix Overview

This appendix provides the full construction details, evaluation protocols, and complete results that the main paper summarizes, organized to mirror the structure of the paper.

- **Appendix A** documents the data engine and per-sport corpus statistics.
- **Appendix B.1–B.3** cover the perception tasks (T1–T3).
- **Appendix C.1–C.3** cover the reasoning tasks (T4–T6).
- **Appendix D.1–D.2** cover the simulation tasks (T7–T8).
- **Appendix E.1** covers the agentic task (T9).

For each task we provide the construction pipeline, prompt templates, per-sport statistics, full per-model results, and qualitative examples. For the LLM-judged tasks (T1, T4, T6, T9), the corresponding sections additionally report judge-reliability analyses, including human-agreement studies and cross-judge robustness.

A Data Engine

A.1 Data Sources and Scale

The SVI-Bench corpus is built from game-replay archives, league statistics providers, and sports media outlets. We collect full-game broadcasts, structured event annotations, and textual game coverage across three sports, 64 leagues, and seven years of competition (2018–2025).

Basketball. We gather full-game videos from 17 leagues (including 3 women's leagues) spanning 2018–2024, totaling approximately 25K games. Leagues range from the NBA and NCAA Division I to international competitions such as EuroLeague and EuroBasket. Women's leagues account for approximately 6K games.

Hockey. We collect approximately 35K games from 44 leagues across North America and Europe over 2019–2024, covering both professional (e.g., NHL, Liiga, DEL) and developmental leagues (e.g., NCAA Division I).

Soccer. We collect 4K games from three leagues—the English Premier League, Spain's La Liga, and NCAA Division I Soccer—spanning 2021–2025. Soccer's smaller league count is offset by long match durations (90+ minutes per game) and high inter-agent density (22 players), making per-game evidence density comparable to the other sports.

Multimodal resources. Each game in the corpus includes at least three core resources:

- **Full-game video**: the complete broadcast recording, typically 1.5–3 hours depending on the sport.
- **Play-by-play log**: a timestamped sequence of annotated events (e.g., shots, passes, fouls, goals) aligned to the video timeline, sourced from league statistics providers.
- **Structured metadata**: team- and player-level box-score statistics aggregated from the play-by-play log, broken down by period.

For leagues with high-quality written coverage and professional broadcasts, we additionally collect two supplementary modalities:

- **Game reports**: professionally written post-game summaries from sports journalism outlets, aligned to the corresponding game via identifiers (teams, date, final score). Coverage spans the NBA and NCAA Division I (basketball), the NHL (hockey), and the EPL and La Liga (soccer).
- **Expert commentary**: transcripts extracted from professional broadcast audio using Whisper [62] automatic speech recognition. We cover the NBA, NCAA Division I, and EuroLeague (basketball); the NHL (hockey); and the EPL and La Liga (soccer).

A.2 Data Construction Pipeline

This section provides implementation details for our four-stage data engine.

Entity resolution details. Player names in commentary and game reports are matched against league roster databases using a combination of exact-match, fuzzy-string, and contextual disambiguation (e.g., resolving *"Curry"* to Stephen or Seth based on team context). Resolved entities are organized into identity graphs that capture both static attributes (position, team, height, dominant hand) and dynamic relationships (teammate, opponent, matchup) for each game. These identity graphs are reused across tasks: the same canonical player IDs underlie T1 captions, T2 question generation, T4 commentary-to-question grounding, and T9 agent search.

LLM-assisted instance generation details. For each task, the LLM receives the relevant slice of the corpus—play-by-play for short-horizon tasks, play-by-play plus commentary plus reports for full-game tasks—and generates instances under a sport-specific prompt template. The LLM operates over human-derived primary evidence. It does not itself produce the ground-truth labels, which are anchored in the play-by-play logs, official statistics, or journalist reports. Per-task prompts are included in §B.1–§E.1.

A.3 Quality Control

Automated validity checks. We cross-check question–answer consistency against the play-by-play log and structured metadata. For example, a T2 question asking which

player took a shot is automatically discarded if the play-by-play log does not record any shot in the cited 10-second window. We additionally apply an LLM-based filter to remove samples with strong language bias, formatting issues, or trivially obvious answers. Task-specific filters and shortcut mitigations are described in their respective sections (§B.1–§E.1).

Human expert review. Domain-knowledgeable annotators review a stratified subset of instances—sampled to ensure coverage across all sports, pillars, tasks, and question types—to verify that each instance is (i) correctly grounded in the available evidence, and (ii) linguistically clear and natural. Instances failing either criterion are removed, and systematic error patterns identified during review are used to refine the automated filters applied to the full dataset.

B Pillar 1: Dynamic Scene Understanding (T1–T3)

B.1 T1: Structured Play Description

T1: Data Construction and Statistics

We align each 10-second video segment with the play-by-play log and generate an initial caption using a structured template: *"Player X ⟨identity⟩ does ⟨action⟩ with ⟨attributes⟩."* Attributes include spatial information and fine-grained action details (play type, shot type, etc.). All actions within the time window are concatenated in sequential order, followed by game-state context (teams and current score). To improve linguistic diversity and grammatical quality, we use GPT-4o mini to polish these template-based captions into fluent natural language using the following prompt:

```
Polish this ⟨Sport⟩ game clip caption. Output only the improved caption---no
explanations or extra text. Preserve every specific detail from the original,
including shot type, hand used, assist, defender, play style, teams, and current
score.
```

Unlike prior datasets that emphasize broad, high-level captions, our captions are detail-oriented: they explicitly describe all players involved in each event and include the contextual details needed to interpret the play at an expert level.

In total, we construct 500K video–caption pairs per sport (1.5M across basketball, hockey, and soccer). For practical training and evaluation, given computational cost and API usage constraints, we use a subset: 280K samples for training (100K basketball, 80K soccer, 100K hockey), 3K for validation (1K per sport), and 3K for testing (1K per sport). Figure 23 illustrates the distribution of word lengths in the structured sports de-

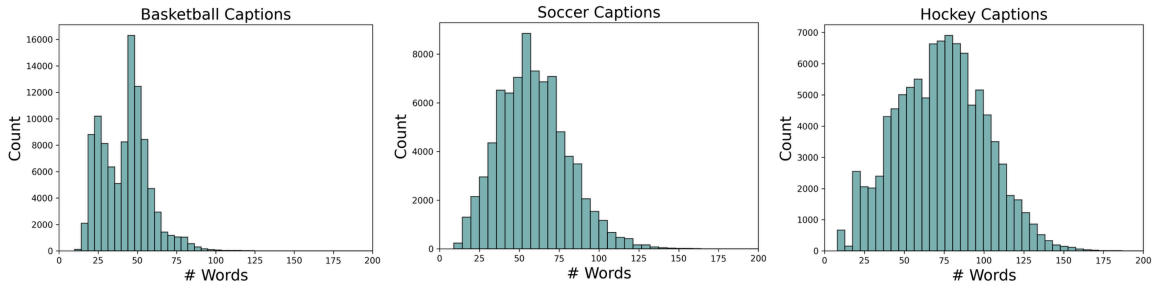


Figure 23. T1 caption length distribution (in words) across basketball, soccer, and hockey.

scriptions across the three sports. On average, basketball captions contain 42 words, soccer captions 59 words, and hockey captions 73 words.

T1: Evaluation Protocol

We adopt an LLM-as-a-judge protocol to evaluate generated captions. Figure 24 shows the full prompt. For each prediction, the judge assigns Likert scores from 0–5 along six axes: action accuracy, identity accuracy, causality/outcome, spatial understanding, temporal understanding, and contextual details. These axes capture both factual correctness and coverage of fine-grained visual details. Our primary metric, Average, is the mean of the six axis scores. During evaluation, the judge is instructed to ground its decisions in verifiable visual evidence and to penalize hallucinated or unsupported details. We use GPT-5.2 as the primary judge.

T1: Baselines and Training Details

We evaluate frontier proprietary models (GPT-5.2, Gemini-3-Flash) and open-source models (LLaVA-Video-7B, Qwen3-VL-32B). During inference, we provide two in-context examples to demonstrate the expected structure, style, and level of detail. In addition to few-shot evaluation, we fine-tune LLaVA-Video-7B on our dataset. The model is trained on a combination of 280K structured play descriptions (100K basketball, 80K soccer, 100K hockey) and 300K fine-grained action QA samples (100K per sport), yielding 580K total training samples. For each video, we uniformly sample 16 frames at 336×336 resolution. Optimization uses AdamW (lr $= 1\times10^{-5}$, β=$(0.9, 0.98)$, weight decay 0.05) with a cosine learning rate schedule and BFloat16 mixed-precision training. Training runs for 1 epoch on 4$\times$ NVIDIA H100 GPUs with a per-GPU batch size of 2.

T1: Analysis

Human–judge agreement methodology. The 0.40 human–judge MAE reported in the main paper is computed on 60 randomly sampled captions independently scored by three

LLM-as-a-judge Prompt for T1 (Structured Play Description)

```
You are an expert **soccer analyst** and a meticulous **video caption evaluation
assistant**. Your task is to compare a Generated Caption (Pred) against a human-annotated
Ground-Truth Caption (GT) across multiple dimensions of understanding.

You MUST follow a strict Chain-of-Thought (CoT) procedure for every category:
1. Analyze GT: Extract the relevant facts from the Ground Truth.
2. Analyze Pred: Extract the relevant facts from the Generated Caption.
3. Compare and Justify (CoT): Highlight matches, omissions (in GT but not Pred),
   contradictions, and hallucinations (in Pred but not GT). Provide clear reasoning.
4. Score: Assign 0 (completely wrong) to 5 (perfect), strictly following the rubric.

Your output must be a single, strictly formatted JSON object containing the analysis for
all categories.

### Example Evaluation Categories and Rubrics

1. Action Accuracy and Specificity
Focus: correctness of verbs, movements, action types, and attributes.
- 5: All actions and attributes match exactly in sequence and specificity.
- 4: Nearly all correct; only minor specificity mistakes.
- 3: Main actions correct; specificity or key attributes wrong/missing.
- 2: At least one major action correct; many other errors/omissions.
- 1: Vaguely related; most actions wrong or hallucinated.
- 0: All actions wrong/irrelevant.

2. Entity and Identity Accuracy
Focus: actors (players, teams) and their roles.
- 5: All identities and roles correct.
- 4: Almost all correct; minor omissions or a secondary error.
- 3: Main identities correct; secondary roles wrong/missing.
- 2: Some correct IDs, but major errors (e.g., wrong main actor or team).
- 1: Identities mentioned but mostly misidentified.
- 0: All identities wrong or missing.

(Additional categories - Causality Outcome, Spatial Understanding, Temporal Understanding,
Contextual Details, and Final Holistic Score - follow the same 0-5 rubric structure
with category-specific definitions)

### Output Format (Strict JSON)

{
  "action_accuracy": {
    "gt_analysis": "...",
    "pred_analysis": "...",
    "justification_cot": "...",
    "score": X
  },
  "identity_accuracy": { ... },
  "causality_outcome": { ... },
  "spatial_understanding": { ... },
  "temporal_understanding": { ... },
  "contextual_details": { ... },
  "final_holistic_score": { ... }
}

GT: <ground-truth caption>
Pred: <model-generated caption>
```

Figure 24. T1 LLM-as-a-judge prompt. The judge model compares a generated structured play caption to a human ground-truth caption along multiple dimensions (e.g., actions, entities, causality), and outputs a JSON report with per-category chain-of-thought analysis and 0–5 scores. The full prompt will be released with the code for reproducibility.

human annotators using the same six-axis rubric. Beyond the overall MAE, we also measure the score-band spread—the maximum gap in MAE across low (0–1), mid (2–3), and high (4–5) score bands—which is 0.21, indicating that agreement quality is approximately uniform across the scoring range and not driven by easy extremes.

Independent judge replication. We repeat the full T1 evaluation using GPT-5-mini as an independent judge instead of GPT-5.2. Model rankings are fully preserved, with GPT-5-

Judge Pair	mean A	mean B	bias	MAE	Spearman ρ
GPT-5.2 vs Gemini	1.896	2.568	-0.673	0.852	0.729
GPT-5.2 vs Qwen3	1.896	1.905	-0.009	0.625	0.692
GPT-5.2 vs DeepSeek	1.896	1.849	$+0.046$	0.613	0.731
Gemini vs Qwen3	2.568	1.905	$+0.663$	0.924	0.660
Gemini vs DeepSeek	2.568	1.849	$+0.719$	0.932	0.716
Qwen3 vs DeepSeek	1.905	1.849	$+0.056$	0.609	0.677

Table 16. Pairwise agreement between four LLM judges on the T1 0–5 scoring rubric. We report each judge's mean score, signed bias (mean A minus mean B), mean absolute error, and Spearman rank correlation.

mini assigning slightly higher scores overall (2.79 vs. 2.59 Average for the best model).

Cross-judge agreement. We compare four independent LLM judges—GPT-5.2, Gemini-3-Flash, Qwen3-235B, and DeepSeek-V3.2—to evaluate whether T1 rankings depend on judge choice. Table 16 reports pairwise agreement. The judges differ in calibration: Gemini scores systematically higher than the others, with bias up to $+0.72$ relative to DeepSeek. All pairs preserve moderately strong rank agreement, with Spearman correlations of 0.66–0.73, indicating that T1 model rankings are not driven by any single judge's scoring preference.

Self-preference check. To rule out GPT-family self-preference, we re-score GPT-5.2's generated captions with three non-GPT judges (Qwen3-235B, Gemini-3-Flash, DeepSeek-V3.2). The GPT-5.2 judge assigns GPT-generated captions 0.21 points lower on the 0–5 scale than the mean of the non-GPT judges. In other words, the GPT judge is slightly stricter on its own outputs.

Additional qualitative examples. Figure 25 shows additional predictions on soccer and hockey. The fine-tuned model captures spatial layout and temporal structure well but misattributes player identities across both sports. Hockey is the most challenging setting in our benchmark, with lower scores across nearly every axis, reflecting faster camera dynamics, smaller jersey numbers, and greater inter-player visual similarity.

B.2 T2: Fine-Grained Action QA

T2: Data Construction and Statistics

We define question types across three sports (Table 17), organized into six capability categories: *action recognition* (identifying the key event), *temporal ordering* (sequencing events within the window), *play analysis* (fine-grained interpretation, e.g., distinguishing a layup from a floater), *spatial reasoning* (relative positions and trajectories), *player identi-*

Figure 25. Additional T1 qualitative examples on soccer (top) and hockey (bottom). Each panel shows the ground-truth caption (blue), the fine-tuned LLaVA-Video-7B-ft prediction (red), and per-axis LLM judge scores (gray). The pattern observed on basketball (Fig. 5) holds across sports: spatial layout and temporal structure are captured reasonably well, but player identities and fine-grained actions are misattributed. Hockey is the most challenging setting, with low scores across nearly every axis.

fication (associating actions with specific players), and *OCR* (reading on-screen text such as scoreboards and shot clocks).

In total, we construct 500K video–question pairs per sport (1.5M total) using the data engine described in §A.2. Throughout generation, we carefully balance the answer distribution within each question type. For instance, for basketball shot-type identification, the play-by-play log distinguishes seven shot types. We ensure each appears with similar frequency, reducing label bias and preventing models from exploiting skewed priors. Due to computation constraints, we use a subset for experiments: 300K samples for training (100K per sport), 30K for validation (10K per sport), and 30K for testing (10K per sport).

T2: Evaluation Protocol

We formulate all questions as multiple-choice and evaluate performance using accuracy—the percentage of correctly answered questions. We report both overall accuracy and breakdowns per sport and per question type to analyze performance across different sports and reasoning categories.

T2: Baselines and Training Details

We evaluate both frontier proprietary models (GPT-5.2, Gemini-3-Flash) and open-source video-language models (LLaVA-Video-7B, Qwen3-VL-32B) in a zero-shot setting. We additionally evaluate a fine-tuned LLaVA-Video-7B model. This model is trained on a mixture of 280K structured play descriptions and 300K fine-grained action QA samples (580K total), producing a single model shared with T1 (see training details in §B.1).

T2: Analysis

Below we provide additional ablations, the full human study details, and additional qualitative examples.

Effect of caption supervision. The fine-tuned model is trained jointly on T1 captions (280K samples) and T2 multiple-choice QA (300K samples). Removing the captions and training on QA alone drops T2 accuracy from 73.91% to 67.70% overall. Table 18 shows this holds across all sports: basketball drops from 73.43% to 65.05%, soccer from 75.48% to 67.59%, and hockey from 72.83% to 70.45%. The effect is largest on basketball and soccer, where captions carry the most fine-grained spatial and identity information.

Cross-sport transfer. Joint multi-sport training matches or outperforms single-sport fine-tuning. Training only on basketball (100K QA + 100K captions) gives 71.67% on basketball, whereas joint training on all sports reaches 73.43% on basketball. The same holds for soccer (75.08% single-sport vs. 75.48% joint) and hockey (72.49% vs. 72.83%). Visual–temporal primitives such as passes, shots, and action sequences transfer across sports, and the additional cross-sport training data provides a small but consistent gain.

Effect of training data scale. Within the multi-sport QA+caption setting, scaling from 25K to 50K to 100K samples per sport yields monotonic improvements across all sports: 65.24% $\rightarrow$ 69.69% $\rightarrow$ 73.91% overall. We do not observe saturation at the largest scale tested.

Human study. We recruited participants with at least five years of experience in their respective sports. Each participant answered a subset of 50 T2 multiple-choice questions and provided a confidence rating from 1 to 3. Six participants completed the basketball subset and achieved 78.33% accuracy, three participants completed the soc-

cer subset and achieved 74.00%, and one participant completed the hockey subset and achieved 74.00%. Human accuracy increased monotonically with self-rated confidence: low-confidence responses were 30% correct, moderate-confidence responses 75%, and high-confidence responses 90%, indicating that participants' uncertainty was well-calibrated and that they were aware when they encountered a perception bottleneck.

Additional qualitative examples. Figure 26 shows additional T2 examples on basketball (temporal ordering) and hockey (play analysis). Together with the main-paper soccer example (Fig. 6), the three cases span three distinct capability categories and illustrate where fine-tuning helps and where it does not. Fine-tuning provides the most reliable gains on technique discrimination and action-sequence ordering. Player identification remains difficult for every evaluated model regardless of training.

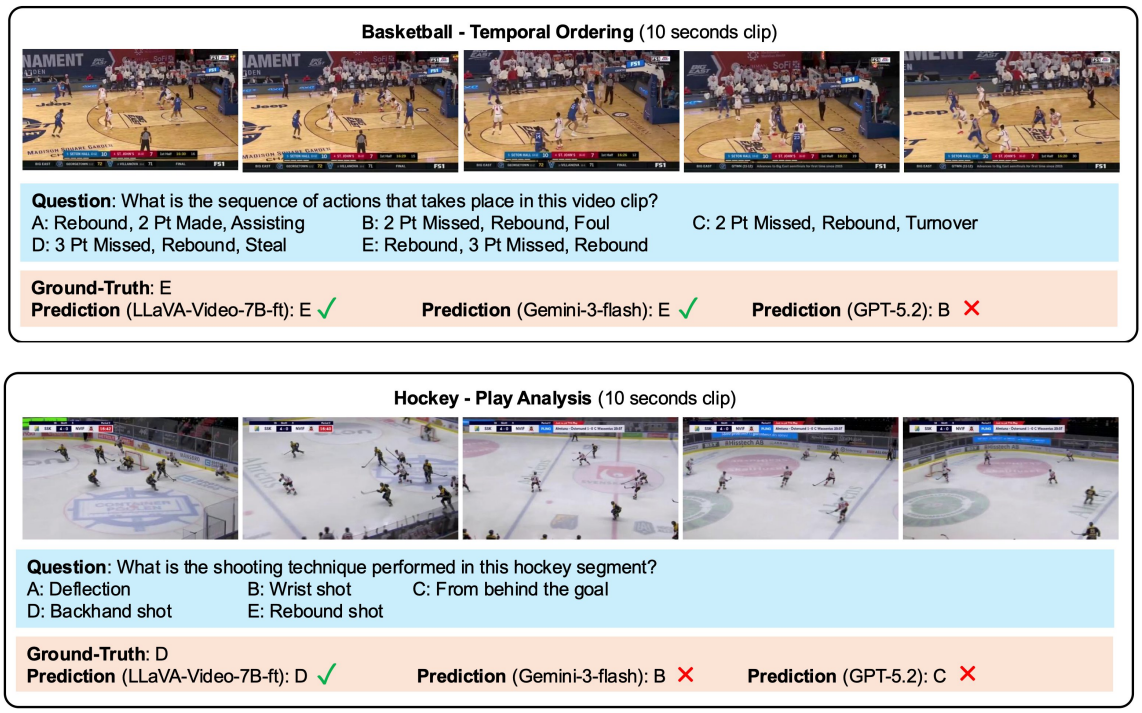


Figure 26. Additional T2 qualitative examples. For the basketball clip, the fine-tuned LLaVA-Video-7B and Gemini-3-Flash correctly identify the action sequence (Rebound, 3 Pt Missed, Rebound), while GPT-5.2 selects a plausible but incorrect ordering. For the hockey clip, only the fine-tuned model correctly recognizes the backhand shot technique. GPT-5.2 and Gemini-3-Flash confuse it with visually similar alternatives (wrist shot, from behind the goal).

