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], and Gedas Bertasius[1]

[1]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]Northeastern University, Boston, MA, USA

⋆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 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 higher levels of the stack. 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. We release the full benchmark to catalyze progress toward AI systems capable of strategic intelligence in complex, dynamic multi-agent environments.

Code	`github.com/texaser/svi-bench`
Data	`huggingface.co/mvp-group/svi-bench`
Website	`svi-bench.github.io`
Extended Paper	`svi-bench.github.io/svi_bench_extended.pdf`

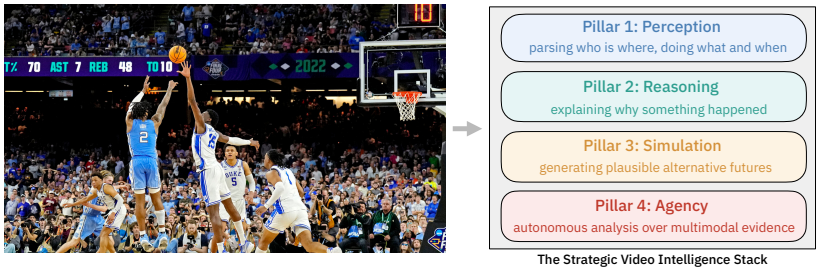

Fig. 1: Overview of SVI-Bench, illustrated through a single play from the 2022 NCAA Final Four. SVI-Bench is the first large-scale video benchmark evaluating the full SVI stack: Perception (describing what happens), Reasoning (explaining why), Simulation (generating plausible alternatives), and Agency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

1 Introduction

In the final seconds of Game 6 of the 1998 NBA Finals, Michael Jordan receives the ball trailing by one. Jordan perceives the defensive formation shifting 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—moving 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, predict *what would have happened* had the guard driven left instead of right, or recommend the *optimal response* given the defensive configuration. 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 spanning *perception* (parsing who is where and doing what), *causal reasoning* (explaining why actions lead to outcomes), *simulation* (generating futures and goal-directed strategies), and *agentic synthesis* (autonomously integrating multimodal evidence) (Figure 1).

Progress on SVI has been limited by the absence of suitable benchmarks. Existing video benchmarks [20, 44, 78, 94] cover perception and temporal reasoning, but none evaluates the full stack from perception to agency (Table 1). Synthetic datasets [3, 88] provide verifiable ground truth for causal questions but without real-world complexity. Real-world benchmarks [20, 44], by contrast, offer visual richness but no verifiable ground truth for causal or strategic reasoning.

Team sports offer a *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 reason-

ing 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 outcome labels for causal and strategic questions.

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 perception but decline sharply on reasoning and agentic tasks (Figure 2). Our contributions are:

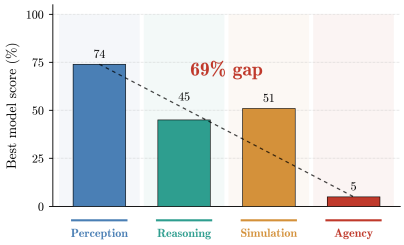

Fig. 2: The performance cliff. Best task score per pillar, normalized to a 0–100% range. From the strongest Perception result (74%) to Agentic Synthesis (5%), performance drops by 69 points, with Reasoning and Simulation falling in between.

1. **The Strategic Video Intelligence framework**, formalizing video understanding as a progressive perception-to-agency stack.
2. **A data engine** that aligns five modalities via temporal alignment, cross-modal entity resolution, LLM-assisted generation, and 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.
4. **Reference methods and analysis** for every task, localizing where and why performance degrades.

Extended version. This is the compact version of the paper. An extended version provides full dataset construction details, per-task statistics, task-specific prompts, additional analysis, and complete results for every task.

2 Related Work

Video Benchmarks. Early benchmarks targeted action recognition [10, 32, 66] and temporal localization [9, 30]. Video QA benchmarks [44, 81, 82, 90] added

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.

					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**	✓	✓	✓	✓

language-grounded reasoning over short clips with predominantly perceptual questions. Long-video benchmarks [20, 78, 94] extend temporal scope but lack verifiable causal ground truth. Egocentric planning [14] and temporal reasoning [35, 42] benchmarks target single-agent or low-complexity settings.