	Question Type	Category
Shared across all sports (6 types)		
1	Atomic Action Recognition	Action recog.
2	Action Sequence	Temporal ordering
3	Player Name Identification	Player identif.
4	Participating Teams Identification	OCR
5	Current Period Identification	OCR
6	Spatial Understanding	Spatial reasoning
Basketball-specific (12 types)		
7	Shot Type Identification	Play analysis
8	Dribble Move Identification	Play analysis
9	Play Type Identification	Play analysis
10	Contested Shot Identification	Play analysis
11	Drive Direction Identification	Play analysis
12	Shooting Hand Identification	Play analysis
13	Player Number Identification	Player identif.
14	Player Position Identification	Player identif.
15	Player Skill Level Identification	Player identif.
16	Scoreboard Recognition	OCR
17	Shot Clock Recognition	OCR
18	Remaining Time Recognition	OCR
Hockey-specific (4 types)		
19	Penalty Type Identification	Play analysis
20	Goalie Stance Recognition	Play analysis
21	Shot Type Recognition	Play analysis
22	Player Number Identification	Player identif.
Soccer-specific (9 types)		
23	Fine-Grained Action Recognition	Action recog.
24	Pass Outcome Identification	Play analysis
25	Pass Height Identification	Play analysis
26	Attack Flank Direction	Spatial reasoning
27	Shot Expected Goal Estimation	Play analysis
28	Shot Body Part Identification	Play analysis
29	Shot Outcome Recognition	Play analysis
30	Player Position Identification	Player identif.
31	Remaining Time Recognition	OCR

Table 17. T2 question type taxonomy (31 types total). Each type is assigned to one of six capability categories: action recognition, temporal ordering, play analysis, spatial reasoning, player identification, and OCR. Six types are shared across all sports; the remainder are sport-specific.

Training Configuration	Basketball	Soccer	Hockey	Overall
Zero-shot baseline				
No training (LLaVA-Video-7B)	35.76	38.72	36.56	37.01
Single-sport training (100k QA + 100k Caption)				
Basketball only	71.67	–	–	–
Soccer only	–	75.08	–	–
Hockey only	–	–	72.49	–
All-sport training: data type				
QA only	65.05	67.59	70.45	67.70
QA + Caption	**73.43**	**75.48**	**72.83**	**73.91**
All-sport training: data scale (QA + Caption)				
25k per sport	62.05	70.06	63.61	65.24
50k per sport	68.04	72.96	68.08	69.69
100k per sport	**73.43**	**75.48**	**72.83**	**73.91**

Table 18. T2 ablation study on training data composition. Accuracy (%) of LLaVA-Video-7B under different training configurations. *Zero-shot baseline* reports the untrained model. *Single-sport training* fine-tunes on one sport (100K QA + 100K captions) and evaluates on that sport. *Data type* compares training on QA alone against the joint QA + caption mixture, both on all three sports. *Data scale* varies the per-sport sample count (25K, 50K, 100K) under the joint QA + caption setting. Best entries are bolded. The QA + caption configuration at 100K per sport is the model used throughout the paper.

B.3 T3: Compositional Video Retrieval

T3: Data Construction and Statistics

Table 19 summarizes the T3 data statistics. The benchmark contains approximately 291K training samples and 15K evaluation queries. Each evaluation query is paired with one positive video and 5,000 negatives. The test and validation splits contain 72K clips in total, with a mean clip duration of $\sim$10 s. Each sport contributes 1,000 validation and 4,000 test queries. Within Tier 2 and Tier 3, queries are evenly distributed across five hard-negative buckets, treated as right-open intervals: [0,100), [100,200), [200,300), [300,400), and [400,500), with 70 validation and 270 test queries per bucket.

Attribute taxonomy. To construct compositional video-text pairs, each video is annotated with sport specific attributes organized into four high-level categories. *Entity* attributes identify the participants, including player names, jersey numbers, and team names. *Dynamics* attributes describe the action occurring in the clip. *Context* attributes

	Basketball	Hockey	Soccer
Training set			
Queries / clips	100,000	100,000	91,126
Hours (est.)	∼281	∼282	∼260
Validation set			
Queries	1,000	1,000	1,000
Positive clips	1,000	1,000	1,000
Negative clips	8,000	8,000	8,000
Hours	25.30	25.38	25.67
Test set			
Queries	4,000	4,000	4,000
Positive clips	4,000	4,000	4,000
Negative clips	16,000	16,000	16,000
Hours	56.23	61.23	57.10
Avg. clip length	10.12 s	10.15 s	10.28 s

Table 19. T3 dataset statistics per sport. Each training query is paired with a single clip. Each evaluation query is paired with one positive and 5,000 negative clips drawn from the same sport.

capture sport-specific conditions surrounding the action, such as play type, shooting hand, defensive pressure, goalie stance, ball possession, field flank, or playing position. *Spatial* attributes localize the event on the playing surface, such as court region in basketball or rink zone in hockey. The full sport-specific attribute taxonomy is listed in Table 20.

Query generation. For each positive video, its ground-truth attribute annotations are filled into sport-specific natural-language templates to produce a structured caption (e.g., *"Laura Spreafico performs a 3 pt missed from the wing, right-handed, on a screen off play"*). These captions are then paraphrased with GPT 5.2 to yield fluent and diverse natural-language queries (e.g., *"Laura Spreafico misses a right-handed 3-point attempt from the wing on a screen play."*).

Hard negative mining. A hard negative is a same-sport video that differs from the positive in exactly one attribute while sharing all others. This construction applies to Tier 2 and Tier 3 queries (Tier 1 queries have only one attribute). Queries are grouped into difficulty buckets by their hard-negative count—queries with more hard negatives are harder because the candidate pool contains more near-duplicates.

Training samples Training samples are built based on these annotations. Each video is paired with a full natural-language caption that describes all annotated attributes. We

Category	Sport	Attributes
Entity	Basketball	player name, jersey number, team name
	Hockey	player name, jersey number, team name
	Soccer	player name, team name
Dynamics	Basketball	action type
	Hockey	action type
	Soccer	action type
Context	Basketball	play type, shooting hand, shot type, drive, dribble move, defensive pressure, game score
	Hockey	shot type, goalie view clarity, goalie stance, penalty type
	Soccer	action detail, shot body part, goal zone, shot on target, keeper involved, pass target, duel target, ball possession, field flank, possession outcome, playing position
Spatial	Basketball	court region
	Hockey	rink zone
Total attributes per sport	Basketball: 12	Hockey: 9 Soccer: 14

Table 20. T3 attribute taxonomy. Each video is annotated with attributes from four categories: *Entity* (who is involved), *Dynamics* (what action occurs), *Context* (surrounding circumstances), and *Spatial* (where on the playing surface). Soccer has no spatial attribute as field position is captured under context (field flank).

also construct attribute-dropout caption variants by progressively removing attributes from the full caption, yielding captions at different levels of specificity. For example, the full caption "Elease Stafford, number 0, attempts a contested jumper from the wing but misses the two-point shot with the score at 8 to 6" yields variants such as "a player in number 0 attempts a contested jumper from the wing but misses the two-point shot" after dropping the player name, and "Elease Stafford, number 0, is on the court" when retaining only entity-level attributes. These variants expose the model to diverse attribute combinations and support compositional video-text training.

T3: Evaluation Protocol

We report Recall@K for $K \in \{1, 5, 10, 50, 100, ..., 500\}$, with Recall@1 as the primary metric. For the category- and composition-level analyses, we report R@100, as R@1 values are generally too low to yield meaningful comparisons.

T3: Baselines and Training Details

We fine-tune InternVideo2-Stage2 1B [83], which comprises a ViT-based vision encoder (40 transformer blocks, d_{model}=1408, 16 attention heads, patch size 14×14) and a BERT-Large text encoder. We initialize from the publicly available InternVideo2-Stage2_1B-224p-f4 checkpoint. Training uses AdamW (lr $= 1\times10^{-5}$, β=$(0.9, 0.98)$, weight decay 0.05) with cosine learning rate scheduling (minimum LR multiplier $= 0.01$, no warmup) and BFloat16 mixed precision. Only the video-text contrastive (VTC) loss is used. We sample 16 frames per clip at 224×224 resolution, using uniform random sampling during training and middle-frame sampling at test time, with linear interpolation of positional embeddings to accommodate 16-frame inputs. Training runs for 5 epochs on 8$\times$ NVIDIA A6000 GPUs with a per-GPU batch size of 2 and a maximum text length of 200 tokens. Data from all three sports are combined into a single training pool. We compare two training regimes:

- **Full caption**: each video is paired with its complete natural-language description containing all annotated attributes.
- **Full caption with attribute dropout**: for each video, one caption variant from the attribute-dropout set described above is randomly sampled at each training step.

T3: Analysis

Effect of hard negatives. Retrieval performance degrades substantially as the number of visually similar distractors in the candidate pool grows (Figure 27). The sharpest decline occurs at low hard-negative counts: moving from 0–100 to 100–200 hard negatives per query, R@100 drops from 54.8% to 23.3% in Basketball, from 56.3% to 44.1% in Soccer, and from 76.3% to 39.1% in Hockey. Beyond 200 hard negatives, performance continues to degrade but at a slower rate, plateauing at substantially lower levels. These results confirm that the model's ability to distinguish fine-grained visual differences remains fragile when many near-duplicate candidates are present.

Category-level analysis. Table 21 aggregates performance by attribute category. Entity-only compositions yield notably higher R@100 than context-only compositions (e.g., 42.1% vs. 19.4% in basketball Tier 2), likely because entity attributes often correspond to more visually distinctive cues, such as player identity, jersey numbers, team uniforms, and venue appearance. In contrast, context attributes such as play type or shooting hand are harder to distinguish among visually similar clips.

Composition-level analysis. Table 22 lists the easiest and hardest attribute compositions per sport, reinforcing the category-level trends. The easiest compositions are dominated by entity attributes. *player name + team name* achieves 89.3% R@100 in hockey and 73.4% in basketball. The hardest compositions are typically context-heavy and often

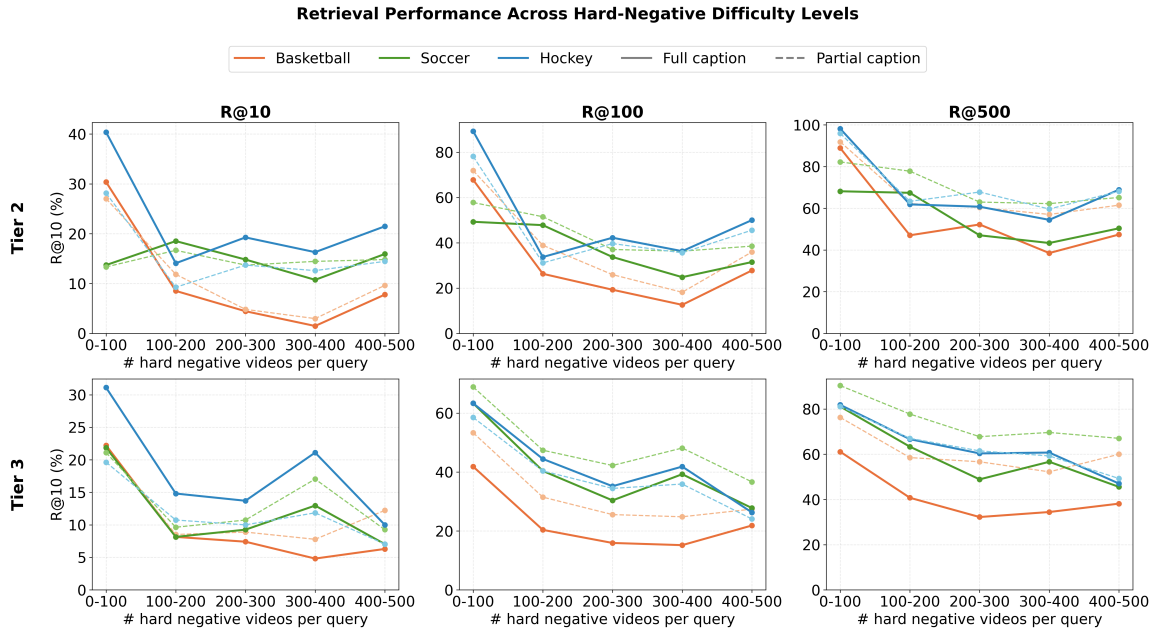


Figure 27. T3 retrieval performance as a function of hard-negative count per query. Performance degrades sharply at low hard-negative counts—e.g., R@100 drops from 76.3% to 39.1% in hockey when moving from 0–100 to 100–200 hard negatives—then plateaus at substantially lower levels, indicating that even a moderate number of visually similar distractors is sufficient to challenge current retrieval models.

include spatiotemporal attributes. In basketball, all five hardest compositions achieve 0.0% R@100 and involve attributes such as court region, play type, shot type, shooting hand, and defensive pressure. A similar pattern holds in soccer, where the hardest queries involve fine-grained contextual attributes such as playing position, possession, shot body part, pass target, and duel target. Notably, the presence of entity attributes does not guarantee easier retrieval when context attributes dominate the composition.

Additional qualitative examples. Figure 28 shows additional T3 examples on basketball and soccer. Together with the main-paper hockey example (Fig. 8), the three cases span all three sports and illustrate that the model captures broad scene context but struggles to distinguish fine-grained compositional details among visually similar clips.

Tier	Categories	Avg.	Basketball	Hockey	Soccer
Tier 1: single attribute category					
	E	24.9	20.8	28.3	22.6
	D	22.0	20.2	22.6	24.0
	C	21.9	19.3	21.1	24.1
Tier 2: two attribute categories					
	E	54.5	42.1	65.8	51.4
	E+S	39.8	22.6	54.0	50.0
	E+C	37.0	25.3	39.8	47.0
	D+S	35.7	37.5	41.2	0.0
	C+S	34.4	19.6	38.9	41.7
	E+D	33.9	25.6	38.1	40.6
	D+C	32.4	25.5	33.0	33.8
	C	27.7	19.4	0.0	29.4
Tier 3: three or more attribute categories					
	E	62.9	54.5	66.7	–
	E+D	52.1	48.0	53.9	0.0
	E+S	45.7	50.0	45.7	–
	E+D+S	41.9	45.5	41.5	–
	E+C	39.4	23.8	38.4	47.4
	E+D+C	36.8	27.8	38.3	41.0
	D+C	33.3	23.0	41.7	37.1
	E+D+C+S	32.3	24.8	37.6	75.0
	C	31.5	19.6	33.3	35.8
	E+C+S	30.3	18.9	38.6	100.0
	C+S	23.2	17.4	57.1	52.4
	D+C+S	22.8	16.9	38.2	35.3

Table 21. T3 category-level retrieval performance (R@100), sorted by average performance within each tier. E = Entity, D = Dynamics, C = Context, S = Spatial. Compositions involving entity attributes consistently rank highest, while context-only compositions rank lowest, reflecting that entity attributes produce more visually distinctive cues than context attributes.

Tier	Cat.	Composition	R@100	MedRank
Basketball — Easiest				
2	E	player name, team name	73.4	24
2	E+C	game score, team name	72.7	14
2	E+C	game score, player name	54.7	79
3	E	player name, jersey number, team name	54.5	45
2	E+C	def. pressure, team name	50.0	210
Basketball — Hardest				
3	E+C+S	court region, player name, play type	0.0	520
3	E+C+S	court region, player name, shot type	0.0	1232
3	E+D+C+S	action type, def. pressure, court region, shooting hand, player name, play type	0.0	671
3	E+D+C+S	action type, court region, shooting hand, player name, play type	0.0	972
3	E+D+C+S	action type, court region, player name, play type	0.0	1165
Soccer — Easiest				
3	E+C	player name, playing position, possession, secondary action	75.0	36
3	D+C	pass target, possession, primary action	71.4	41
3	C	pass target, playing position, possession, secondary action	70.0	33
2	E+C	pass target, team name	64.7	38
3	E+C	playing position, secondary action, team name	64.3	38
Soccer — Hardest				
2	E+C	player name, shot body part	0.0	314
3	C	field flank, playing position, possession	10.0	752
3	C	pass target, playing position, secondary action	12.5	2242
2	C	duel target, playing position	14.3	306
2	D+C	playing position, primary action	14.3	963
Hockey — Easiest				
2	E	player name, team name	89.3	14
3	E+S	player name, team name, ice zone	85.7	10
3	E+D+S	action type, jersey number, player name, team name, ice zone	77.8	19
3	E+D+C+S	action type, goalie view, jersey number, ice zone	71.4	21
3	E+C+S	goalie stance, jersey number, player name, ice zone	70.0	30
Hockey — Hardest				
3	E+D+C	action type, goalie stance, goalie view, jersey number	0.0	987
3	E+D+C	action type, goalie stance, team name	0.0	530
3	E+D+S	action type, jersey number, ice zone	7.1	636
3	E+C+S	goalie stance, goalie view, shot type, team name, ice zone	12.5	902
3	E+C+S	jersey number, shot type, ice zone	14.3	1118

Table 22. T3 five easiest and hardest attribute compositions per sport (min. 5 samples), sorted by R@100 (ties broken by median rank). E = Entity, D = Dynamics, C = Context, S = Spatial. Entity-dominated compositions consistently rank easiest (e.g., player name + team name: 89.3% in hockey), while context-only or context-heavy compositions are hardest, often achieving 0%.

Figure 28. T3 Compositional Video Retrieval: given a natural-language query specifying multiple attributes, the model must retrieve the matching clip from 5,001 candidates (1 positive and 5000 negatives). Attributes in the query are color-coded by category: entity (player name, jersey number, teams playing), dynamics (action type), context (shooting hand, contest status, playing position), and spatial (field/court location). We show the retrieved videos for two queries. **Basketball (top):** the correct video is retrieved at the rank-29, not appearing in the top 3. The top-2 retrieved clips match the action but show the wrong shooting hand and contest status (left hand instead of right hand, and uncontested instead of contested). **Soccer (bottom):** the correct video is retrieved at rank 1. The rank-2 result comes from the same game but shows different playing position (midfielder) and action (foul instead of duel); the rank-3 result matches all attributes except one action detail (aerial instead of ground-level contact). Together, these examples illustrate that the model captures game-level context and scene layout but struggles with fine-grained distinctions.

C Pillar 2: Causal Reasoning (T4–T6)

C.1 T4: Strategic Reasoning QA

T4: Data Construction and Statistics

Given a full-game video, the model must answer an open-ended strategic reasoning question with a free-text response. Questions require understanding complex interactions across extended temporal spans, including tactical decisions, causal chains, and counterfactual reasoning. We define six question types:

- **Tactical & strategic analysis**: reasoning about tactics and strategies employed by teams and players, how these decisions interact, and how they affect the game state.
- **Player role & skill assessment**: identifying player archetypes and evaluating the execution quality of specific roles and associated skills.
- **Causal & counterfactual reasoning**: understanding causal links between events and outcomes, and reasoning how alternative scenarios could change the game state.
- **Anomaly & novelty detection**: identifying what makes certain events, tactics, or outcomes unusual and explaining those novelties.
- **Spatiotemporal & relational reasoning**: reasoning about spatial structures, player positions, and how these relate to specific outcomes or strategies.
- **General**: questions that require a combination of the above capabilities.

T4 requires rigorous quality control. Strategic reasoning questions are uniquely susceptible to language-prior shortcuts: a question that *sounds* like it requires game understanding may in fact be answerable from general sports knowledge alone. We therefore construct the dataset through a multi-stage pipeline with aggressive filtering at each stage, prioritizing question quality over quantity.

- **Initial construction** (Prompt 29): GPT-5.2 generates candidate QA pairs, along with supporting evidence, from professional commentary and game reports. To ground the model against transcription errors, we also provide team rosters.
- **Revision** (Prompt **??**): to mitigate the effect of language priors embedded in question formulation, we revise each candidate using GPT-5.2, instructing it to rephrase the question so that it does not reveal its answer. This preserves the question's intent while removing superficial cues that could allow a model to guess correctly without actually watching the game.
- **Automated quality check** (Prompt 31): GPT-5-mini removes candidates not supported by the evidence from Stage 1. This retains 69.2% of the initial set.
- **Language-bias filtering**: to verify that the remaining questions genuinely require video understanding, we present each candidate to GPT-5.2 and Gemini-3-Flash in a *blind* setting—the models receive only the question, without any video or game con-

QA Generation Prompt

```
You will act as a teacher in a class called 'Sports Video Understanding.' Given a
Question Category, team rosters, subtitles with timestamps, and game reports, your task
is to generate difficult and diverse questions and corresponding answers for
your students about the video, to later be used in a short answer setting.

The provided Question Category is not a strict format, and should instead be used
as inspiration to generate questions of similar or more quality and depth.

**Guidelines and Restrictions:**
- Ensure each question does NOT give away its answer.
- Ensure each question and answer DO NOT contain any subtitles, timestamps, or quotes.
- Do NOT reference information that cannot be directly observed from watching the games
themselves.
- Remember that students will only be provided with the game video without audio to
answer the questions.
- The question and answers should NOT include any assumed details.
- The question should require more than simple action recognition or stats to answer.
- The question and answers MUST be in plain text, no formatting.
- The answers MUST be at most 50 words long.

**Output Requirements:**
Generate the five highest quality questions and corresponding answers based on the
provided data. For each answer, provide evidence: timestamps for relevant subtitles
and relevant quotes from the game report.

**Question Category:**
<question category and examples>

**Rosters:**
<team rosters>

**Subtitles with timestamps:**
<professional commentary>

**Game reports:**
<game report>
```

Figure 29. T4 question-answer generation prompt. For a given game, the model is given professional commentary, game reports, and question type specific examples to produce diverse and difficult questions relating to the specified question type.

text. If either model scores 3/5 or higher, the candidate is removed. Only 5.2% of the initial set survives this stage, demonstrating that the vast majority of initially plausible strategic reasoning questions can in fact be answered without watching the game.

- **Human review**: expert annotators verify that each surviving candidate is (1) fully supported by the evidence and (2) not answerable using general sports knowledge alone. Borderline questions receive manual refinements. The final validated set consists of 398 basketball, 333 soccer, and 269 hockey samples for a total of 1000 samples (2.4% of the initial candidate pool).

Quality Check Prompt

```
You are a sports video understanding expert. Given a question, answer,
and evidence determine if all of the following criteria are met.

**Criteria**
- The question and answer only involve a single game, not multiple.
- The question and answer only reference in game information, not interviews, game
reports, etc.
- The question and answer do not reference any commentary of the game.
- The answer is COMPLETELY supported by the evidence.

You must respond in this form:

Decision: [Yes/No]
Explanation: [Reasoning]

Question: <question>

Answer: <answer>

Subtitle evidence:
<commentary evidence>

Game report evidence:
<game report evidence>
```

Figure 31. T4 automated quality check prompt. GPT-5-mini is given the question, answer, and evidence to determine whether the question is valid and if the answer is supported.

T4: Evaluation Protocol

Strategic reasoning questions may admit more than one valid answer grounded in different aspects of the game. To account for this, we allow each model to produce up to k candidate answers per question and report the *top-k score*: each candidate is independently evaluated against our ground-truth answer by an LLM judge (DeepSeek-V3, selected for reproducibility as an open-source model), and the highest-scoring candidate is used. The judge (Prompt 32) prioritizes alignment on key ideas and reasoning traces over minor factual details. We use $k{=}5$ by default.

T4: Analysis

Effect of response count k. As discussed in the main paper, we allow each model up to k candidate answers per question and report the best-scoring response. We evaluate Qwen3-VL-32B (768 frames) at $k \in \{1, 3, 5, 7, 10\}$. Performance at $k{=}1$ (1.48) is substantially worse than at $k{=}3$ (1.99), indicating that models' top-ranked predictions often do not align with the ground-truth reasoning. A secondary jump occurs between $k{=}5$ (1.99) and $k{=}7$ (2.29). However, even at $k{=}10$, the performance only reaches 2.43, a modest gain of +0.44 over $k{=}5$. This confirms that the questions cannot be solved by exhaustive guessing: the overall score remains well below 3/5 regardless of how many attempts the model is given.

Cross-judge agreement. To assess whether the T4 LLM-as-a-judge evaluation is sensitive to the choice of LLM evaluator, we compare scores from three separate LLM judges:

Judge Pair	mean A	mean B	bias	MAE	Spearman ρ
DeepSeek-V3.2 vs Gemini-3-Flash	2.518	1.798	$+0.720$	0.720	0.829
DeepSeek-V3.2 vs GPT-5.2	2.518	1.858	$+0.660$	0.660	0.822
Gemini-3-Flash vs GPT-5.2	1.798	1.858	-0.060	0.128	0.883

Table 23. Cross-judge consistency for T4 evaluation. We compare the scores assigned by DeepSeek-V3.2, Gemini-3-Flash, and GPT-5.2 on the same evaluation subset. For each judge pair, we report the mean score from each judge, signed bias, mean absolute error (MAE), and Spearman rank correlation. The results show that judges differ in calibration in assigning absolute scores, but preserve relative rankings across judges.

DeepSeek-V3.2, Gemini-3-Flash, and GPT-5.2. Table 23 presents the pairwise agreement on the same evaluation subset from T4. While DeepSeek-V3.2 assigns higher scores than the other two judges, the ordering remains consistent across evaluators as indicated by the pairwise Spearman correlations – all are above 0.82. This indicates that despite calibration differences in absolute scoring among judges, the T4 evaluation is robust in terms of model ranking of answers.

Human study. To verify that T4 is solvable by humans, we conduct an expert study. We sample five T4 questions per sport and ask participants with at least 10 years of experience watching or playing the corresponding sport to answer them using the same free-form format as model responses given the full-game video. Two basketball experts complete the basketball subset and achieve an average score of 4.3/5, while one soccer expert and one hockey expert achieve 4.3/5 and 4.0/5. These results indicate that expert humans can identify and reason about the relevant strategic explanations with high accuracy, going beyond surface-level perception to reason about what happened, why it happened, and how specific tactical decisions affected the outcome.

Additional qualitative examples. Figures 33 and 34 provide additional T4 qualitative examples illustrating two common failure modes. Figure 33 tests tactical causal reasoning that challenges the model to identify not only the defensive strategy, but also the change of tactics relative to earlier possessions. The answers produced by models are plausible in generic basketball language, but reverse the actual causal mechanism in the game. Figure 34 tests performance assessment that challenges the model to grade a player's execution and justify the grade with specific evidence. The models capture the salient late-game three-pointer but miss the earlier passed-up shots that explain the low grade, producing an evaluation that is correct on the visible highlight but incomplete on the underlying reasoning.

LLM-as-a-judge Prompt

```
You are an expert sports analyst and a meticulous free-response evaluation assistant. Your task is to
compare a **Generated Answer (Pred)** against a **Ground Truth Answer (GT)** and assign a score from 0 to
5 based strictly on content accuracy and coverage of the explicitly requested information.

## Evaluation Process

1. **Isolate the question's explicit request**: Identify precisely what the question asks. Discard any
information in the GT that, while true, does not directly answer the specific question.
2. **Extract the core answer from GT**: Determine the minimal set of facts that directly and completely
answers the question.
3. **Compare Pred against this core answer**: Check if the Pred provides the same essential outcome,
event, or causal explanation.
4. **Crucial Rule on Conciseness**: A response that is concise but factually correct and answers the core
question fully must be scored highly. Do not penalize for missing descriptive context or narrative flair
that the question did not explicitly request.
5. Penalize only factual contradictions, inaccurate statements, or omissions that leave the specific
question unanswered.

## Scoring Rubric

### 5 (Perfect)
- Directly and accurately answers the question with the exact core outcome/event.
- No factual errors or contradictions.
- May be concise; verbosity is not required.

### 4 (Good)
- Answers the question correctly in substance.
- May contain minor, non-essential factual errors that do not alter the core answer.
- No major contradictions.

### 3 (Fair)
- Partially answers the question or provides a vague but broadly correct direction.
- May include inaccuracies or omit a critical component of what the question asked for.
- May include limited contradictions.

### 2 (Poor)
- Touches on the topic but misses the main point of the question.
- Contains multiple factual errors or major omissions regarding the question's target.
- Shows partial but shallow understanding.

### 1 (Very Poor)
- Minimal overlap with the core answer.
- Largely incorrect or irrelevant to the question asked.
- May contain significant contradictions.

### 0 (Completely Wrong)
- No meaningful overlap with the core answer.
- Entirely incorrect or unrelated.

### Output Format (Strict JSON Structure)
The JSON must follow this structure, including the analysis steps (gt_analysis, pred_analysis,
justification_cot) within the JSON object.
{{
    "gt_analysis": "...",
    "pred_analysis": "...",
    "justification_cot": "...",
    "score": X
}}
"""

Question: <question>
GT: <ground truth answer>
Pred: <predicted response>
```

Figure 32. T4 LLM-as-a-judge evaluation prompt. The judge scores each predicted response against the ground-truth answer on a 0–5 scale, prioritizing alignment on key ideas and reasoning traces over peripheral details.

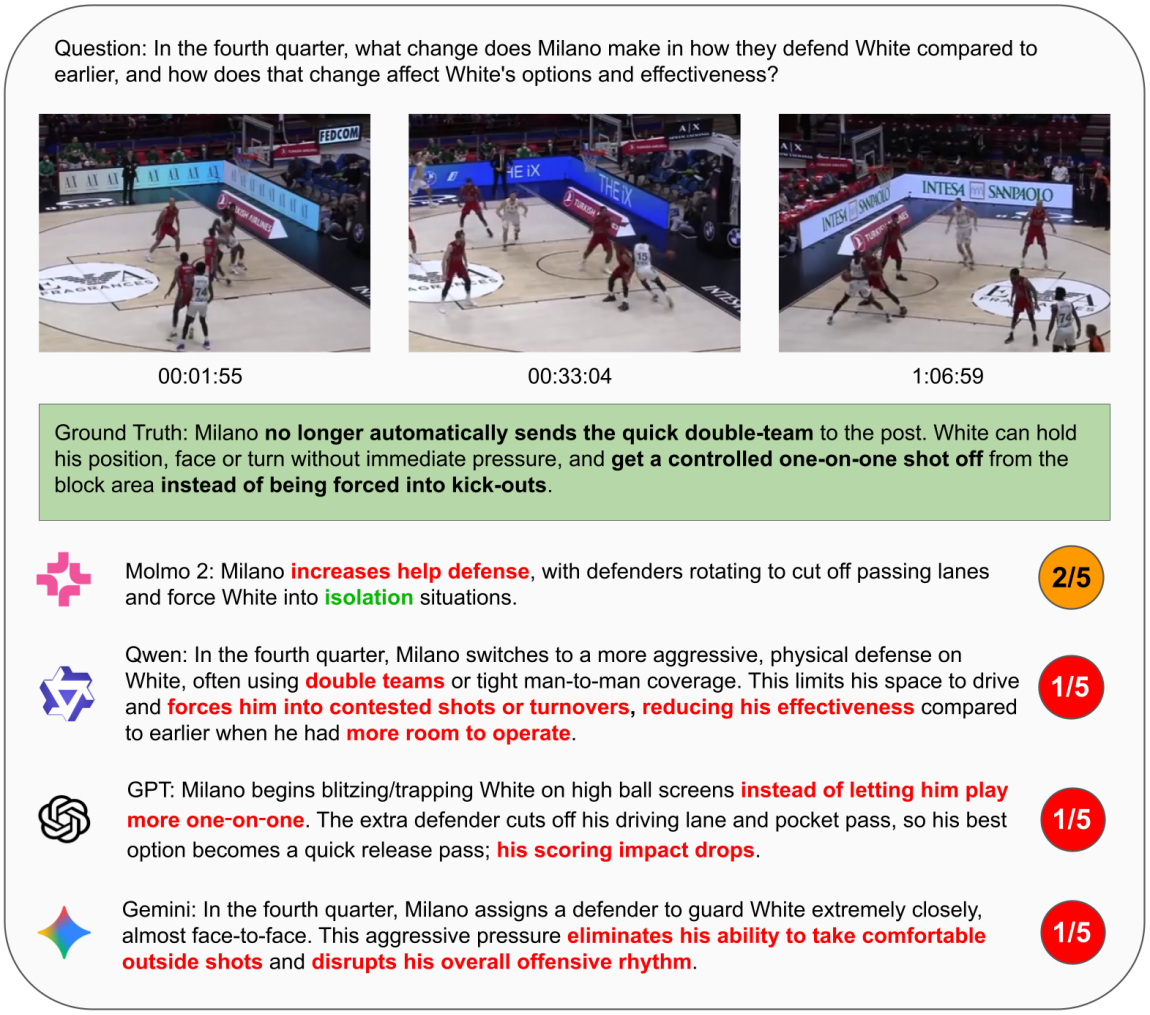
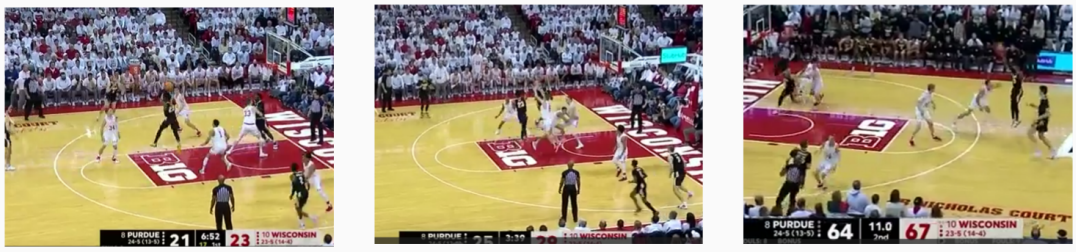


Figure 33. T4 Strategic Reasoning QA (Example 1 — tactical analysis): given a full-game video (∼100 minutes), the model must answer an open-ended strategic question. We show each model's best response (out of k=5 responses) with its LLM judge score (0–5). Here, the model must identify how Milano changed its defensive strategy and what effect this had on a specific player. The ground truth identifies that Milano stopped double-teaming, giving the player controlled one-on-one opportunities. Several models claim the opposite—increased help defense—arriving at a contradictory conclusion about both the tactical change and its downstream effect.

Question: How would you grade Jaden Ivey's shot selection and decisiveness (good, fair, or poor) earlier in the game versus in the clutch and why?



00:32:17 00:40:31 1:51:12

Ground Truth: Early: **poor decisiveness**, shown by **passing up shots and taking strange attempts**. Late: **good decisiveness**, shown by confidently **pulling up to hit the game-tying three**.

Molmo 2: Ivey's early shot selection was **fair**, with a **mix of mid-range and three-point attempts**. His decisiveness was inconsistent, sometimes taking rushed shots. In the final minute, **he was decisive, hitting a crucial three-pointer to tie the game**, demonstrating his composure in critical moments. **3/5**

Qwen: **Good in the clutch**. Ivey **made a crucial three-pointer** with under a minute left, demonstrating improved shot selection and composure under pressure. **3/5**

GPT: Earlier: fair—got downhill at times, but **too many perimeter looks and not enough rim pressure** (graphic shows Purdue 3PT 0–2 when he's highlighted early). In the clutch: poor—**final possession ends in a low-percentage, off-balance long three** as Wisconsin wins 70–67. **1/5**

Gemini: I grade Ivey's shot selection and decisiveness as poor **throughout**. Earlier, he **forced contested shots**, like a tough step-back three at 12:44. In the clutch, he inexplicably drove for a missed layup when down three (58:12) and **hesitated on a final desperation three** (59:45). **2/5**

Figure 34. T4 Strategic Reasoning QA (Example 2 — performance assessment): same setup as Figure 33. The model must evaluate a player's execution quality with concrete justification. Several models correctly identify that Ivey made a crucial three-pointer late in the game, warranting a high execution grade. However, the same models fail to recognize why his earlier performance deserved a low grade—namely, repeatedly passing up open shots. Across both examples, a recurring pattern emerges: models produce plausible-sounding analysis that captures surface-level events but misses the deeper tactical or evaluative reasoning required for a high-quality answer.

Sport	# Train	# Test	# Hours	Avg. Length
Basketball	43,466	7,000	5,530	431.8 s
Hockey	43,585	7,123	6,804	497.7 s
Soccer	10,328	2,593	945	525.9 s

Table 24. T5 dataset statistics across three sports.

C.2 T5: Outcome Forecasting

T5: Data Construction and Statistics

T5 contains 114K forecasting video–question pairs spanning 13,277 hours of video (Table 24), with average observation-clip lengths of 432–526 seconds across sports.

Data construction. We construct the forecasting benchmark from full-game videos with game-aligned play-by-play logs. We define 15 question types across three sports (Table 25), organized into three forecasting categories. For *performance forecasting*, we sample (i) a target entity (player or team), (ii) a target statistic (e.g., points, rebounds), and (iii) an observation window in the video. The ground-truth label is computed by aggregating the target statistic in the play-by-play log over the period *after* the observation window ends, up to a specified future horizon (e.g., end of game). For *game state evolution*, we focus on outcomes such as the final score and ensure that observation windows end at least five minutes before the game concludes, when the outcome remains uncertain. For *strategic intention*, we follow a similar procedure: given a target entity and observation window, we use the play-by-play log to identify the dominant strategy (e.g., most-used play type, preferred attacking flank) adopted by the target in the period beyond the observation window.

- **Performance forecasting**: predicting future player- or team-level statistics (e.g., how many points a player will score from the current moment to the end of the game, or which team will first reach a statistical milestone).
- **Game state evolution**: anticipating how the game state will change, including final scores, possession shifts, and overall outcomes.
- **Strategic intention**: inferring the most likely tactics from observed play patterns (e.g., which play type a team will attempt most frequently, or from which flank a team will predominantly attack).

Preventing shortcut solutions. We take two measures to ensure the task requires genuine forecasting from visual evidence rather than exploiting statistical priors. First, all questions are explicitly formulated about *future* game progression relative to the observation window, not cumulative totals. For example, rather than asking for a player's total game points (which could be looked up or estimated from seasonal averages), we ask: *"Focus*

	Prediction Type	Category
Shared across all sports (6 types)		
1	Player Statistics Prediction	Performance
2	Team Statistics Prediction	Performance
3	Player Statistics Milestone	Performance
4	Team Statistics Milestone	Performance
5	Game Outcome Prediction	Game state
6	Next Player Action Prediction	Game state
Basketball-specific (4 types)		
7	Player Most Attempted Play Type	Strategic intention
8	Player Most Attempted Shot Type	Strategic intention
9	Team Most Attempted Play Type	Strategic intention
10	Team Most Attempted Shot Type	Strategic intention
Hockey-specific (2 types)		
11	Player Most Attempted Shot Type	Strategic intention
12	Team Most Attempted Shot Type	Strategic intention
Soccer-specific (3 types)		
13	Team Most Attempted Attack Flank	Strategic intention
14	Next Team to Score	Game state
15	Team Most Possession	Strategic intention

Table 25. T5 forecasting question type taxonomy (15 types total). Each type maps to one of the three forecasting categories defined in §C.2: performance forecasting, game state evolution, or strategic intention. Six types are shared; the remainder are sport-specific.

on the player who makes a 3-point shot during 4:13–4:23. How many points will this player score from the end of this segment until the conclusion of the game?" While a player's historical averages may provide a coarse prior, the precise answer depends on game-specific factors observable only in the video—matchup dynamics, foul trouble, momentum shifts, rotation patterns, and score differential—making statistical priors alone insufficient for reliable prediction.

Second, we avoid revealing target player names. Instead, we use indirect references grounded in the observation clip, such as *"the player who makes a 3-point shot during 4:13–4:23."* This forces the model to visually identify the relevant player from the described

Model	Sampling	FT	Bball.	Soccer	Hockey	Overall / CE
Qwen3-VL 8B	1 FPS	–	35.8	39.2	34.9	36.0 / 0.22
Qwen3-VL 8B	0.2 FPS	–	37.4	39.6	35.3	36.9 / 0.23
Qwen3-VL 8B	0.2 FPS	✓	**46.1**	**45.1**	**43.5**	**44.8 / 0.01**

Table 26. T5 temporal sampling ablation. Accuracy (%) and calibration error for Qwen3-VL 8B with different frame sampling rates. Increasing frame density from 0.2 FPS to 1 FPS does not improve forecasting accuracy. Fine-tuning substantially improves both accuracy and calibration.

action and time window before any prior knowledge can even be applied.