Causal and Counterfactual Reasoning in Video. Synthetic environments [3–5, 88] provide verifiable ground truth for causal and physical reasoning but involve single objects in simplified worlds without multi-agent behavioral complexity. Real-video causal QA [13, 36] attempts causal reasoning but relies on subjective annotations without objectively verifiable outcomes.

Sports Video Analysis. Prior work targets action detection [15,18,22], tracking [17], trajectory prediction [1,19], and skill analysis [53,84]. SPORTU [80] and SportR [79] add rule comprehension but omit simulation or agentic reasoning. NBA tracking datasets [11] lack video or language.

Multimodal LLMs and Agentic AI. Recent multimodal LLMs [21, 40, 41, 51, 91] demonstrate strong video description but exhibit brittle temporal reasoning [20, 44]. Agentic AI systems [62, 72, 87] combine LLMs with tool use, and recent work extends this to video [52, 58, 74, 86, 92]. SVI-Bench's Pillar 4 introduces the first agentic video evaluation at corpus scale.

Strategic Game AI and World Models. Superhuman game AI [63, 64, 69] operates in fully observable, discrete-state environments. SVI-Bench targets continuous, partially observable, real multi-agent video. World models span model-based RL [25–27], diffusion [2], transformer [47,59], and foundation [8,29, 50] approaches, but none target strategic reasoning in 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 for evaluating the full SVI stack. It is

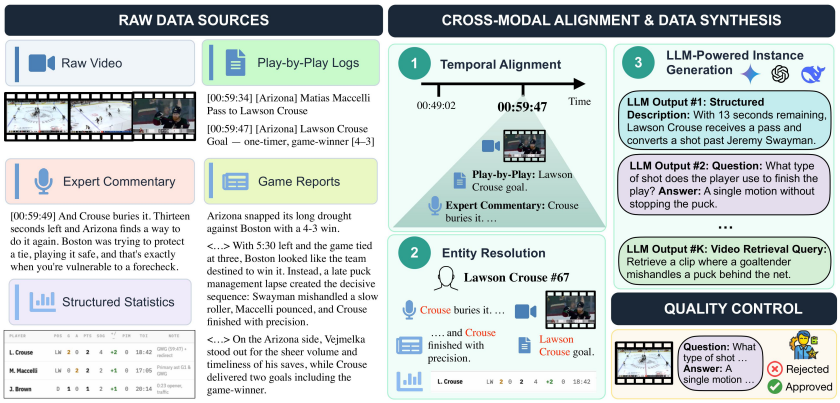

Fig. 3: The SVI-Bench data engine. Five raw sources are transformed into a cross-referenced corpus via (1) temporal alignment, (2) cross-modal entity resolution, (3) LLM-based instance generation, and (4) automatic and human quality control.

built on two principles: (i) primary evidence is human- or league-derived (play-by-play logs, official statistics, journalist reports, broadcast commentary), and (ii) LLMs *scale* task-instance generation from these grounded sources, with manual verification on a representative subset of every task. This yields supervision at a scale and density unmatched by prior sports video resources.

3.1 Data Sources and Scale

SVI-Bench spans three 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 comprises five synergistic modalities (Figure 3, left), all temporally aligned and cross-referenced through shared game and player identifiers: **broadcast video** (∼35K hours spanning multiple seasons); **play-by-play logs** (∼15M timestamped event records from official league data feeds, including shots, passes, fouls, and substitutions with player identities and spatial coordinates); **expert commentary** (∼15K hours of broadcast commentary and analyst narration collected via ASR); **game reports** (∼23K post-game journalist recaps and editorial analyses); and **box-score statistics** (∼103K records of player/team performance metrics and standings).

3.2 Data Construction Pipeline

Our data engine (Figure 3) transforms these five raw sources into a cross-referenced 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 linking 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

Table 2: Summary of SVI-Bench benchmark tasks. Nine tasks across four cognitive pillars 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

linked across modalities and organized into identity graphs capturing relationships (teammate, opponent) and attributes (position, statistics, role). **(3) LLM-powered instance generation:** Using the assembled multimodal context, LLMs synthesize instances guided by pillar-aware prompt templates, generating 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 automatic checks against event logs, followed by human expert review by domain-knowledgeable annotators across a balanced subset spanning all sports, pillars, and difficulty levels.