T5: Evaluation Protocol

We report **accuracy** and **calibration error (CE)** following [42]. Predictions are grouped into $B{=}5$ equally spaced confidence bins, and CE is computed as:

$$\mathrm{CE} = \frac{1}{B}\sum_{i=1}^{B} |\mathrm{acc}(i) - \mathrm{conf}(i)|$$

This metric penalizes overconfidence and underconfidence. $\mathrm{CE} = 0$ indicates perfect calibration.

T5: Baselines and Training Details

We evaluate five models in a zero-shot setting: GPT-5.2, Gemini 3.0 Pro, Qwen3-VL 8B, BIMBA, and Molmo 2 8B. For GPT-5.2, we sample frames at 0.5 FPS (up to the 500-image API limit) and disable reasoning mode to enable log-probability extraction for CE computation. For Gemini 3.0 Pro, we provide the full video and report accuracy only (the API does not expose log probabilities). For Qwen3-VL 8B, we evaluate at two temporal sampling rates (1 FPS and 0.2 FPS). Molmo 2 8B is evaluated at 0.2 FPS, and BIMBA uses 32 uniformly sampled frames. We additionally fine-tune Qwen3-VL 8B on the T5 training set using LoRA fine-tuning on the combined three-sport training data at 0.2 FPS.

Temporal sampling ablation. We ablate the effect of temporal sampling using Qwen3-VL 8B at 1 FPS and 0.2 FPS. Increasing the sampling rate does not improve forecasting performance: overall accuracy is 36.0% at 1 FPS and 36.9% at 0.2 FPS, with similar minor differences across basketball, soccer, and hockey. This suggests that T5 is not primarily limited by the number of sampled frames. In other words, observing more frames does not lead to better reasoning about future outcomes.

Additional qualitative examples.

Figure 35 provides a qualitative example of the prediction-horizon effect on T5. We ask the same game-outcome question—which team will win—at four observation times

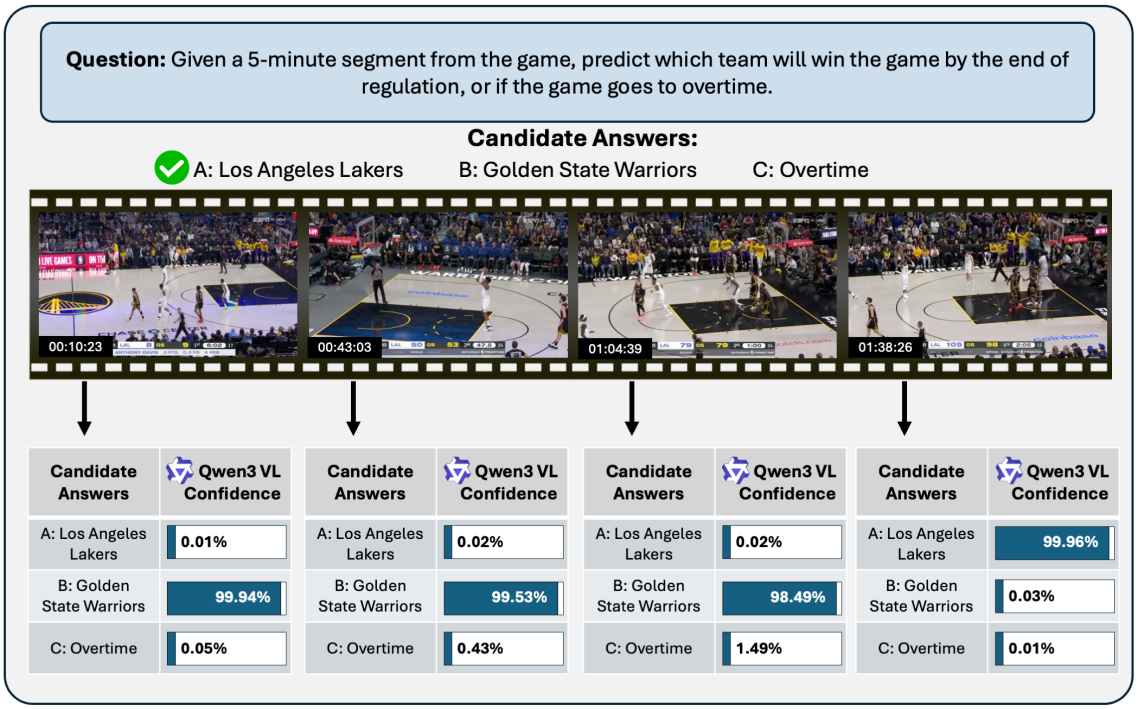


Figure 35. T5 Outcome Forecasting — prediction horizon analysis: given an observation clip and a multiple-choice question, the model must predict a future game outcome with calibrated confidence. Here, the same game-outcome question is asked at four different time points in the same game, with the observation window shifting from early (00:10:23) to late (01:38:26). The correct answer is Los Angeles Lakers. Qwen3-VL predicts Golden State Warriors with >98% confidence at the first three time points, then abruptly flips to Lakers with 99.96% confidence near the end — remaining near-certain throughout regardless of whether its prediction is correct. This reveals a fundamental calibration failure: the model does not adjust its confidence to reflect the genuine uncertainty of long-horizon predictions.

within the same game, with the window shifting from early (00:10:23) to late (01:38:26). The correct answer is the Los Angeles Lakers. Qwen3-VL predicts Golden State with over 98% confidence at the first three observation windows, then abruptly flips to the Lakers with 99.96% confidence at the final window. The model is near-certain at every horizon, including the early windows where the outcome is genuinely undecided and its prediction is wrong. This illustrates that the model is poorly calibrated to the forecasting task. It does not lower its confidence when the question is far from decisive, and the late-window correction reflects the game state becoming obvious rather than improved forecasting.

	Template	Scope
Shared across all sports (8 templates)		
1	Game Summarization	Single-game
2	Player Performance Summarization	Single-game
3	Team Strategy Analysis	Single-game
4	Player Performance Period Comparison	Single-game
5	Team Multi-Game Summarization	Multi-game
6	Player Multi-Game Performance Summarization	Multi-game
7	Two-Team Matchup Comparison	Multi-game
8	Player Performance Comparison	Multi-game
Basketball-specific (templates 9–10)		
	Game Final Moments	Single-game
	Player Face-to-Face Matchup Comparison	Multi-game
Hockey & soccer-specific (templates 9–10)		
	Goal Development Analysis	Single-game
	Team Home–Away Performance Analysis	Multi-game

Table 27. T6 report template taxonomy. Each sport uses the 8 shared templates plus 2 sport-specific templates, yielding 10 templates per sport (5 single-game, 5 multi-game). Hockey and soccer share the same sport-specific templates.

C.3 T6: Long-Form Narrative Synthesis

T6: Data Construction and Statistics

Given a full-game video (or multiple game videos) and a topic-specific writing prompt, the model must generate a detailed game report that accurately captures what happened, highlights the most important developments, and adheres to the specified writing constraints. We organize report prompts into two settings and five topic categories:

- **Single-game reports** require the model to watch one full game and produce a report on a specified topic. Topics include an overall game summary, a targeted player performance analysis, a team strategy breakdown, a comparison of a player's performance across periods, or a narrative focused on the game's decisive moments.
- **Multi-game reports** require synthesizing 2–10 games linked by a coherent theme—such as a recurring team, a head-to-head matchup, or a player's trajectory across consecutive games—into a single analytical piece. Topics include multi-game team sum-

Sport	#Train	#Test	Videos	Hours	Avg. Len.	Avg. Words
Basketball	8,375	350	5,224	7,792 h	88 min	471
Hockey	7,780	350	1,835	3,324 h	108 min	467
Soccer	1,560	300	249	414 h	100 min	495

Table 28. T6 long-form narrative synthesis dataset statistics. Videos refers to unique full-game videos. Avg. Len. is the mean video duration. Avg. Words is the mean ground-truth report length. Because multiple report samples can target the same game and multi-game samples reuse component videos, the number of unique videos is smaller than the number of samples.

maries, player performance trends, and matchup comparisons.

Each prompt additionally specifies a target word count, a narrative perspective (e.g., analyst, beat reporter), and formatting constraints. This task is uniquely demanding because it requires the model to integrate perception over hours of video, reason about which events are most consequential, and distill this understanding into a structured, coherent narrative under space constraints—capabilities that go well beyond those tested by short-form QA or captioning.

For each sport, we design ten report-writing templates—five for single-game settings and five for multi-game settings (Table 27). Templates range from overall game recaps to player performance comparisons, with multi-game templates spanning 2–10 games linked by a coherent theme (e.g., a recurring team, matchup, or target player).

To construct ground-truth reports, we leverage the aligned multimodal resources for each game: full-game video, play-by-play logs, structured metadata, and journalist-written game reports. However, journalist reports frequently incorporate information not directly observable from the game video—such as pre- or post-game interviews, seasonal context, and discussion of upcoming matchups. To ensure that ground-truth reports match the intended topic and are grounded only in game-observable evidence, we provide GPT-5 with the aligned multimodal resources and the corresponding writing template, instructing it to synthesize a report that follows the template while remaining faithful to the underlying game content. Dataset statistics are summarized in Table 28.

T6: Evaluation Protocol

We evaluate generated reports along three dimensions using an LLM-as-a-judge framework with Qwen3-235B-A22B Thinking as the judge. We select this open-source model to ensure reproducible and transparent evaluation, unlike proprietary APIs that may change unexpectedly.

- **Factual accuracy** (Prompt 36): following FActScore [53], we decompose each report into atomic verifiable facts and verify each against ground-truth resources (game re-

port, structured metadata, play-by-play logs). The score is the proportion of *Supported* facts among all verifiable facts (*Supported* + *Contradicted*).

- **Saliency** (Prompt 37): we decompose the ground-truth report into atomic statements and measure what proportion is covered by the generated report. This captures whether the model identifies and prioritizes the most important game developments.
- **Writing quality** (Prompt 38): rubric-based 1–5 scoring for stylistic adherence, narrative coherence, and overall quality.

T6: Baselines and Training Details

We evaluate the following models and methods:

- **Qwen3-VL 8B**: The frames are sampled at 1 FPS and then provided to the model together with the writing instruction for end-to-end report generation.
- **LLoVi** [99]: a two-stage pipeline that first uses a LLaVA-Video model (fine-tuned on Pillar 1 perception tasks) to generate dense 10-second captions for each input video, then feeds these captions with the writing instructions to GPT-5 for final report generation. This tests whether strong short-horizon perception can bootstrap long-form narrative synthesis.
- **GPT-5**: 500 uniformly sampled frames (budget split evenly across games for multi-game settings).
- **Gemini 3.1 Pro**: each video is downsampled to fit within the 1-hour API input limit; for multi-game settings, the budget is split evenly across videos which are then concatenated into a single input sequence.
- **Oracle**: to estimate the perception bottleneck, we replace video input with the detailed play-by-play logs (textual event records with timestamps, player actions, and game state) and feed them to GPT-5 together with the writing instructions.

T6: Analysis

Caption-then-aggregate baseline. To test whether strong short-clip perception is sufficient for Pillar 2 reasoning, we feed the descriptions produced by our fine-tuned perception model into a language-model aggregator, following the two-stage structure of LLoVi [99]. We first use the LLaVA-Video model fine-tuned on T1 and T2 to generate structured play descriptions for consecutive 10-second segments of the full-game video, following the T1 captioning format. We then use GPT-5 as an aggregator, conditioning it on the generated captions and the task-specific writing instruction. The pipeline reaches 25.20% factual accuracy and 3.41% saliency. These results indicate that strong short-clip perception does not transfer to Pillar 2 performance. Even with dense perceptual descriptions as input, the aggregator cannot perform the evidence integration and event prior-

Judge Pair	mean A	mean B	bias	MAE	Spearman ρ
Factual accuracy					
DeepSeek vs Gemini	0.454	0.594	−0.140	0.171	0.857
DeepSeek vs GPT-5.2	0.454	0.433	+0.021	0.102	0.886
DeepSeek vs Qwen3	0.454	0.475	−0.021	0.112	0.859
Gemini vs GPT-5.2	0.594	0.433	+0.162	0.181	0.888
Gemini vs Qwen3	0.594	0.475	+0.119	0.167	0.856
GPT-5.2 vs Qwen3	0.433	0.475	−0.042	0.114	0.859
Saliency					
DeepSeek vs Gemini	0.0260	0.0277	−0.0017	0.0021	0.952
DeepSeek vs GPT-5.2	0.0260	0.0257	+0.0002	0.0024	0.950
DeepSeek vs Qwen3	0.0260	0.0273	−0.0013	0.0031	0.951
Gemini vs GPT-5.2	0.0277	0.0257	+0.0019	0.0027	0.962
Gemini vs Qwen3	0.0277	0.0273	+0.0004	0.0032	0.950
GPT-5.2 vs Qwen3	0.0257	0.0273	−0.0015	0.0028	0.943
Writing style					
DeepSeek vs Gemini	4.552	4.324	+0.229	0.343	0.743
DeepSeek vs GPT-5.2	4.552	3.610	+0.943	0.962	0.749
DeepSeek vs Qwen3	4.552	4.162	+0.390	0.505	0.745
Gemini vs GPT-5.2	4.324	3.610	+0.714	0.867	0.731
Gemini vs Qwen3	4.324	4.162	+0.162	0.410	0.749
GPT-5.2 vs Qwen3	3.610	4.162	−0.552	0.933	0.752

Table 29. Pairwise agreement between different LLM judges for T6 evaluation. We report the mean score assigned by each judge in the pair, bias (mean A minus mean B), mean absolute error (MAE), and Spearman rank correlation. All three evaluation dimensions show low MAE and high rank correlation across judges.

itization that T6 requires.

Cross-judge agreement. We evaluate the stability of T6 LLM-as-a-judge evaluation by comparing four independent LLM judges: DeepSeek-V3.2, Gemini-3-Flash, GPT-5.2, and Qwen3-235B-Thinking. Table 29 reports pairwise agreement for the three T6 evaluation dimensions: factual accuracy, saliency, and writing style. All three dimensions show strong agreement across judges, with low pairwise MAE and consistently high Spearman rank correlation, all over 0.73. These results support the reliability of T6 evaluation and are highly stable across judges.

Factual Evaluation Prompt

```
You will perform a two-step task on <Sport> game text. Your job is
(1) to extract atomic, observable in-game facts, and then
(2) to verify each fact against the provided ground-truth resources.

STEP 1 — EXTRACT OBSERVABLE IN-GAME FACTS
In this step, your job is to break down every sentence and extract every stat and every
discrete in-game event described.

The extracted items must strictly reflect information that is visually or audibly
verifiable from the live game broadcast. Do not include any contextual information that
cannot be observed from the game itself, including: interviews; commentary about
season framing; player career context; transactions; historical or team background;
or any narrative not grounded in the on-court action.

Key events include—but are not limited to—scoring plays, lead changes, momentum plays,
clutch shots, defensive events, blocks, steals, or any other on-court actions that would
be observable from watching the game.

Each extracted fact must:
<Atomic Statement Instruction>

Do not keep compound statements. Whenever a sentence contains multiple facts (player +
multiple stats, action + location + score, etc.), split into the smallest independently
verifiable units. Each unit should be testable as Supported or Contradicted.

Example:
<Compound Statement Examples>

If the report contains no observable in-game stats or events, write:
"No extractable game facts."

Your output should list each extracted fact as a separate line, each
containing one complete sentence. For example:
<Atomic Statement Examples>

STEP 2 — VERIFY EXTRACTED FACTS AGAINST GROUND TRUTH
After extracting the facts, compare each one against all provided ground-truth resources:
(1) the official game report, (2) the period-by-period box score statistics,
(3) play-by-play Game Log.

Assign one of three possible classifications:
 Supported — The fact is confirmed or logically entailed by the ground truth.
 Contradicted — The fact conflicts with the ground truth.
 Inconclusive — The ground truth does not provide enough information to confirm or
 contradict the fact.

Instructions:
<Additional Instructions>

<INPUT>

Your output format must be EXACTLY as follows (do not deviate):
[Fact text] — Supported
[Fact text] — Contradicted
[Fact text] — Inconclusive
...

IMPORTANT: You MUST end your response with the overall score in this
EXACT format:
Overall Score: X/Y
Where X = number of Supported facts and Y = total number of Supported facts + Contradicted
facts. Do not include Inconclusive facts in the score.
```

Figure 36. T6 factual accuracy evaluation prompt. The judge decomposes the generated report into atomic facts and labels each as Supported, Contradicted, or Inconclusive against the ground-truth resources. The score is $\frac{\text{Supported}}{\text{Supported}+\text{Contradicted}}$; Inconclusive facts are excluded.

Coverage Evaluation Prompt

```
You will evaluate the factual coverage of a GENERATED REPORT against a fixed list of
GROUND-TRUTH ATOMIC FACTS.

For each ground-truth fact, determine whether the generated report contains the
same meaning.

## DEFINITIONS

Covered: The generated report explicitly states the fact OR clearly implies it with
equivalent meaning, without altering any key details.

Not Covered: The fact is missing, too vague to confirm, contradicted, or altered in
any key way.

## EQUIVALENCE RULES (STRICT)

A ground-truth fact is Covered ONLY if ALL of the following conditions are met:

1. Entities: Correct player/team names are used (no substitution with wrong person/team).
2. Directionality/Polarity: The meaning is preserved—scored vs missed, won vs lost,
   led vs trailed. No flipping.
3. Numbers: If the fact includes any number (<Sport Specific Stats>),  the generated
   report  must match the number EXACTLY or express the  same value in words.
   Different numbers = Not Covered.
4. Timing/Half: If the fact specifies a half, extra time, stoppage time, time remaining,
   or sequence, the generated report must match that timing. (If the ground-truth
   fact has no timing constraint, timing does not need to match.)
5. Specificity: Generic statements do NOT cover specific facts.
   <Sport Specific Example>

## YOUR TASK

1. Read the GROUND-TRUTH FACT LIST below.
2. Read the GENERATED REPORT below.
3. For EACH fact in the ground-truth list, output whether it is "Covered" or
   "Not Covered".
4. Compute the coverage score: Covered facts / Total facts.

## OUTPUT FORMAT (STRICT - NO EXPLANATIONS)

List each fact with its classification on a separate line:
[Fact text] — Covered
[Fact text] — Not Covered
...

IMPORTANT: You MUST end your response with the overall score in this EXACT format:
Overall Score: X/Y

Where X = number of Covered facts and Y = total number of facts.

---

GROUND-TRUTH FACT LIST:
{fact_list}

GENERATED REPORT:
{generated_report}
```

Figure 37. T6 saliency evaluation prompt. The judge receives a set of atomic statements extracted from the ground-truth report and determines which are covered by the generated report. The saliency score is the proportion of ground-truth statements covered.

Writing Style Evaluation Prompt

```
You are an expert sports analyst and a meticulous long-form report evaluation assistant.
Your task is to evaluate how well a Report satisfies the REPORT_INSTRUCTION
specifically in terms of writing style, formatting, persona adherence, narrative coherence,
and writing flow.

Crucial Constraints:
- You DO NOT generate your own report. You only analyze and score the Pred report.
- You MUST follow a strict Chain-of-Thought (CoT) procedure for every evaluation category below:
  • task_input_analysis: Identify style, formatting, persona, tone, and perspective required
    by the REPORT_INSTRUCTION.
  • report_analysis: Analyze the Pred report's tone, structure, flow, formatting, and stylistic elements.
  • justification_cot: Explicitly compare the Pred report against the stylistic/formatting
    instructions. Provide step-by-step reasoning.
  • score: Assign a score from 1 to 5 based on the rubric.

EVALUATION CATEGORIES AND PROCEDURES

## STYLISTIC & PERSONA ADHERENCE

FOCUS:
- Does the report successfully adopt the requested Writing Style (e.g., Analytical, Journalistic,
  Dramatic)?
- Does it successfully adopt the requested Perspective/Audience (e.g., Scout, Fan, Neutral Analyst)?
- Does it adhere to formatting constraints (word count, paragraph form, bullets, etc.)?

PROCEDURE:
- task_input_analysis: Identify required style, persona, formatting.
- report_analysis: Analyze tone, vocabulary, narrative approach, formatting.
- justification_cot: Evaluate consistency with required style/persona; check formatting
  constraint adherence.

SCORING GUIDE:
<Scoring Rubric>

## NARRATIVE COHERENCE & QUALITY

FOCUS:
- Coherence: Is the report logically structured?
- Flow: Do ideas transition smoothly?
- Clarity & Fluency: Is writing professional, readable, and non-repetitive?
- Synthesis: Does the report integrate ideas into a cohesive narrative?

PROCEDURE:
- report_analysis: Evaluate structure, transitions, clarity, grammar, repetition, and synthesis.
- justification_cot: Provide detailed reasoning on coherence, flow, and clarity.

SCORING GUIDE:
<Scoring Rubric>

## FINAL HOLISTIC SCORE

The final_overall score MUST reflect overall stylistic and narrative quality.

ERROR SEVERITY:
- Minor Error: Small style/flow/formatting issues.
- Major Error: Broken flow, inconsistent persona, major formatting failures.
- Critical Error: Completely incoherent or disregards all stylistic and formatting rules.

SCORING GUIDE:
<Scoring Rubric>

<Output Format Instruction>

Do NOT include any extra keys.
Do NOT output explanations outside the JSON object.
Ensure the JSON is valid.

REPORT TO ANALYZE:
{report}

REPORT_INSTRUCTION:
{instruction}
```

Figure 38. T6 writing-style evaluation prompt. Given a generated report and its writing instruction, the judge scores (i) stylistic/persona adherence and (ii) narrative coherence/quality, then assigns a final holistic score.

D Pillar 3: Strategic Simulation (T7–T8)

D.1 T7: Motion-Conditioned Generation

T7: Data Construction and Statistics

Constructing the dataset involves two stages: extracting player trajectories from broadcast video, and producing player-removed background videos. We describe each below.

Trajectory extraction. Player trajectories are extracted using the MixSort multi-object tracking framework [19], which is specifically designed for tracking players in sports scenes. We use the publicly released model off the shelf without fine-tuning on our data. MixSort detects players in each frame and associates detections across frames to form continuous tracks, assigning each player a unique identity maintained throughout the clip. Tracks with short duration or unstable identities are discarded.

Player inpainting. To isolate motion control from appearance synthesis, we construct a *player-removed background video* by inpainting out all player regions from the original footage using Gen-OmniMatte [36], with player masks derived from the tracking results. The resulting video preserves the court layout, audience, lighting, and camera motion while containing no foreground players in any frame. By decoupling background reconstruction from player synthesis, this setup allows the generation model to focus entirely on trajectory-conditioned rendering rather than reconstructing the entire scene. For players entering the scene after the first frame, we retain a crop from their first visible bounding box as an appearance reference; in all subsequent frames, the entering player is removed through inpainting like all other players.

Filtering and quality control. We apply two stages of filtering to ensure data quality. First, we filter clips based on the number of visible players per frame: basketball clips require at least eight and soccer clips at least ten. This removes close-ups, replays, and other non-gameplay shots where tracking is unreliable, discarding 81.3% of the original basketball clips and 47.4% of the soccer clips. Second, after inpainting, we re-run MixSort on the player-removed videos: since all players should have been removed, any detected bounding box indicates a residual inpainting artifact. Clips containing such detections are filtered out, removing a further 15.2% of inpainted outputs and ensuring that background videos used for generation are clean.

Dataset statistics. Table 30 summarizes the final dataset after all filtering. The dataset contains 123K soccer clips and 166K basketball clips, with an average clip duration of 6 seconds. Soccer clips contain substantially more visible players per clip (20.6 on average) than basketball clips (10.0), reflecting the different spatial scales and broadcast views of the two sports.

Sport	Split	Hours	Clips	Players / Clip
Soccer	Train	172.4	103K	20.7
	Val	16.8	10K	19.9
	Test	16.7	10K	19.8
	Total	205.9	123K	20.6
Basketball	Train	244.5	146K	10.0
	Val	16.7	10K	10.0
	Test	16.8	10K	10.0
	Total	278.0	166K	10.0

Table 30. T7 motion-conditioned generation dataset statistics. Soccer clips contain roughly twice as many visible players as basketball (20.6 vs. 10.0), reflecting different spatial scales and broadcast views, and making soccer the more challenging setting for multi-agent generation.

T7: Evaluation Protocol

We evaluate motion-conditioned generation using two complementary metrics: *Video mIoU*, which measures trajectory fidelity, and *Temporal feature similarity*, which evaluates the visual consistency of generated players over time. For both metrics, the ground-truth reference is derived from the original broadcast video *before* the player-removal inpainting step, with ground-truth trajectories obtained using the same MixSort tracking pipeline employed during dataset construction. In addition to these two automatic metrics, we also conduct a human evaluation to assess perceptual quality and trajectory-following realism.

Video mIoU. To evaluate motion accuracy, we first run the MixSort tracking pipeline on each generated video to obtain predicted player trajectories. We then establish a one-to-one correspondence between predicted and ground-truth players via Hungarian matching of bounding boxes in the first frame (using IoU as the matching criterion). Once this correspondence is established, the identity mapping is kept fixed for the remainder of the video. Video mIoU is computed by averaging the IoU between matched predicted and ground-truth bounding boxes across all frames within each trajectory, and then averaging across all matched trajectories in the evaluation set.

Temporal feature similarity. To measure the visual fidelity of generated players over time, we compute temporal feature similarity between generated and ground-truth player crops. For each matched trajectory, player regions are cropped from both the generated and ground-truth videos using the ground-truth bounding boxes. The cropped regions are resized to a uniform spatial resolution and encoded using the SigLIP v2 vision encoder (ViT-SO400M, 384×384) [76]. Cosine similarity is computed between the L2-normalized feature embeddings at each frame. In cases where the generated video does

not contain a valid player appearance at the corresponding location (e.g., the player is missing or the generated content does not overlap with the ground-truth bounding box), the similarity score for that frame is set to zero. Scores are averaged across all frames within each trajectory and then across all player trajectories in the evaluation set.

Evaluation subset. Due to the substantial computational cost of trajectory-conditioned video generation, we evaluate the two baselines, ATI [78] and MagicMotion [40] (described below), on a subset of 100 videos. Our method is evaluated on both the same 100-video subset (for direct comparison) and on a larger set of 1,000 videos (to assess performance stability at scale).

Evaluation constraint for baselines. Our method and the ground-truth reference include trajectory information for players that enter the scene after the first frame. In contrast, ATI and MagicMotion only receive trajectory conditions for players visible in the first frame, since their models take only the first frame as input and therefore can only control players present in that frame. To ensure a fair comparison, both Video mIoU and temporal feature similarity are computed only for first-frame-visible players when evaluating the baselines.

T7: Baselines and Training Details

Our method. Our model is built on Wan 2.1-Fun-1.3B-Control [75], a controllable video generation model that originally accepts a single reference frame and a control signal. We extend this architecture to additionally condition on a player-removed background video and multi-player bounding-box trajectories, and fine-tune it on our dataset using LoRA with rank 32 and $\alpha = 32$, applied to the attention and feed-forward layers of the DiT backbone (`q, k, v, o, ffn.0, ffn.2`). The resulting model takes three inputs: (1) the first frame of the original video as an appearance reference, (2) a player-removed background video providing scene context, and (3) per-player bounding-box trajectory conditions. To encode motion control, each bounding-box trajectory is rendered into a per-frame RGB control image where each player identity is represented by a distinct color. All player trajectories are composited into a single control image per frame. These color-coded control images are encoded into 16-channel latent representations using the pretrained VAE and injected into the DiT through the model's control branch. The background video is encoded through a separate VAE pathway and combined with the reference frame to form the image-conditioning tensor.

Training is performed at a resolution of 832×480 for 81 frames (≈5.4 s at 15 fps). We train separate models for basketball and soccer using their respective training splits. Optimization uses AdamW with a learning rate of 1×10^{-4} for 3 epochs. Text conditioning uses a fixed sport-specific prompt (*"a realistic basketball game video"* or *"a realistic soccer*

game video"). Each model is trained on 8× NVIDIA A6000 GPUs, taking approximately two weeks per sport.

ATI. We compare against ATI [78], using the Wan2.1-14B variant released by the authors, which supports generation of up to 81 frames. ATI represents motion control through point trajectories. Since our dataset provides per-player bounding-box trajectories, we convert each box trajectory into a point trajectory by taking the center of the bounding box in each frame. We use the official zero-shot image-to-video inference pipeline without additional fine-tuning on our dataset.

MagicMotion. We compare against MagicMotion [40], using the CogVideoX-5B-I2V variant released by the authors. The public code supports only 49-frame generation; we use its bounding-box trajectory control configuration to match our task setup. To generate 81-frame sequences for fair comparison, we adopt a two-stage strategy: we first generate a 49-frame video, then take the generated frame at timestep 32 as a new starting frame and run the model again to produce the remaining frames. The two segments are concatenated to obtain an 81-frame video. We use the official zero-shot inference pipeline without additional fine-tuning. For fair comparison, all methods receive the same initial frame and trajectory conditions.

T7: Analysis

Additional qualitative examples. We provide additional qualitative comparisons between ATI and our method on a representative basketball clip, alongside the ground-truth video in Figure 39, using the same task setup and visual annotations as Figure 16. Compared to soccer, basketball scenes are substantially more challenging due to tighter player spacing, faster interactions, and more frequent occlusions. Our method follows the prescribed trajectories more faithfully overall, whereas ATI often exhibits noticeable trajectory violations, generates unrealistically large players, and introduces background artifacts in later frames. Nevertheless, our method still shows occasional identity inconsistencies, such as jersey color changes visible in the zoomed-in comparisons.

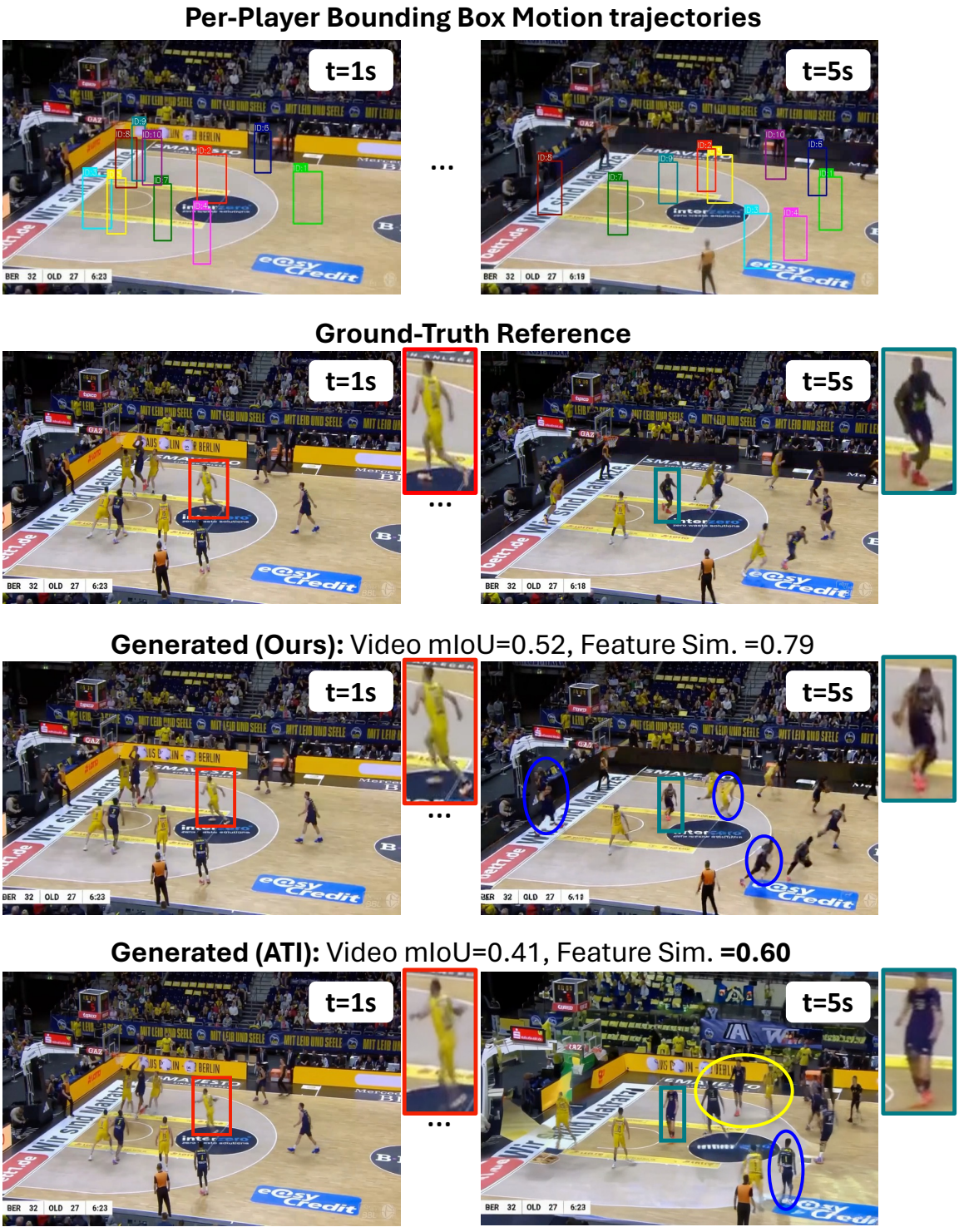


Figure 39. T7 Motion-Conditioned Generation (basketball): same task setup and visual annotations as Figure 16. Basketball scenes are more challenging than soccer due to tighter player spacing and more frequent occlusions. **Yellow** circles highlight trajectory deviations; **blue** circles mark correct motion with incorrect appearance. The zoomed crop compares generated and ground-truth player detail.

	Train	Val	Test	Total
Clips	64,003	5,000	5,000	74,003
1 player / clip	92.5%	95.3%	95.4%	92.9%
2 players / clip	7.3%	4.6%	4.4%	6.9%
3 players / clip	0.2%	0.2%	0.2%	0.2%

Table 31. T8 goal-conditioned action generation dataset overview. Each clip is 81 frames at 15 fps (≈5.4 s) at 832×480 resolution. In the vast majority of clips the generation target is a single player. The remaining 7% specify multiple target players.

	Minimal	Short	Medium	Long
	< 0.05	0.05–0.15	0.15–0.30	≥ 0.30
% of actions	5.3	24.5	38.6	31.6

Table 32. T8 spatial displacement distribution between start and end bounding boxes (in normalized frame coordinates). The mean displacement is 0.247 (median 0.222, std 0.150), and the mean bounding-box area is 0.015 of the frame, confirming that players occupy small regions of the wide broadcast view. Over 70% of actions involve medium or long displacements, indicating that the benchmark primarily targets dynamic player motion.

D.2 T8: Goal-Conditioned Action Generation

T8: Data Construction and Statistics

The goal-conditioned dataset contains 74,003 basketball clips, split into 64,003 training, 5,000 validation, and 5,000 test clips (Table 31). Each clip specifies one or more target players, each with an associated action outcome and spatial constraints (start and end bounding boxes), for 79,448 target players in total. Most clips involve a single target player (92.9%), while 6.9% contain two and 0.2% involve three. All clips are 81 frames at 15 fps (≈5.4 s) at 832×480 resolution.

Instruction specification. Instructions follow a standardized template (*"Simulate Player #XX executing ..."*) and can incorporate several elements of increasing specificity:

- **Target player identity.** Each participating player is identified by jersey number and localized via a bounding box drawn on the first frame. Instructions may involve up to three players: the primary actor, a teammate (e.g., a passer), and a defender.
- **Desired action outcome.** The core objective is an action label specifying the intended play outcome. The taxonomy covers 14 basketball event types spanning scoring (e.g., *2 Pt Made*, *3 Pt Missed*, *Free Throw*), ball movement (*Assist*, *Steal*, *Turnover*, *Rebound*), set plays (*Pick-and-Roll*, *Screen*, *Post*), and *Foul*. Scoring events may additionally specify

Action type	%	Action type	%
Shooting	47.2	*Set plays*	14.5
3 Pt Missed	12.6	Pick-and-Roll	10.5
2 Pt Missed	11.9	Screen	3.5
2 Pt Made	10.3	Post	0.5
3 Pt Made	8.2	*Possession change*	14.4
Free throw	3.6	Rebound	9.8
Other scoring	0.6	Steal	2.8
Playmaking	14.2	Turnover	1.8
Accurate pass	8.0	*Fouls*	7.8
Assist	6.2		

Table 33. T8 action type distribution. Percentages are computed over all player-action annotations in the dataset. Actions are grouped into five categories: *Shooting* (attempts to score), *Playmaking* (passes and assists), *Set plays* (coordinated team actions such as screens), *Possession change* (turnovers, steals, and rebounds), and *Fouls* (rule violations). Free throw combines successful and unsuccessful attempts; other scoring includes compound outcomes such as scoring while simultaneously drawing a foul.

compound outcomes, such as a made shot that also draws a foul.

- **Tactical context.** The instruction may specify the offensive scheme in which the action occurs (e.g., *Pick-and-Roll*, *Isolation*).
- **Shot attributes.** For shooting actions, additional attributes may include the court location (e.g., *Wing*, *Corner*), shot type (e.g., *jump shot*, *layup*), dribble moves (e.g., *step back*, *crossover*), and shooting hand.
- **Defensive pressure.** Instructions may describe the level of defensive contest on the shot (e.g., *contested*, *uncontested*).
- **Player interactions.** Multi-player interactions may be specified explicitly, such as passing relationships (*assisting Player #11*) or defensive matchups (*guarded by Defender #5*).

Instructions range from simple to highly specific. A minimal example:

> *"Simulate Player #23 scoring a three-pointer from the wing."*

A maximally detailed example combining all elements:

> *"Simulate Player #23 scoring a three-pointer in an isolation play from the wing, using a step-back jump shot with the right hand under defensive pressure from Defender #5."*

Spatial constraints. In addition to the textual instruction, each sample provides spatial constraints for the participating players. Each player is specified by a bounding box

Attribute	Coverage	Most common values (%)
Court position	67.6%	Restricted Area (37.1), Wing (31.8)
Shot type	67.6%	Jumper (65.0), Lay-up (21.3)
Shooting hand	67.5%	Right (85.0), Left (13.4)
Play type	68.6%	Catch&Shoot (30.0), Catch&Drive (11.6)
Contest	64.3%	Contested (87.1), Uncontested (12.9)
Dribble move	13.3%	Fake shot (22.7), Step back (21.2)
Drive direction	10.9%	Right (50.9), Left (49.1)

Table 34. T8 shot-detail annotation coverage for scoring attempts. Coverage indicates the percentage of scoring events with the attribute annotated. Core attributes (court position through contest) are available for ∼65% of events; specialized attributes (dribble move, drive direction) are rarer but provide additional tactical detail when present.

in the first frame (starting position) and a bounding box in the final frame (target position), defining the spatial endpoints that the generated motion must satisfy. Samples may contain up to three player specifications: the primary actor, a teammate, and a defender. Table 32 summarizes the distribution of spatial constraints in T8. The average normalized bounding-box area is 0.0153, indicating that players occupy relatively small frame regions. The mean spatial displacement between start and end boxes is 0.247 in normalized coordinates (median 0.222). Most actions involve medium or long spatial movements: 38.6% fall in the 0.15–0.30 range, 31.6% exceed 0.30, and only 5.3% involve minimal movement ($<$0.05). This confirms that the benchmark primarily targets dynamic player actions rather than static poses.

Action type distribution. Following the taxonomy in Table 33, actions are grouped into five categories: *shooting* (47.2%), *playmaking* (14.2%), *offensive set plays* (14.5%), *possession change* (14.4%), and *fouls* (7.8%). Shooting actions form the largest portion, with the most frequent events being 3 Pt Missed (12.6%), 2 Pt Missed (11.9%), and 2 Pt Made (10.3%). Playmaking is dominated by accurate passes (8.0%) and assists (6.2%). Among offensive set plays, Pick-and-Roll (10.5%) and Screen (3.5%) are most common. Possession-changing events are led by rebounds (9.8%), followed by steals (2.8%) and turnovers (1.8%). This distribution reflects a realistic mix of offensive, defensive, and possession-related events in basketball gameplay.

Annotation richness. For scoring-related actions, we provide additional structured annotations describing fine-grained play details (Table 34). Several fields have high coverage: *court position* (67.6%), *shot type* (67.6%), *shooting hand* (67.5%), and *play type* (68.6%). Among these, the most common shot type is *Jumper* (65.0%), followed by *Lay-up* (21.3%);

Restricted Area (37.1%) and *Wing* (31.8%) are the most frequent court locations. Defensive contest information is available for 64.3% of scoring actions, with the majority being contested shots (87.1%). More specialized annotations—dribble moves (13.3%) and drive directions (10.9%)—appear less frequently but provide additional tactical detail when present. These structured attributes are incorporated into the natural-language instructions when available.