4 The SVI-Bench Evaluation Suite

SVI-Bench comprises 9 tasks organized into a four-pillar hierarchy (Table 2), from perception through causal reasoning and simulation to agency. Construction details, prompts, per-task statistics, and complete results are in the extended version. 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).

T1: Structured Play Description

Task formulation. Given a 10-second video, the model must generate a dense, structured caption that describes the actions, player identities, spatial positioning, and game context. Unlike standard captioning benchmarks [33, 75, 85, 95] describing a single salient action per clip, T1 requires describing dynamic scenes with 10+ coordinated agents, parallel sub-actions, and game-state context.

Data and construction. We extract 10-second segments centered on specific events, with captions synthesized from play-by-play annotations and refined via GPT-4o-mini for linguistic diversity and narrative coherence.

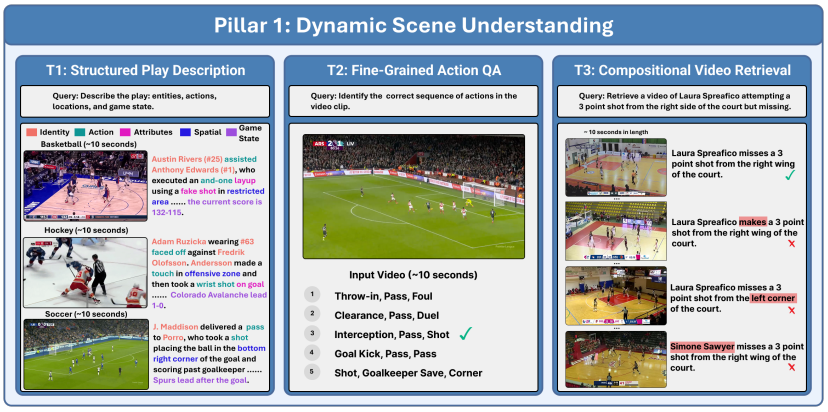

Fig. 4: Overview of Pillar 1: Dynamic Scene Understanding. This pillar evaluates foundational perceptual capabilities through three tasks: structured play description (T1), fine-grained action QA (T2), and compositional video retrieval (T3).

Evaluation. We use an LLM-as-a-judge protocol with scoring rubrics, assigning Likert scores from 0–5 along six axes: action, identity, causality/outcome, spatial understanding, temporal understanding, and contextual details. We report the average score across all axes as the primary metric. To verify judge reliability, three annotators independently scored 60 randomly sampled instances. The mean absolute difference between human and LLM-judge scores is 0.40 on the 0–5 scale, indicating agreement within normal inter-annotator variation.

Baselines and findings. GPT-5.2 achieves only 1.61/5.00 average score, and Gemini 3 Flash reaches 1.67. Fine-tuned LLaVA-Video-7B reaches 2.17 (+1.28 over zero-shot), demonstrating the value of domain-specific training. Per-axis analysis reveals that models perform relatively well on spatial understanding (2.74) and temporal understanding (2.72) but struggle with identity recognition (1.11) and causality/outcome reasoning (1.82).

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 [44,81,82,90] 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 six categories (action recognition, temporal ordering, play analysis, spatial relationships, player identification, and OCR).

Data and construction. We segment full-game footage into 10-second clips centered on individual plays with dense play-by-play annotations, covering 31 question types across three sports organized into six categories.

Evaluation. We report average accuracy across all sports and question types.

Baselines and findings. Gemini 3 Flash achieves 58.75% accuracy while GPT-5.2 achieves 52.91%. Fine-tuned LLaVA-Video-7B reaches 73.91% (+36.9% over

zero-shot). Player identification questions are most challenging, while action recognition is easier. Sport-experienced humans reach 75.78% overall (Section 5).

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. Unlike standard video retrieval [28, 33, 60, 85] where visually diverse candidates allow coarse features to suffice, all T3 candidates depict the same sport and many share nearly identical visual elements, differing only in their specific composition of attributes.

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.

Evaluation. We report Recall@K with R@1 as the primary metric.

Baselines and findings. We fine-tune InternVideo2 [76] using a video-text contrastive loss. The model achieves an aggregate R@1 of 3.0% and R@10 of 13.3%, highlighting the difficulty of fine-grained retrieval at scale. When the proportion of near-duplicate distractors increases, R@100 drops by nearly half, confirming that distinguishing visually similar compositions remains a core challenge.

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.

4.2 Pillar 2: Causal Reasoning (T4–T6)

This pillar evaluates causal reasoning over long-form video spanning minutes to hours, testing whether models can explain the causes behind game events (T4), predict how a situation will unfold (T5), and synthesize extended developments into coherent narratives (T6) (Figure 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 dynamics such as momentum shifts. Compared to long-form video QA benchmarks [20, 67, 94] whose questions are 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 expert commentaries and game reports across all three sports, totaling 825 unique games. A multi-

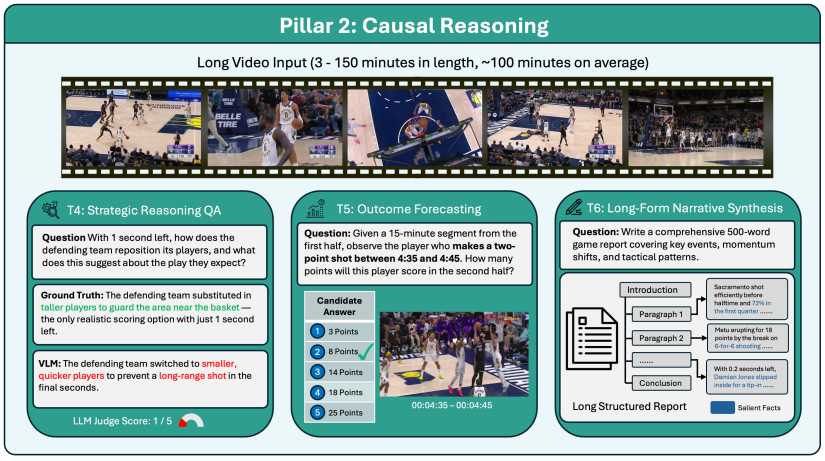

Fig. 5: Overview of Pillar 2: Causal Reasoning. This pillar evaluates the ability to reason about game-level context through three tasks: strategic reasoning QA (T4), outcome forecasting (T5), and long-form narrative synthesis (T6).

stage pipeline generates open-ended question–answer pairs followed by bias-mitigation filtering and human validation of factual validity and question quality.

Evaluation. We use an LLM-as-a-judge protocol that scores each response on a 0–5 scale, assessing factual consistency with the reference answer and reasoning coherence. To validate judge reliability, a human annotator scored all model responses for 25 evaluation instances. The mean score difference between human and LLM judge is 0.12 on the 0–5 scale, confirming strong alignment.

Baselines and findings. We evaluate proprietary models (GPT-5.2, Gemini 3.1 Pro) and open-source models (Qwen3-VL-32B, Molmo 2-8B). All models score near 2/5 on average. Gemini 3.1 Pro is strongest overall (2.17), driven primarily by soccer (2.49); GPT-5.2 (2.06) leads on basketball (1.99). Qwen3-VL-32B and Molmo 2-8B reach 2.01 and 1.82 respectively. These low scores reflect the difficulty of T4, which requires reasoning about causal relationships and strategic intent, not just recognizing individual events.

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 [24,39,45,61] that predicts short-horizon spatial positions, T5 targets semantically rich outcomes (who will score, which strategy will be used, how a game state will evolve) requiring understanding of complex causal mechanisms rather than trend extrapolation. Questions span *performance forecasting* (predicting player or team statistical accomplishments), *game state evolution* (anticipating scores and possessions), and *strategic intention* (identifying the most probable tactical shifts).

Data and construction. We curate 114K multiple-choice questions spanning 15 question types across three sports, using game videos and dense play-by-play event annotations with video segments ranging from 3 to 15 minutes.

Evaluation. We use top-1 accuracy and additionally report calibration error to measure alignment between predicted confidence and empirical correctness.