T8: Evaluation Protocol

We evaluate goal-conditioned generation along three dimensions—spatial correctness, visual fidelity, and instruction-following—using automatic metrics and a human study.

Final-frame mIoU. We compute the intersection-over-union between the generated and ground-truth bounding boxes of the target player in the final frame, measuring whether the player successfully reaches the intended spatial goal.

Final-frame feature similarity. We compute cosine similarity between SigLIP v2 embeddings of the generated and ground-truth player crops in the final frame, assessing whether the appearance of the generated player is visually consistent.

Goal accuracy. We measure instruction-following ability using a video-language QA model fine-tuned on sports video understanding (described below). Given the generated video and the corresponding instruction, the QA model predicts whether the intended action outcome is successfully achieved. When the instruction contains additional play details (shot location, shot type, dribble move, shooting hand, defensive contest), the QA model also evaluates whether these attributes are correctly reflected. Goal accuracy is computed as the percentage of generated videos for which the QA model confirms that the specified objective and relevant attributes are satisfied.

Evaluation subset. Due to computational cost, ATI and MagicMotion are evaluated on a 100-clip subset. Our method is evaluated on both the same 100-clip subset (for direct comparison) and on 1,000 clips (to assess performance at scale).

T8: Baselines and Training Details

Our method. Our generation model uses the same Wan 2.1-Fun-1.3B-Control architecture and LoRA configuration as T7, with the addition of textual instruction conditioning via the text prompt. Baselines are ATI and MagicMotion in the same configurations as T7, without fine-tuning on our data.

Goal-conditioned action evaluator. To evaluate whether generated videos satisfy the specified goals, we fine-tune a video-language QA model as an automatic instruction-following evaluator. We adopt LLaVA-Video-7B-Qwen2 [102] as the backbone, consisting of a SigLIP-

SO400M vision encoder (ViT-SO400M/14 at 384×384) [98], a two-layer MLP projector with GELU activation, and a Qwen2-7B language model [92]. The vision encoder extracts spatiotemporal features from video frames, which are projected into the language space and processed by the language model for video QA.

We perform *full fine-tuning* of the entire model with a differentiated learning-rate schedule: the vision encoder uses 2×10^{-6}, while all other parameters use 1×10^{-5}. Training is conducted on 4$\times$ NVIDIA H100 GPUs using DeepSpeed ZeRO Stage-3 [64] for memory-efficient sharding. We use a per-device batch size of 2 with gradient accumulation over 2 steps (effective global batch size of 16), for 2 epochs with a cosine learning-rate scheduler (warmup ratio 0.03) under BF16 mixed precision. Each input video is uniformly sampled to 16 frames using the AnyRes strategy (up to 9 tiles per frame with spatial pooling stride 2), allowing a maximum sequence length of 32,768 tokens.

The training data contains 125,122 video QA pairs spanning eight question types that correspond to key action attributes used in T8 instructions: atomic action recognition (56,696 samples), shot type (13,499), spatial position (13,461), contested shot (12,969), shooting hand (12,019), play type (11,599), dribble move (2,684), and drive direction (2,195). In each training example, the target player is highlighted with a red bounding box, and the model answers a multiple-choice question about the player's action outcome or associated attributes.

T8: Analysis

Per-attribute goal realization. Table 35 measures how often the generated video successfully realizes each attribute specified in the instruction, as assessed by the fine-tuned QA evaluator. To establish that the evaluator is a credible verifier, we first report its accuracy on real (ground-truth) videos (GT column): it achieves 73–87% across most attributes, confirming that the evaluator reliably recognizes the target attributes when they are genuinely present. The main exception is dribble move (60.3%), indicating that this attribute is inherently difficult to recognize even in real footage—generation accuracy for this attribute (22.7%) should be interpreted with this lower ceiling in mind.

Turning to the generation results, our method achieves the highest accuracy on most attributes. The model is most successful at realizing appearance-related attributes: contested shot (86.0% vs. 87.4% GT ceiling), shooting hand (81.1% vs. 81.4%), and shot type (69.8% vs. 86.8%)—attributes largely determined by player pose and configuration in the final frames. For contested shot and shooting hand, our method nearly matches the ground-truth ceiling, suggesting that these attributes are effectively solved by the current approach. In contrast, attributes requiring coherent motion *throughout* the video are much harder to realize: dribble moves are correctly generated only 22.7% of the time (0% for both baselines; 60.3% GT ceiling), and the correct play type is produced 46.8%

Attribute	GT	ATI	MagicM.	Ours
Atomic action	73.4	32.4	20.6	**35.6**
Shot type	86.8	55.6	44.4	**69.8**
Shooting hand	81.4	77.8	77.8	**81.1**
Contested shot	87.4	77.8	77.8	**86.0**
Play type	76.2	33.3	22.2	**46.8**
Dribble move	60.3	0.0	0.0	**22.7**

Table 35. T8 goal realization accuracy (%) per attribute, as assessed by the fine-tuned QA evaluator. Our method is evaluated on 1,000 clips, while baselines are evaluated on a 100-clip subset. GT reports the QA evaluator's accuracy on real (ground-truth) videos measured on the same 1,000 clips, serving as an approximate upper bound and demonstrating that the evaluator is a credible verifier. Appearance-related attributes (contested shot, shooting hand) are most reliably realized; attributes requiring coherent motion throughout the video (dribble move, play type) remain challenging for all methods. MagicM. = MagicMotion.

of the time (vs. 33.3% for ATI and 22.2% for MagicMotion; 76.2% GT ceiling). This gap between static-appearance and dynamic-motion attributes highlights a key limitation: current generation models can place players in plausible configurations but struggle to produce the fine-grained motion sequences that define specific basketball actions.

Spatial accuracy by category. Table 36 breaks down our method's performance by action type, grouped into four categories. *Set plays* achieve the highest mIoU (0.335), likely because coordinated actions such as pick-and-rolls follow structured spatial patterns with predictable endpoints. *Ball handling* yields slightly lower spatial accuracy (0.306) but competitive feature similarity (0.408), suggesting that while player appearance is realistically generated, precise endpoint localization is harder for actions involving rapid changes of possession. *Scoring* actions exhibit the lowest spatial accuracy (0.294), as shooting motions involve larger displacements and more diverse final positions.

Goal realization by category. The pattern reverses for action recognition: *scoring* actions are most reliably realized (47.0%), likely because shot outcomes produce distinctive visual signatures (ball trajectory, net interaction) that are comparatively easy for the model to generate. *Set plays* are the hardest to realize correctly (16.7%), because coordinated multi-player actions like pick-and-rolls require generating realistic interactions between multiple agents—a challenge that goes beyond single-player motion fidelity.

Individual actions. At the per-action level, *Post* and *Pick&Roll* achieve the highest spatial accuracy (mIoU of 0.411 and 0.398), consistent with these being sustained positional plays near fixed court locations. *Free throws* achieve the highest feature similarity

(0.504), reflecting the spatially constrained and visually consistent nature of free-throw scenarios. The most striking contrast is between spatial and recognition metrics for set plays: Pick&Roll has high mIoU (0.398) but very low goal realization (9.2%), indicating that the model places players in roughly the right positions but fails to generate the coordinated motion pattern that makes the action recognizable as a pick-and-roll.

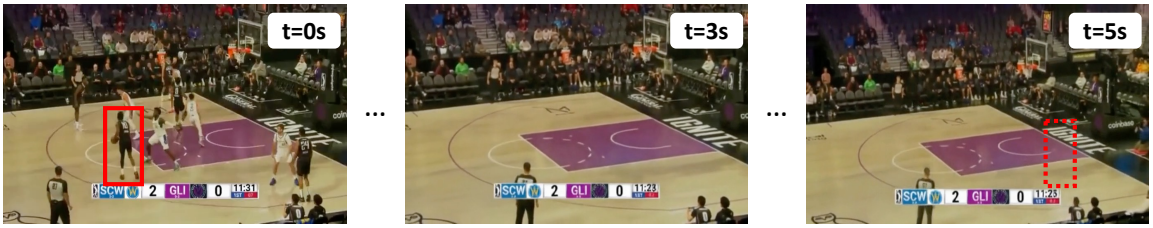
Additional qualitative examples. We provide additional qualitative comparisons between ATI, MagicMotion, and our method on another representative basketball example in Figure 40, using the same task setup and visual annotations as Figure 18. Unlike Figure 18, which presents a multi-player setting with simultaneous spatial constraints, this example focuses on a simpler single-player scenario. Despite the reduced complexity, all methods still struggle to fully satisfy the task requirements. Our method is the only approach that successfully places the player at the target location; however, the generated player appearance remains blurry and the action does not fully match the intended motion. ATI and MagicMotion both fail to move the player to the correct final position.

			Generation		QA
Category	**Action**	n	**mIoU**	**Feat.**	**Acc.**
Scoring	3 Pt Missed	125	.304	.428	40.7
	2 Pt Made	87	.247	.337	42.7
	2 Pt Missed	72	.321	.422	69.5
	3 Pt Made	59	.292	.398	24.2
	Free throw	16	.362	.504	87.5
	Subtotal	*364*	*.294*	*.403*	*47.0*
Ball Handling	Steal	91	.240	.323	68.1
	Turnover	32	.400	.521	12.5
	Post	29	.411	.551	10.3
	Subtotal	*152*	*.306*	*.408*	*45.4*
Set Plays	Pick&Roll	114	.398	.504	9.2
	Screen	39	.281	.398	11.1
	Assist	79	.272	.356	30.5
	Subtotal	*232*	*.335*	*.436*	*16.7*
Def./ Foul	Foul	202	.311	.422	23.0
	Rebound	50	.304	.414	60.0
	Subtotal	*252*	*.309*	*.420*	*30.3*
Overall		**1000**	**.309**	**.415**	**35.6**

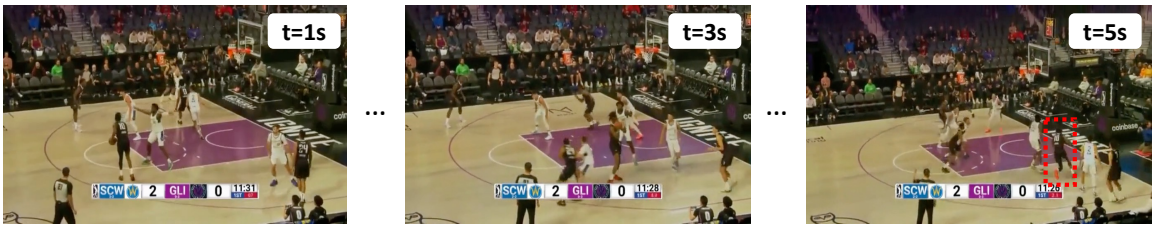
Table 36. T8 per-action breakdown for our method (1,000 clips). We report last-frame mIoU (spatial accuracy), feature similarity (visual fidelity), and goal realization accuracy (%). Set plays achieve the highest spatial accuracy (mIoU = .335) but the lowest goal realization (16.7%), indicating that the model places players in structured positions but fails to generate the coordinated motion that makes the action recognizable. Scoring actions are most reliably generated (47.0%), likely because shot outcomes produce distinctive visual signatures.

Goal: Simulate the player in the **red** bounding box assisting another player in making a two-point shot, with the player ending at the location marked by the **dotted red** bounding box.

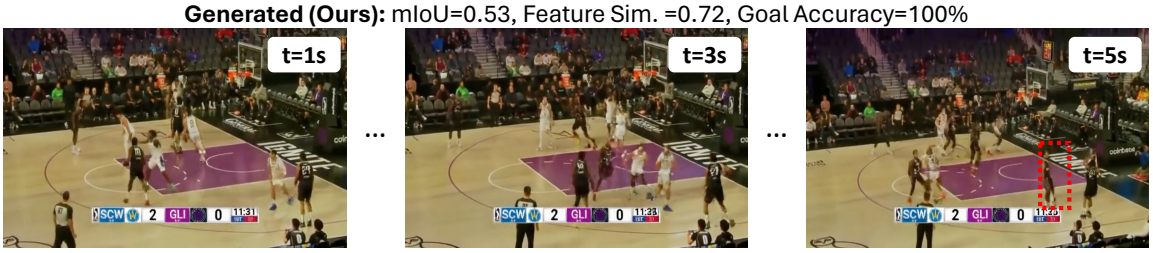
Initial Frame + Player-Removed Background Video w/ a Spatial Target



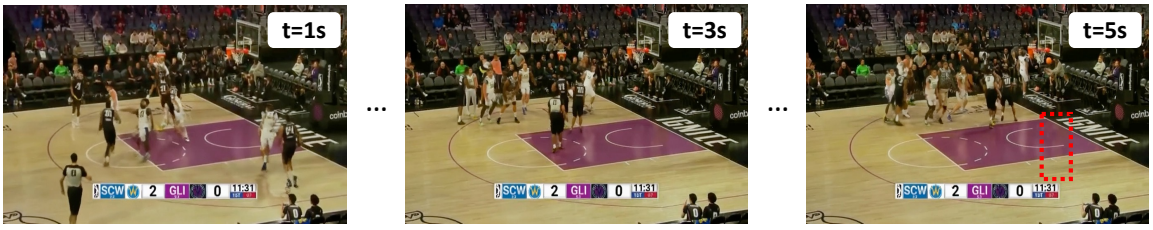
Ground-Truth Reference



Generated (Ours): mIoU=0.53, Feature Sim. =0.72, Goal Accuracy=100%



Generated (ATI): mIoU=0, Feature Sim. =0, Goal Accuracy=0%



Generated (MagicMotion): mIoU=0, Feature Sim. =0, Goal Accuracy=0%

Figure 40. T8 Goal-Conditioned Action Generation. The **red** bounding box marks the player's starting position; the **dotted red** bounding box marks the required target position in the final frame.

	Basketball	Hockey	Soccer	Total
General				
Seasons	5	4	4	—
Games	3,307	2,482	1,641	7,430
Players	952	1,533	2,031	4,516
Teams	30	32	51	113
Documents				
Game reports	3,307	2,482	1,070	6,859
Statistical documents	9,547	9,130	7,771	26,448
Video clips				
Total clips	854,949	715,574	189,707	1,760,230
Avg. clips / game	258.5	288.3	115.6	236.9
Avg. clip duration (s)	10.9	12.1	12.9	11.6
Total hours	2,587.6	2,404.8	677.4	5,669.9

Table 37. T9 environment statistics across basketball, hockey, and soccer. The corpus comprises 1.76M video clips of 5,670 hours in total and 33K documents across five types.

E Pillar 4: Agentic Synthesis (T9)

E.1 T9: Cross-Corpus Agentic Reasoning

T9: Data Construction and Statistics

Corpus statistics. Table 37 summarizes the corpus. The database contains 7,430 games across the three sports. The video corpus consists of 1.76M clips (avg. 237 per game, avg. duration 11.6 s) of 5,670 hours. The document corpus contains 33,307 documents: 21,719 game-level documents (game reports and per-player/per-team game statistics) and 11,588 entity-level seasonal documents (11,151 player–season and 437 team–season entries across 4,516 unique players and 113 teams).

Question generation. We generate candidate questions using GPT-5.2 with a structured prompt (Prompt 41), instantiated per sport. For each game, the model receives the corresponding game report, player and team statistics (game-level and season-level), and the play-by-play game log, together with sport-tailored in-context examples. The prompt enforces four constraints: (1) every answer must require video evidence, (2) each question must identify its target game through a compound description drawing on multiple source types, (3) usage must be balanced across answer-attribute categories and reasoning patterns, and (4) answer-value distributions must prioritize rare attribute values rather than mirror corpus-level frequencies, preventing statistical priors from making

questions trivially guessable. Per-game yield varies by sport (10, 5, and 20 candidate questions per game for basketball, hockey, and soccer respectively), reflecting differences in event density and annotation granularity. To control the diversity and difficulty of the generated questions, we annotate each one along four complementary facets:

Search patterns — the reasoning primitives required to solve the question. We define 14 primitives that span three levels of complexity. *Basic retrieval* primitives include textual search (matching descriptions in game reports), visual search (locating events in video), temporal filtering (restricting to a time window), and sequencing (ordering events along a timeline). *Reasoning* primitives include state tracking (monitoring evolving game state), aggregation (counting across events), comparison (contrasting quantities), backtracking (tracing events backward from a reference point), and pattern categorization (classifying events into predefined types). *Compositional* primitives include logic (deductive reasoning from constraints), loop (iterating over sets of events or players), math (arithmetic), conditional branching (following different reasoning paths based on intermediate results), and visual analysis (fine-grained interpretation of player actions). Each question combines an average of 4.0 of these primitives across the tri-sport pool (range 2–7).

- **Game identification** — how the question specifies which game it refers to. Rather than naming the game directly (e.g., "Lakers vs. Celtics on Jan. 15"), each question identifies its target game through a compound description that the agent must resolve. For example, a question might begin: *"In the game where Charlotte completed a fourth straight win and Kemba Walker doubled his season three-point average ..."* — requiring the agent to cross-reference a narrative detail (winning streak) from game reports with a statistical anomaly from player stats to pinpoint the correct game. These identifiers combine report-based narrative details, statistical anomalies, and temporally distinctive patterns, ensuring that game identification itself is a non-trivial retrieval problem.
- **Required sources** — which modalities the agent must consult to arrive at the answer. Every question requires video evidence, and each additionally draws on a different combination of textual game reports, player/team game statistics, and season-level statistics. This ensures that no single source type is sufficient and that multi-source retrieval is always necessary.
- **Answer attribute types** — the category of the final answer, which determines what the agent must ultimately recognize in the video. We enforce a balanced distribution across eight types: *spatial position* (e.g., "from the wing"), *shot type* (e.g., "floater"), *play type* (e.g., "pick-and-roll"), *player name*, *player number*, *game scores*, *dribble move* (e.g., "step back"), and *numeric counts* (e.g., "how many threes"). Balancing across these types prevents the benchmark from being solvable through answer-type priors.

Question statistics. The 1,000 final questions span all three sports (334 basketball / 333

hockey / 333 soccer), drawn from hundreds of unique games and covering several hundred distinct search-pattern combinations. On average, answering a question requires 2–5 video clips together with a couple of textual or statistical documents, confirming that both visual and non-visual multi-hop retrieval are necessary.

Question validation. We apply a five-stage validation pipeline to filter the 3,280 candidates down to 1,000 high-quality questions:

- *Stage 1: Document-statistics filter.* Each candidate is sent to GPT-5.2 with the game report and box-score statistics but without the play-by-play log. Questions answered correctly are removed, as they do not require video-level evidence.
- *Stage 2: Full-context review.* Remaining questions are sent to GPT-5.2 with all available sources including the game log. Questions answered incorrectly despite complete information are flagged as potentially malformed or ambiguous and manually reviewed.
- *Stage 3: Blind LLM guessability.* Each question is presented to five LLMs (Qwen3-235B-A22B-FP8, DeepSeek-R1-Distill-Llama-70B, Gemma3-27B-IT, GPT-5.2, and Llama-3.3-70B) with no game context. Any question answered correctly by any model is removed, as it is answerable from the model's knowledge alone.
- *Stage 4: Balanced subsampling.* We subsample to the target per-sport question count using a water-fill allocation algorithm that balances counts across attribute categories and answer values. Within each value group, questions are selected to maximize unique game coverage.
- *Stage 5: Manual refinement.* We manually check for questions not uniquely tied to a specific game and replace them with high-quality alternatives from the remaining pool.

T9: Evaluation Protocol

We use exact-match accuracy graded by an LLM judge (GPT-5.2) using the Prompt 42. Each model is evaluated in two modes: *default mode*, which uses the full visual pipeline, and *oracle mode*, which provides ground-truth event captions in place of video frames. Comparing the two modes isolates the contribution of visual perception from the agent's reasoning and planning capabilities.

T9: Baselines and Implementation Details

Agent framework. All models use a ReAct-style [94] agent loop with global planning. At each turn, the agent first reasons in a structured `<think>` block (decomposing the query, tracking progress, planning the next action), then executes exactly one tool call. The loop continues for up to 30 LLM calls or until the agent produces a final answer.

The system prompt (Prompt 43) enforces five critical rules: (1) no use of internal (pretrained) knowledge—answers must be grounded in tool observations; (2) a search-

first policy requiring at least one search before answering; (3) strict ID integrity (never guess or modify document/video IDs); (4) consistency between the reasoning plan and executed actions; and (5) a mandatory fallback statement when evidence is insufficient.