Baselines and findings. We benchmark open-source (Qwen3-VL-8B [57], Molmo 2-8B [16], BIMBA [31]) and proprietary models (GPT-5.2, Gemini 3.0 Pro). No model exceeds 43.18% under zero-shot evaluation. Fine-tuning Qwen3-VL-8B improves accuracy to 44.82% (+7.9% over zero-shot). Even frontier models like GPT-5.2 are poorly calibrated, with confidence exceeding accuracy by 28 points.

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 [12, 54, 93] 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.

Data and construction. We define 10 report templates per sport (e.g., game narrative, player impact, team strategy evolution), five targeting single-game analysis and five requiring synthesis across multiple games. Reference reports are generated by an LLM from play-by-play logs, box scores, and journalist reports, ensuring consistent structure and verifiable factual content.

Evaluation. We evaluate along three metrics using an LLM-as-a-judge (Qwen3-235B Thinking [56]): *factual accuracy* via atomic fact decomposition [48]; *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 quality*, rated on a 1–5 scale for coherence, topic adherence, and length compliance.

Baselines and findings. We benchmark Qwen3-VL-8B, GPT-5, and Gemini 3.1 Pro. Models achieve relatively high factual accuracy at 73.01%, but struggle with saliency (7.33%). Writing quality is the least problematic dimension, with most models scoring above 4.5/5.

Pillar 2 Takeaway (Causal Reasoning). Models degrade sharply on reasoning tasks. T4 strategic reasoning scores 2/5 and T5 forecasting stays below 45%. On T6 models reach 73.0% factual accuracy but only 7.33% saliency, describing what happened while failing to cover salient events. Across all three tasks, causal and strategic reasoning emerges as the main bottleneck.

4.3 Pillar 3: Strategic Simulation (T7–T8)

Understanding why something happened is distinct from reasoning about what *could* happen. This pillar evaluates whether models can simulate alternative futures through video generation. Given a short clip (5–10s), both tasks require

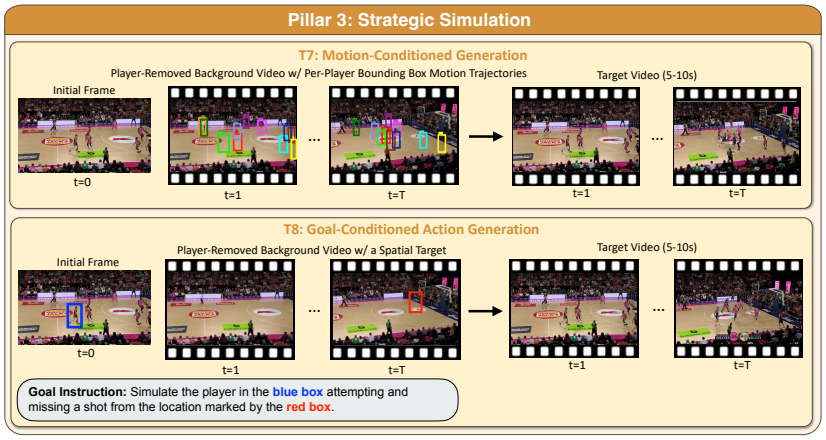

Fig. 6: Overview of Pillar 3: Strategic Simulation. This pillar tests the ability to simulate alternative futures through two video generation tasks: motion-conditioned generation (T7), where players follow prescribed trajectories, and goal-conditioned action generation (T8), where the model plans actions toward a specified goal.

generating a realistic video of how the scene evolves if players follow specified trajectories (T7) or execute an action toward a goal (T8) (Figure 6).

T7: Motion-Conditioned Generation

Task formulation. Given an initial frame showing all players in their starting positions, a *player-removed background video* (the original footage with all players digitally erased via video inpainting, leaving only the court or pitch and static elements), and a set of 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. Prior trajectory-conditioned generation [43,49,73,89] typically involves one or two objects in simple scenes; T7 targets multi-agent coordination where 10+ players move simultaneously, interact physically, and occlude one another.

Data and construction. Each instance consists of: (1) an initial frame, (2) per-player motion trajectory as bounding-box sequences, and (3) a player-removed background video generated via video inpainting [34]. We apply explicit quality filtering to remove instances with unstable tracking, severe occlusion, or visible inpainting artifacts (e.g., residual player silhouettes, texture bleeding), ensuring that generation models operate on clean background inputs.