Context management. Each agent uses its model's default supported context window, capped by what our serving node allows. GPT-5.2 runs via the OpenAI API and uses its native context. The open-source Qwen agents (Qwen3-235B, Qwen3-Omni-30B, Qwen3-32B) are served locally with vLLM on an 8×A6000 (48GB) node, all with a 40K-token context window; MiniMax M2.5 runs on a 4×H100 (80GB) node with a 128K-token window.

Search backends. Document search uses Elasticsearch with hybrid retrieval: semantic similarity via BGE-M3 [13] embeddings and BM25 lexical matching, fused via reciprocal rank fusion (RRF). The top-20 results are returned with 500-character highlighted snippets. Video search uses fine-tuned InternVideo2 video embeddings in default mode and BGE-M3 text-caption embeddings in oracle mode, both returning the top-100 clips.

Document analysis. When the agent calls `document_qa`, the full document content is passed to Qwen3-VL-8B with the prompt shown in Prompt 44. The model returns a JSON response containing the answer, a confidence score, and supporting evidence quoted from the document. Responses are capped at 32,000 tokens.

Video analysis. In default mode, `video_qa` extracts 16 frames per clip at 336×336 resolution and sends them to a visual QA model (a fine-tuned LLaVA-NeXT-Video or Qwen3-Omni-30B-Thinking) with the prompt shown in Prompt 45. In oracle mode, video frames are replaced with ground-truth structured event captions from the database (e.g., *"Pascal Siakam performed 2 PT Made from Restricted Area with a Floater during a Hand off..."*), providing a performance ceiling that measures the agent's reasoning ability independently of its visual perception.

Evaluated models. We evaluate five agent models, each in both default and oracle mode:

- **GPT-5.2**: accessed via the OpenAI API with medium reasoning effort, temperature 1.0, max output 4,096 tokens.
- **Qwen3-235B**: Qwen3-235B-A22B-FP8-Thinking served via vLLM with prefix caching, temperature 0.6, top-p 0.95, max output 4,096 tokens.
- **Qwen3-Omni-30B**: Qwen3-Omni-30B-A3B-Thinking served via vLLM with prefix caching, temperature 0.6, top-p 0.95, max output 4,096 tokens.
- **Qwen3-32B**: Qwen3-32B-A3B-Thinking (text-only) with identical serving configuration as Qwen3-Omni-30B.
- **MiniMax-M2.5**: served via vLLM with prefix caching, temperature 0.6, top-p 0.95, max output 4,096 tokens.

All models use the same system prompt, tool definitions, and hyperparameters.

Benchmark QA-Pair Generation Prompt

```
Role: You are an expert benchmark generator for Video-Language Agentic Search.
Your task is to generate N high-quality, complex, and diverse QA pairs for the single-game data.

Search Vocabulary: textual search, visual search, visual analysis, temporal
filtering, loop, state tracking, backtracking, conditional branching,
aggregation, logic, comparison, pattern categorization, math, sequencing.

Game Identifier & Query Corpus Vocabulary: report, video, player_game_statistics,
team_game_statistics, player_season_statistics, team_season_statistics.

## S1. STYLE & OUTPUT CONSTRAINTS
1. Implicit Sources: Never mention "report," "statistics," or "video" in the
   question. Phrases such as "according to the report" or "as annotated" are prohibited.
2. No Report Answers: Final answers must always derive from Statistics or Video.
3. Game-Log Timing: Timestamps denote remaining time within a period.
4. Video Necessity: Every question must functionally require video.
5. No Yes/No: Each question must elicit a specific value.
6. Natural Phrasing: Vary sentence structures.
7. No Hallucinations: All terms must be explicitly present in the source data.

## S2. COMPOUND GAME IDENTIFICATION
Each question identifies the target game via a compound identifier composed
of distinct source types. Direct game IDs are prohibited. Rotate among:
1. Report-Based Facts: narrative details that uniquely fingerprint a game.
2. Statistical Anomalies: outlier metrics unlikely to recur.
3. Temporal Patterns: rare in-game sequences.
4. Numerical values use ambiguity markers ("nearly," "at least," "exceeding").

## S3. QUERY COMPLEXITY & DIVERSITY
1. Information Siloing: The final answer must be an attribute that cannot be
   inferred from non-video sources alone.
2. Attribute Rotation: Rotate final-answer attributes across the
   sport-specific final-answer categories (e.g., player name/number, action
   type, shot/play-type, spatial location, and game score).
3. Logical Diversity: Mix sequencing, conditioning (with mandatory
   "otherwise" branch; <=1 per set), backtracking, comparison, temporal
   recurrence, negation/exclusion, looping, and math operations.
4. Balanced Value Distribution: Prioritize rare attribute values.
5. Corpus-Pattern Diversity: Vary source routing across questions.

## OUTPUT FORMAT
A JSON list of N objects, each containing: id, category, question, answer, grounded_evidence,
search_pattern, identifier_corpus_pattern, query_corpus_pattern, and final_query_attribute.

## CONTEXT
{IN_CONTEXT_EXAMPLES}
{CONTEXT}
```

Figure 41. System prompt for benchmark QA-pair generation. The template is instantiated per sport by varying N, in-context examples, and the final-answer attribute schema. It enforces constraints on (1) *style*—natural, implicit phrasing with mandatory video dependency; (2) *game identification*—compound identifiers with ambiguity markers; and (3) *query complexity*—diverse logical operations, attribute rotation, and corpus routing patterns. `{CONTEXT}` and `{IN_CONTEXT_EXAMPLES}` are replaced at runtime with single-game data and sport-tailored exemplar QA pairs.

LLM-as-a-judge Prompt for T9 (Answer Equivalence)

```
You are grading a Q&A pair about sports (basketball, soccer, or hockey).
Interpret domain-specific terms according to their standard meaning in the
relevant sport.

Ground Truth: {gt_answer}
Prediction: {pred_answer}

Task: Decide whether the Prediction conveys the same information as the
Ground Truth.

Rules:
- Wording can differ. Focus on semantic meaning, not exact phrasing.
- Synonyms, paraphrases, and extra context are fine as long as they do not
  change the meaning.
- The Prediction must cover the key information in the Ground Truth. If it
  drops a required detail that changes the meaning, mark Wrong.
- For numerical answers, the numeric value must match exactly.
- If the Prediction says "Not found", "I cannot find...", or similar, and
  the Ground Truth is a specific value, mark Wrong.
- "0" is a valid numeric answer and is NOT equivalent to "not found".

Return exactly one JSON object on a single line:
{"verdict": "Right" | "Wrong", "reason": "<one short sentence>"}
```

Figure 42. T9: LLM-as-a-judge prompt. The judge model evaluates whether a predicted answer conveys the same information as the ground-truth answer for sports Q&A, allowing paraphrases and synonymous wording while requiring exact matches for numerical answers. The judge outputs a single-line JSON object with a binary verdict and a short reason.

System Prompt for T9 Agent

Role. You are a precise research assistant specializing in multimodal information retrieval from the provided Video Database and Document Database.

Critical Rules.

1. **No Internal Knowledge**: Answer strictly based on Observations. Do not use pre-trained facts.
2. **Search-First Policy**: Execute at least one search tool before providing a final answer.
3. **ID Integrity**: Use exact `doc_id` or `video_id` as returned by tools.
4. **Consistency**: Your plan in `<think>` must align with the subsequent tool call.
5. **Fallback**: If no information is found after exhaustive search, state so explicitly.

Execution Protocol: ReAct with Global Planning.

1. **Reasoning** (`<think>` tags): *Initial turn*—create a Global Plan decomposing the query into sub-steps with assigned tools. *Subsequent turns*—refine the plan based on new observations. *Final turn*—collect all evidence and reason toward a conclusion.
2. **Action**: Execute exactly one tool call per turn. If sufficient evidence is gathered, output `<answer>result</answer>`.

Search Instructions. Decompose complex queries into targeted sub-questions. If a search returns no results, adjust keywords or time ranges. Verify partial information with QA tools before concluding.

Answer Format. The content inside `<answer>` must be a concise value (word, number, or short phrase).

Environment. `{{env_desc}}`

Figure 43. T9 system prompt for the ReAct-style agent. The prompt enforces five rules: (1) answers must be grounded in tool observations, not internal knowledge; (2) at least one search must precede any answer; (3) document and video IDs must never be guessed or modified; (4) reasoning plans must be consistent with executed actions; and (5) a fallback statement is required when evidence is insufficient. `{{env_desc}}` is replaced at runtime with the full environment description and tool signatures (Figure 46).

Prompt for Document QA Tool

You are a helpful assistant. Read the provided document content and answer the user's question with high accuracy. If you can't find the relevant information to answer, you should return "there is no evidence" as the answer and confidence level as 1.0.

Document Content: `{{document_content}}`

Question: `{{query}}`

You MUST provide your answer in the following JSON format:

```
{   "answer": "Your detailed answer here.",
"confidence": "Confidence level between 0.0 and 1.0.",
"evidence": "Quote or reference from the text supporting the answer."
}
```

Figure 44. T9 document QA prompt. When the agent calls `document_qa`, the full document content and the agent's query are injected into this template. The model returns a structured JSON response containing the answer, a confidence score, and supporting evidence quoted from the document. `{{document_content}}` and `{{query}}` are replaced at runtime.

Prompt for Video QA Tool

You are a helpful video assistant. Analyze the visual content of the provided video clip and answer the user's question. If you can't find the relevant information to answer, you should return "there is no evidence" as the answer and confidence level as 1.0.

Question: `{{query}}`

Please provide your answer in the following JSON format:

```
{   "answer": "Your detailed answer based on visual analysis.",
"confidence": "Confidence level between 0.0 and 1.0.",
"evidence": "Description of the visual frames that support your answer."
}
```

Figure 45. T9 video QA prompt. When the agent calls `video_qa`, 16 frames are extracted from each specified clip and presented alongside the agent's query. In oracle mode, frames are replaced with ground-truth captions. The model returns the answer and a confidence score. `{{query}}` is replaced at runtime with the sub-question generated by the agent at each reasoning step.

T9 Tool API Signatures

```
search_documents(
    query: str,
    game_ids: list[str],
    doc_type: str,
    teams: list[str],
    players: list[str],
    season: int,
    date: str // "2019-04-28" or "04-01..04-30"
)
if doc_type is "game_report":
→ list[{doc_id, score, game_id, teams, highlights}]
if doc_type is "game_stat_player":
→ list[{doc_id, score, game_id, teams, players,}
  {season, date, highlights}]
if doc_type is "game_stat_team":
→ list[{doc_id, score, game_id, teams, season,}
  {date, highlights}]
if doc_type is "season_stat_player":
→ list[{doc_id, score, teams, players, season,}
  {highlights}]
if doc_type is "season_stat_team":
→ list[{doc_id, score, teams, season, highlights}]
```

```
document_qa(
    doc_ids: list[str],
    query: str
) → list[{doc_id, answer, confidence, evidence}]
```

```
search_videos(
    query: str,
    game_ids: list[str],
    quarter: int,
    teams: list[str],
    players: list[str],
    temporal_boundary: str // "100.5-120.0"
) → list[{clip_id, score, game_id, quarter,}
  {teams, players, timestamps}]
```

```
video_qa(
    video_ids: list[str],
    query: str
) → list[{clip_id, answer, confidence}]
```

Figure 46. T9 tool API signatures. Each `search_*` tool accepts a free-text query with optional structured metadata filters and returns ranked results; returned fields vary by document type. Each `*_qa` tool takes specific item IDs and a question, processing each item independently. In oracle mode, `search_videos` uses ground-truth event captions with text embeddings, and `video_qa` replaces visual frame analysis with ground-truth captions.

	Video QA Model		
Agent	LLaVA-Video-7B-ft	Qwen3-Omni-30B	Δ
GPT-5.2	4.5	5.4	+0.9
Qwen3-Omni-30B	1.8	1.5	−0.3

Table 38. T9 ablation: effect of the video QA model on overall accuracy (%). Neither the fine-tuned nor the larger VLM lifts T9 accuracy above 6%. T9 requires accurate, fine-grained analysis across multiple video clips, where a single misinterpretation is enough to produce an incorrect final answer. These results suggest that video understanding remains the primary bottleneck for robust cross-corpus agentic reasoning.

		GPT-5.2		**MiniMax M2.5**		**Qwen3-235B**		**Qwen3-32B**		**Qwen3-Omni-30B**	
Sport	Tool	Default	Oracle	Default	Oracle	Default	Oracle	Default	Oracle	Default	Oracle
Basketball	`search_doc`	7.75	6.57	8.31	6.52	2.62	2.70	1.59	1.66	1.83	1.79
	`search_video`	4.81	2.56	3.74	2.31	1.16	1.11	1.05	1.06	1.05	1.03
	`doc_qa`	3.32	2.88	3.62	3.05	3.10	3.09	1.36	2.37	1.23	3.51
	`video_qa`	8.89	3.20	7.79	3.85	6.58	4.52	3.04	1.99	1.50	1.28
Hockey	`search_doc`	6.70	5.67	8.58	6.97	2.04	2.04	1.21	1.09	1.29	1.14
	`search_video`	2.77	1.94	2.94	1.86	0.57	0.67	0.57	0.68	0.49	0.62
	`doc_qa`	3.13	2.44	3.64	2.98	1.66	1.55	0.33	0.27	0.59	0.59
	`video_qa`	4.59	2.15	4.85	3.29	1.41	1.52	1.20	1.29	0.40	0.60
Soccer	`search_doc`	8.35	6.24	8.96	7.96	1.75	1.68	1.02	0.92	0.93	0.83
	`search_video`	3.94	1.95	2.95	1.83	0.63	0.70	0.73	0.90	0.68	0.72
	`doc_qa`	2.45	1.88	3.19	2.99	1.42	1.23	0.21	0.18	0.49	0.41
	`video_qa`	5.39	1.90	5.09	2.40	1.79	1.24	1.36	1.54	0.77	0.86

Table 39. T9 average tool calls per question. GPT-5.2 makes ∼21 tool calls per question in default mode across the three sports—roughly 3–5× more than any open-source Qwen model—reflecting substantially deeper exploration of the search space. Under oracle mode, GPT-5.2's total drops to ∼14, as accurate captions reduce the need for repeated video verification. In oracle mode, `video_qa` maps to `video_qa_oracle`.

T9: Analysis

Effect of video QA model. Table 38 examines the effect of replacing the default Video QA model (LLaVA-Video-7B-ft) with Qwen3-Omni-30B on the basketball subset ($n = 334$). With LLaVA-Video-7B-ft, T9 accuracy reaches only 4.8% for GPT-5.2 and 1.8% for Qwen3-Omni-30B. Replacing it with Qwen3-Omni-30B (39.54% on T2) yields 5.4% and 1.5%, respectively. Neither configuration lifts T9 accuracy above 6%, as T9 requires correctly analyzing multiple retrieved clips in sequence, where a single error on any clip cascades into a wrong final answer.

		GPT-5.2		MiniMax M2.5		Qwen3-235B		Qwen3-32B		Qwen3-Omni-30B	
Sport	Outcome	Default	Oracle	Default	Oracle	Default	Oracle	Default	Oracle	Default	Oracle
Basketball	Overall	23.6	15.0	23.6	16.5	10.9	9.7	6.0	5.4	5.2	5.1
	Success	16.3	12.3	15.0	12.9	11.6	8.0	6.0	5.0	5.0	4.6
	Failure	23.9	17.6	24.0	18.7	10.9	10.1	6.0	5.4	5.2	5.2
Hockey	Overall	18.1	13.2	21.2	16.6	6.8	6.9	4.9	5.0	6.1	5.3
	Success	14.3	11.2	20.8	12.8	6.5	6.6	4.2	4.0	3.2	4.0
	Failure	18.4	15.5	21.3	19.3	6.8	7.0	4.9	5.1	6.2	5.4
Soccer	Overall	21.1	13.0	19.5	15.4	6.7	5.9	5.4	5.1	5.7	4.7
	Success	9.7	10.5	9.8	11.1	5.0	5.5	4.4	4.5	3.9	3.6
	Failure	21.5	16.4	19.9	18.2	6.7	6.0	5.5	5.2	5.7	4.9

Table 40. T9 average conversation turns per question by outcome and sport. GPT-5.2 explores more actively ($\sim$21 turns on average across the three sports) but drops to $\sim$14 turns under oracle mode, suggesting that much of its default-mode effort is spent recovering from perception errors. In contrast, the smaller Qwen models take only $\sim$5 turns regardless of mode or outcome, suggesting premature termination with insufficient evidence, consistent with overall T9 results.

Conversation length analysis. Table 40 breaks down the average number of conversation turns by outcome. For GPT-5.2, successfully answered questions require fewer turns ($\sim$13) than failed ones ($\sim$21), suggesting that longer trajectories often reflect unproductive search rather than deeper reasoning. Under oracle, the overall turn count drops from $\sim$21 to $\sim$14, as accurate captions reduce the need for repeated verification of video content. These results indicate that GPT-5.2 spends a significant portion of its budget on failed video retrieval attempts in the default setting.

Cross-judge agreement and self-preference. Table 42 reports per-judge per-sport accuracy on a 300-instance sample (100 instances per baseline model; 33–34 per sport). The same model responses are re-judged with Qwen3-235B, Gemini-3-Flash, and DeepSeek-V3.2. Model rankings remain stable across all four judges, with mean pairwise Spearman $\rho = 0.93$ (≥ 0.87 pairwise) and $\geq 97\%$ pairwise agreement ($\kappa \geq 0.94$). The mean absolute accuracy shift relative to the GPT-5.2 primary judge is at most 1.33 percentage points per alternative judge. Table 43 further compares GPT-5.2's scoring of each baseline model's outputs against the mean scoring by the three non-GPT judges. GPT-5.2 scores its own outputs identically to the non-GPT judges ($\Delta = 0.0$pp), and the mean absolute GPT–non-GPT scoring gap across the three baseline models is only 0.22pp. Overall, these results indicate that T9 evaluation is robust to the choice of LLM judge, with no evidence that the GPT-5.2 judge favors GPT-generated outputs.

Comparison	Agreement (%)	Cohen's κ
Annotator A vs LLM judge	98.0	0.96
Annotator B vs LLM judge	96.7	0.93
Annotator A vs Annotator B	98.7	0.97
Mean (LLM vs human)	**97.3**	**0.95**

Table 41. T9 LLM-judge vs human agreement ($n = 150$). Two annotators independently re-judge a stratified sample of 150 model answers using the LLM-judge instructions. The LLM judge matches human verdicts on 97.3% of items on average, with inter-annotator agreement $\kappa = 0.97$.

Model output	Sport	Judge1	Judge2	Judge3	Judge4
GPT-5.2 outputs	Basketball	50.0	47.1	55.9	47.1
	Hockey	60.6	60.6	60.6	54.5
	Soccer	57.6	57.6	60.6	60.6
Qwen3-32B outputs	Basketball	33.3	33.3	33.3	33.3
	Hockey	32.4	32.4	32.4	32.4
	Soccer	39.4	39.4	39.4	45.5
Qwen3-Omni-30B outputs	Basketball	27.3	27.3	27.3	27.3
	Hockey	39.4	39.4	39.4	39.4
	Soccer	38.2	38.2	38.2	38.2

Table 42. T9 per-judge per-sport accuracy on the cross-judge sample ($n = 300$). Three baseline models (GPT-5.2, Qwen3-32B, Qwen3-Omni-30B; 100 instances each, 33–34 per sport) are evaluated by four LLM judges: GPT-5.2, Qwen3-235B, Gemini-3-Flash, and DeepSeek-V3.2. Rankings are stable across judges (pairwise Spearman $\rho \geq 0.87$; mean 0.93).

Model output	Judge1	Judge2	Judge3	Judge4	Mean(J2–4)	Δ
GPT-5.2 outputs	56.0	55.0	59.0	54.0	56.0	+0.00
Qwen3-32B outputs	35.0	35.0	35.0	37.0	35.7	-0.67
Qwen3-Omni-30B outputs	35.0	35.0	35.0	35.0	35.0	+0.00

Table 43. T9 self-preference check. For each baseline model's outputs (100 instances each), we report accuracy under the primary judge (Judge1) and three alternative judges (Judge1 = GPT-5.2, Judge2 = Qwen3-235B, Judge3 = Gemini-3-Flash, Judge4 = DeepSeek-V3.2). Δ is the Judge1 accuracy minus the mean of Judges 2–4. The mean absolute Δ across the three baseline models is 0.22pp, indicating no evidence of self-preference bias in the T9 judge.