Evaluation. We evaluate with two metrics: *Video mIoU* [6], measuring spatiotemporal alignment between player trajectories in generated and reference videos; and *temporal feature similarity*, comparing SigLIP [68] features from corresponding player regions across frames to assess visual consistency.

Baselines and findings. Our reference method fine-tunes Wan 2.1 [70] on SVI-Bench data, extending it to accept structured input conditions (initial frame, bounding boxes, trajectories, player-removed background). We additionally eval-

uate ATI [71] and MagicMotion [38] off-the-shelf. On basketball, our model achieves Video mIoU of 0.513 vs. 0.466 (MagicMotion) and 0.397 (ATI), with larger gains on soccer (0.611 vs. 0.544 and 0.402). Temporal feature similarity follows the same trend (basketball: 0.787 vs. 0.725/0.617; soccer: 0.804 vs. 0.708/0.507). Yet even our best model's 0.513 mIoU means roughly half of generated player positions deviate significantly from prescribed trajectories, indicating that reliable multi-agent motion control remains far from solved.

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. Unlike T7, which prescribes exact trajectories, T8 requires the model to *plan* intermediate actions to achieve a high-level goal under explicit spatial constraints, requiring implicit understanding of environment dynamics and goal-directed reasoning, going beyond open-ended text-conditioned generation [7, 23, 29, 65].

Data and construction. We pair curated basketball video clips with structured goal specifications derived from annotated actions, covering diverse goal-conditioned behaviors including completing plays at designated locations, executing specific moves, and interaction-aware scenarios.

Evaluation. We evaluate with three complementary metrics: *mIoU* on the final frame, measuring bounding-box overlap between generated and target player positions; *feature similarity* on the final frame, assessing visual fidelity of the realized outcome; and *goal accuracy* via a fine-tuned video-language QA model evaluating whether the generated video achieves the specified objective.

Baselines and findings. We adapt the Wan-based framework from T7 to the goal-conditioned setting, replacing trajectory inputs with textual goal specifications and spatial endpoint constraints. The model achieves final-frame mIoU of 0.344 vs. 0.129 (MagicMotion) and 0.047 (ATI), feature similarity of 0.468 vs. 0.169 (MagicMotion) and 0.067 (ATI), and goal accuracy of 50.2% vs. 31.4% (MagicMotion) and 40.5% (ATI). These results highlight a fundamental gap between trajectory-following (T7) and goal-directed video generation (T8).

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 rather than follow trajectories.

4.4 Pillar 4: Agentic Synthesis (T9)

T9: Cross-Corpus Agentic Reasoning

Task formulation. Given a complex natural-language query, the model must plan a multi-step retrieval strategy, gather evidence from heterogeneous sources

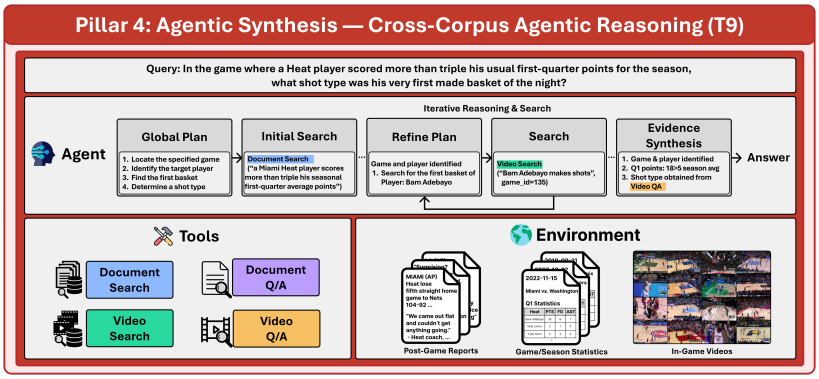

Fig. 7: Overview of Pillar 4: Agentic Synthesis. This pillar evaluates the ability to autonomously gather and integrate multimodal evidence through a single task: cross-corpus agentic reasoning (T9), where the agent plans and executes tool-assisted search across large-scale heterogeneous sources to answer complex strategic queries.