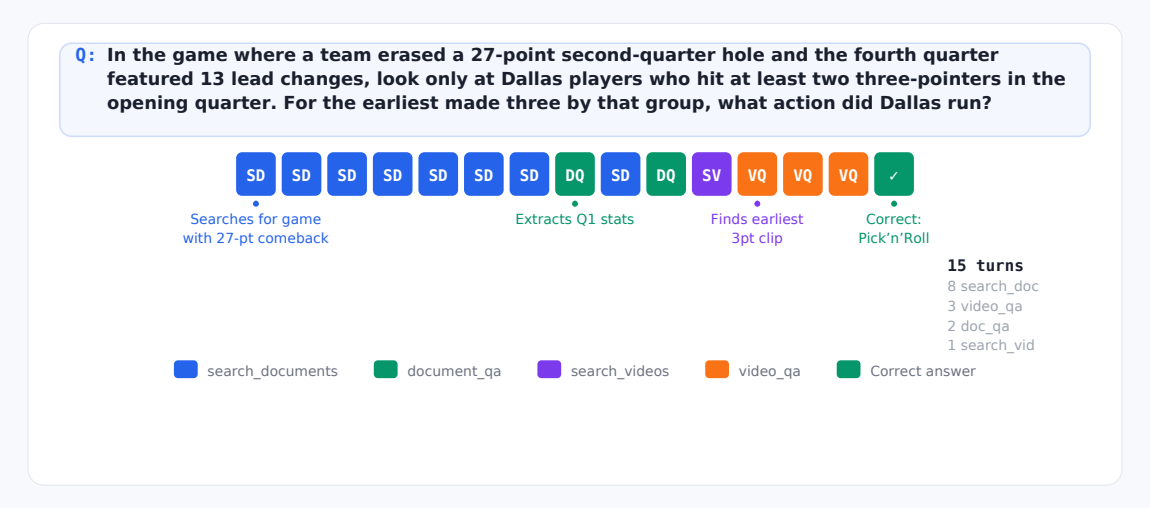


Figure 47. T9 Cross-Corpus Agentic Reasoning: given a complex question and access to a multimodal database of thousands of games, the agent must retrieve and reason over documents and video clips to produce a final answer. This figure shows a **successful** GPT-5.2 trajectory. Each colored block represents one tool call; the timeline reads left to right across turns. The agent identifies the target game via document search (Turns 1–7 and 9), extracts the relevant players' statistics via document QA (Turns 8 and 10), retrieves the earliest three-pointer clip via video search (Turn 11), and determines the play type via video QA (Turns 12–14), arriving at the correct answer in 15 turns. This compact, purposeful trajectory—8 document searches, 1 video search, 3 video QA calls—illustrates the efficient cross-modal reasoning required for T9.

Trajectory case studies. Figures 47–49 provide additional trajectory-level examples of T9 behavior beyond the modality-switch failure discussed in Figure 20. The successful case in Figure 47 shows the intended pattern, where the agent first narrows the search space through document search and document QA, then switches to video search and video QA when visual evidence is needed. The two failure cases highlight complementary bottlenecks. In Figure 48, the agent retrieves the right evidence and selects the appropriate tools, but the video QA model cannot reliably extract the required visual information. In Figure 49, the agent identifies useful textual context but terminates before attempting video retrieval. Together, these trajectories show that T9 failures can arise when agents remain in the wrong modality, stop before exploring the needed modality, or reach the correct clips but fail at visual perception. They also illustrate why T9 requires not only large-scale retrieval, but also effective tool selection, persistence, and reliable grounding across documents and video. Figures 50–52 provide additional qualitative examples showing the detailed conversation at each step, including the agent's reasoning, tool calls, and tool responses.

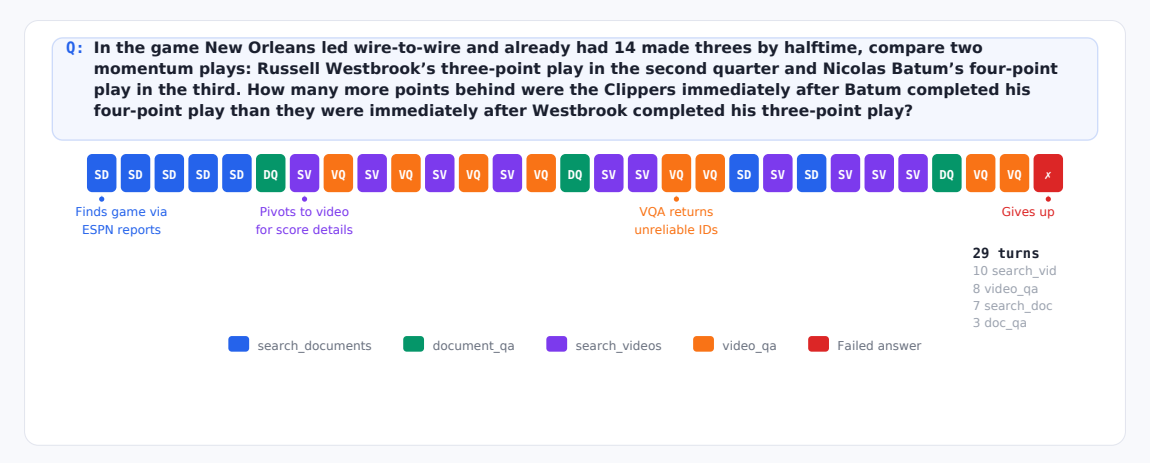


Figure 48. T9 failure case: **video perception bottleneck**. Unlike the modality-switch failure in Figure 20, the agent's planning and tool selection are correct. It efficiently identifies the target game via document search (Turns 1–4), confirms the teams involved, and retrieves the relevant players from statistics. It then correctly pivots to video search and video QA to resolve the score details. But this is where everything breaks down: the agent makes 8 video QA calls asking essentially the same question across different clips, each time receiving an unreliable answer from the VLM. After 29 turns—10 video searches, 8 video QA calls, and several document fallbacks—the agent gives up. Correct retrieval, correct planning, but the visual perception model cannot deliver the answer.

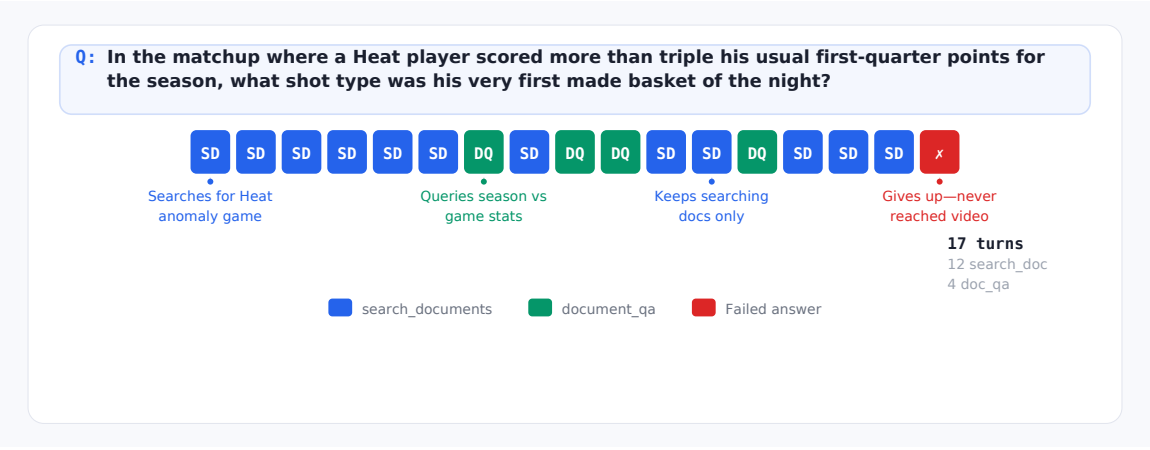


Figure 49. T9 failure case: **early termination without video**. This shows a more subtle variant of modality-switch failure compared to Figure 20. The agent identifies the correct game and queries statistics, but gives up after only 17 turns—well short of the 30-turn limit—without ever issuing a video search. Whereas Figure 20 shows an agent that exhausts its budget in the wrong modality, this case shows an agent that *stops trying* when document evidence proves insufficient, rather than switching to video. The timeline is dominated by document searches (12) and document QA (4), with no video calls at all.

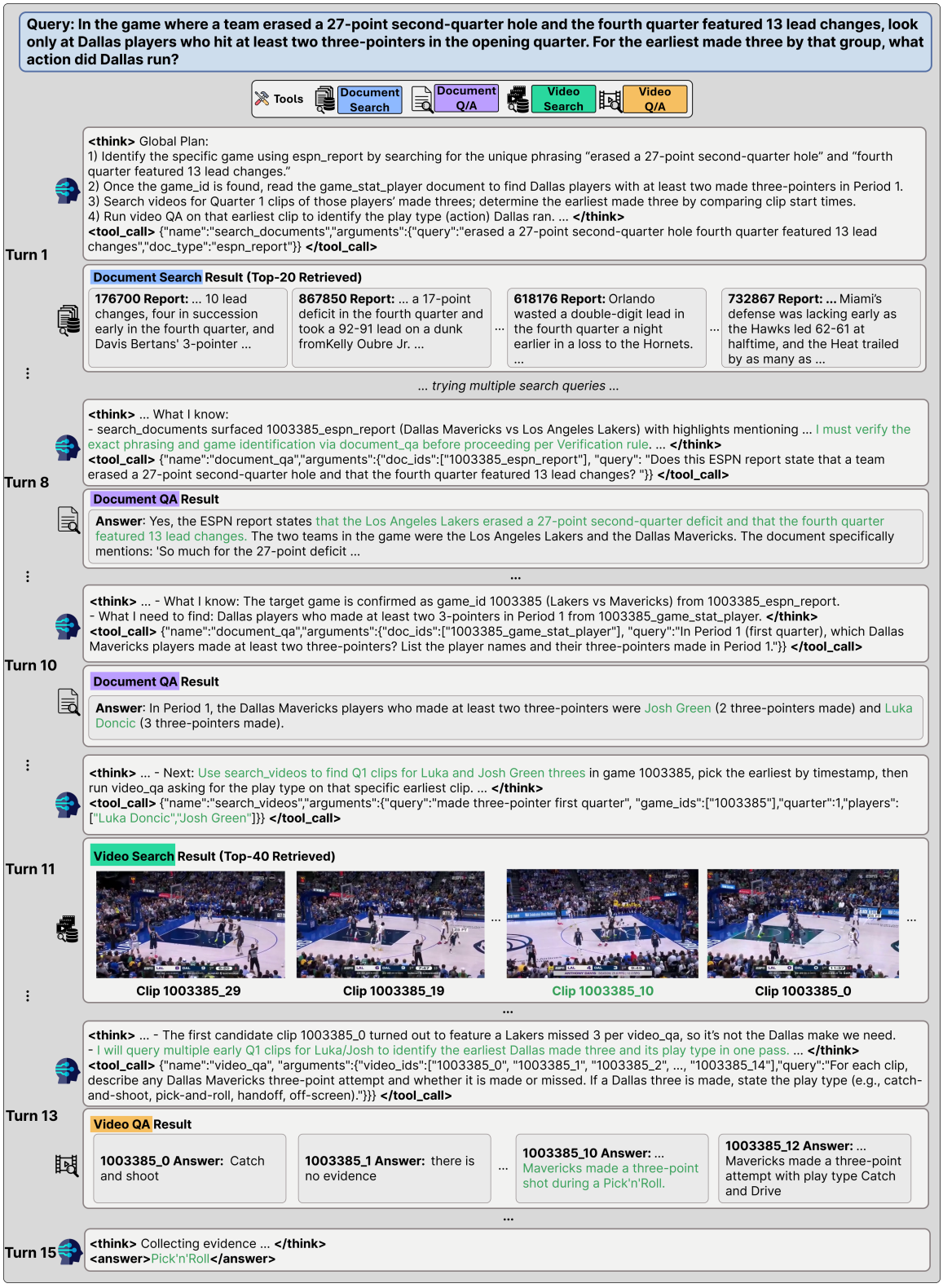


Figure 50. T9 Cross-Corpus Agentic Reasoning: expanded view of the successful trajectory summarized in Figure 47, showing the full reasoning trace and tool-call details at each turn. The agent identifies the target game (the 27-point comeback) via document search and verification (Turns 1–7 and 9), extracts the relevant players' statistics via document QA (Turns 8 and 10), retrieves their three-pointer clips via video search (Turn 11), and determines the play type via video QA (Turns 12–14), arriving at the correct answer in 15 turns. Green highlights mark key reasoning steps and extracted evidence.

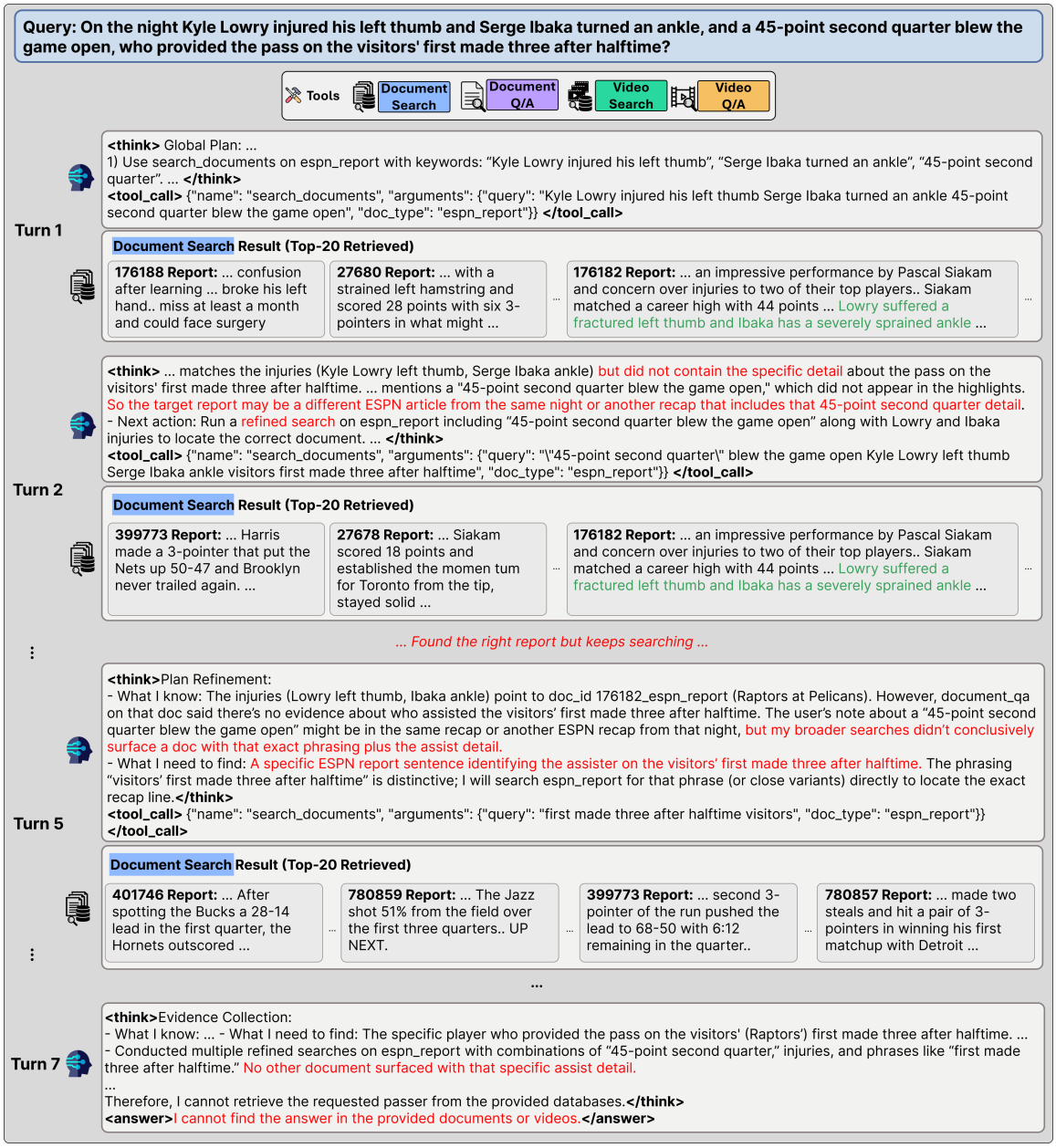


Figure 51. T9 Cross-Corpus Agentic Reasoning: same task setup as Figure 50. This figure shows a common failure mode—modality-switch failure. The agent correctly identifies the target game by matching injury details to a game report (Turn 1), but the requested information (who assisted the visitors' first three-pointer after halftime) is only available in the video clips. Rather than switching to video search, the agent repeatedly refines its document queries (Turns 2–5) and eventually concludes the answer is unavailable (Turn 7), despite the relevant clip being in the corpus. Red highlights indicate unsuccessful reasoning steps.

Figure 52. T9 Cross-Corpus Agentic Reasoning: same task setup as Figure 50. This figure shows a complementary failure mode—video perception bottleneck. Unlike the modality-switch failure in Figure 51, the agent's retrieval and tool selection are correct throughout: it identifies the target game (Turn 1), locates the relevant clips via video search with temporal boundaries (Turns 11, 15), and narrows down to the correct player. However, the video QA model repeatedly misidentifies the dribble move (Turns 16–18), causing the agent to lose confidence and search for alternative evidence. After exhausting all 30 allowed turns, it returns no answer— illustrating that even with correct planning and retrieval, limited visual perception remains a critical bottleneck. Red highlights indicate erroneous reasoning steps and incorrect tool responses.

F Datasheet

Following [22], we provide a datasheet for SVI-Bench. Full details on dataset construction and quality control are provided in §A.1.

Motivation. SVI-Bench was created to evaluate the full spectrum of video intelligence capabilities—from perception through causal reasoning to simulation and agentic synthesis—in a controlled, verifiable domain. Existing video benchmarks focus primarily on recognition and retrieval. SVI-Bench extends evaluation to higher-order reasoning, generation, and multi-step agentic tasks.

Composition. The dataset spans three sports (basketball, hockey, soccer) across nine tasks organized into four cognitive pillars. The corpus comprises approximately 64K games of full-length video, play-by-play logs, structured metadata, game reports, and expert commentary transcripts. Benchmark instances include video–caption pairs (T1), video–question pairs (T2, T4, T5), text-to-video retrieval queries (T3), video–trajectory generation pairs (T7, T8), long-form report generation prompts (T6), and multimodal agentic reasoning questions (T9). All data pertains to professional and collegiate sports competitions. No private or off-field data about individuals is collected.

Collection and preprocessing. Game videos were sourced from game-replay archives, play-by-play logs and statistics from licensed league data providers, and game reports from sports journalism outlets. Expert commentary was extracted from broadcast audio using Whisper ASR [62]. All resources were temporally aligned and processed through the data engine described in §A.2.

Intended uses and limitations. The intended use is benchmarking video-language models and AI agents on strategic video intelligence tasks. The benchmark should not be used for real-world decision-making (e.g., game strategy, player evaluation, or betting) without expert oversight.

Maintenance. The benchmark will be maintained by the authoring research group. We will issue corrections for any errors identified post-release, publish versioned updates with changelogs, and accept community-reported issues through the benchmark's public repository. Future versions may expand to additional sports, leagues, and seasons.

G Broader Impact Statement

Intended uses and benefits. SVI-Bench is designed as a research benchmark to advance video understanding capabilities. By structuring evaluation around a progres-

sion from perception to reasoning to generation to agentic synthesis, it provides a diagnostic tool for identifying specific capability gaps in current models. While instantiated in the sports domain, the core challenges the benchmark targets—fine-grained multi-entity understanding, causal reasoning, long-horizon planning, and calibrated prediction—generalize to domains such as traffic, surgery, or robotics.

Potential risks. We acknowledge two potential risks associated with the benchmark:

- **Dual use in gambling**: outcome forecasting models (T5) could theoretically be applied to sports betting. However, current performance levels (below 43% accuracy with poor calibration) demonstrate that even the strongest models cannot reliably predict game outcomes, and the benchmark's purpose is to measure and expose these limitations.
- **Player privacy**: all data involves professional and collegiate athletes performing in public competitions. No private or off-field data is collected or used.

Bias considerations. The benchmark covers three sports (basketball, hockey, soccer) with data drawn from leagues across multiple continents—including North America, Europe, Asia, and Australia. This provides meaningful geographic diversity, though the distribution is not uniform: European and North American leagues are more heavily represented. Women's leagues are included in the basketball data (3 of 17 leagues, approximately 6K games) but remain underrepresented relative to men's leagues.

Environmental impact. Training and evaluating large video models requires substantial computational resources. We report hardware configurations and training durations for all experiments throughout this supplementary material to enable assessment of computational costs.