(video clips, game reports, statistical records), and reason over it to produce a final answer (Figure 7). The agent is equipped with search and QA tools over a document database (post-game reports, game- and season-level statistics) and a video database (footage segmented into 10–15s clips). While tool-augmented reasoning has been explored in text [37,46,55,96] and single video settings [52,86,92], T9 extends it to multimodal evidence at corpus scale, requiring the agent to integrate evidence across modalities through complex reasoning patterns (looping, backtracking, conditional branching, numerical aggregation) over ∼1.8M clips and ∼33K documents across three sports. T9 adopts the hard-to-find but easy-to-verify principle from prior agentic search work [77]. Each question begins with seed facts (a specific play, score, or event attribute) and adds multi-hop narrative constraints that uniquely identify the relevant 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 across 7,430 games impractical and forces the agent to search, 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 ∼1.8M video clips (∼5,670 broadcast hours). Questions are constructed to require evidence from multiple sources rather than any single modality alone. The final evaluation set contains 1,000 questions, balanced across sports.

Evaluation. We report accuracy, using an LLM judge (GPT-5.2) to compare the agent's response to the ground-truth answer.

Baselines and findings. Even the strongest model (GPT-5.2) achieves only 4.6% accuracy across the three sports, with performance dropping further for smaller open-source models (Qwen3-Omni-30B [83]: 2.1%). Tool-use patterns reveal the same gap: GPT-5.2 issues roughly 21 tool calls per question, while smaller Qwen models make only 3–8 and often terminate early with insufficient evidence. To-

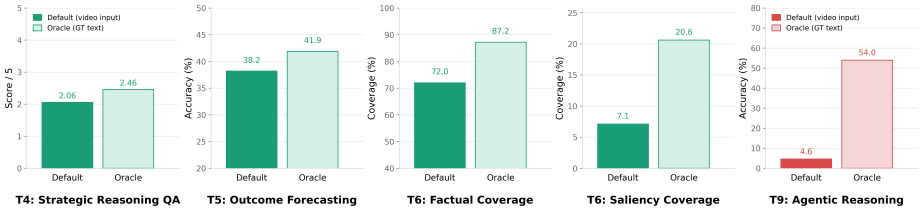

Fig. 8: Oracle performance on reasoning and agentic tasks (T4, T5, T6, T9). The oracle variant replaces video with ground-truth textual descriptions of game events. All tasks use GPT-5.2, except for T6 which uses GPT-5. Gains are small on T4 and T5, moderate on T6, and largest on T9, indicating that strategic reasoning, forecasting, saliency judgment, and multi-step planning remain distinct bottlenecks.

gether, these results show that current agents cannot reliably plan retrieval, gather evidence, and reason over a corpus at this scale.

Pillar 4 Takeaway (Agentic Synthesis). Agentic synthesis is the hardest capability in SVI-Bench, with even the strongest model reaching only 4.6%.

5 Cross-Task Analysis

This section analyzes per-task results to characterize the overall trends in performance from perception through agentic synthesis.

5.1 The Performance Cliff

Figure 2 plots the best task score within each pillar, each normalized to a 0–100% range. The dashed line marks the drop from the strongest perception result to agentic synthesis, with reasoning and simulation 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 perceive dynamic multi-agent environments far more accurately than they can reason about, simulate, or plan within them.

5.2 Oracle Experiments

To test whether perception is the primary bottleneck in reasoning and agentic tasks, we replace video input with ground-truth textual descriptions of game events derived from play-by-play logs, evaluating the same model in default mode (video) and oracle mode (text) (Figure 8). The effect varies sharply across tasks. Oracle access raises T9 accuracy from 4.6% to 54.0% and T6 factual accuracy from 71.99% to 87.19%, with smaller improvements on T4 (2.06 to 2.46 on the 0–5 score), T5 (38.2% to 41.9%), and T6 saliency (7.1% to 20.6%). These results suggest that no single capability accounts for the gap: strategic reasoning (T4), forecasting (T5), saliency judgment (T6), and multi-step planning (T9) each limit performance independently.

5.3 Human Studies

To establish that these tasks are solvable and to measure the human-model gap, we run human studies on T2 (perception), T4 (strategic reasoning), and T5

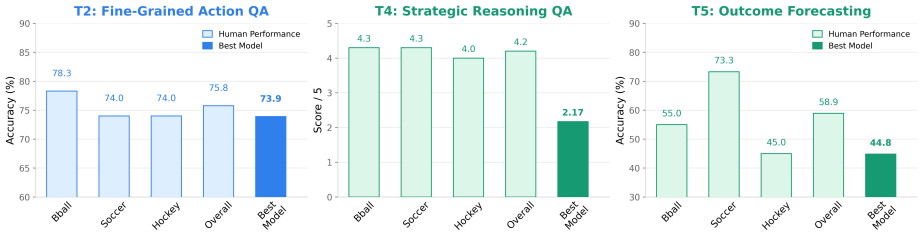

Fig. 9: Human–model comparison on T2, T4, and T5. Bars show per-sport and overall human performance alongside the best model. T2 and T5 use accuracy (%). T4 uses the open-ended 0–5 score. Models nearly match humans on perception but trail them substantially on strategic reasoning and forecasting.

(forecasting), with participants who have 5–10 or more years of experience in their sport (Figure 9). 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%). Humans also know when they are uncertain: their accuracy rises from 30% to 90% across confidence levels on T2 and from 50% to 100% on T5, while GPT-5.2 reports similar confidence on correct and incorrect answers, producing a 28-point gap between average confidence and accuracy. Human performance establishes that these tasks are solvable and that the gap reflects current model limitations rather than task difficulty.

6 Conclusion

We introduced SVI-Bench, the first large-scale benchmark for strategic video intelligence in real-world multi-agent video. Across 9 tasks and four pillars, models perform competently on perception but degrade substantially on higher-level reasoning tasks. Even given perfect visual information on our agentic task, the strongest models reach only 54%, showing that the challenge extends beyond perception to reasoning, planning, and evidence integration.

Scope and limitations. SVI-Bench is a team-sports microworld—a controlled proxy for real multi-agent video, not a claim of cross-domain generalization. Sports-specific properties such as broadcast conventions and fixed rules may not transfer beyond sports. Several tasks rely on LLM judges, which we validate via human-agreement checks, though some bias may persist.

Future directions. The gaps revealed by SVI-Bench point to three directions: video models that ground long-form reasoning in visual evidence (T4–T6), generative models with explicit multi-agent dynamics for goal-directed action (T7–T8), and multimodal agents that plan and reason across corpora at scale (T9).

Acknowledgements

This work was supported by Laboratory for Analytic Sciences via NC State University, ONR Award N00014-23-1-2356, Sony Focused Research award, NSF CAREER Award 2541848, Northeastern University startup funds, and the President Joseph E. Aoun Chair.

References

1. Alcorn, M.A., Nguyen, A.: baller2vec: A multi-entity transformer for multi-agent spatiotemporal modeling. arXiv preprint arXiv:2102.03291 (2021)
2. Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in atari. Advances in Neural Information Processing Systems **37**, 58757–58791 (2024)
3. Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., Girshick, R.: Phyre: A new benchmark for physical reasoning. Advances in Neural Information Processing Systems **32** (2019)
4. Baradel, F., Neverova, N., Mille, J., Mori, G., Wolf, C.: Cophy: Counterfactual learning of physical dynamics. arXiv preprint arXiv:1909.12000 (2019)
5. Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y.F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.Y., et al.: Physion: Evaluating physical prediction from vision in humans and machines. arXiv preprint arXiv:2106.08261 (2021)
6. Bertasius, G., Torresani, L.: Classifying, segmenting, and tracking object instances in video with mask propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9739–9748 (2020)
7. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
8. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024)
9. Caba Heilbron, F., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
10. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019)
11. Cervone, D., D'Amour, A., Bornn, L., Goldsberry, K.: A multiresolution stochastic process model for predicting basketball possession outcomes. Journal of the American Statistical Association **111**(514), 585–599 (2016)
12. Chen, M., Chu, Z., Wiseman, S., Gimpel, K.: Summscreen: A dataset for abstractive screenplay summarization. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8602–8615 (2022)
13. Chen, T., Liu, H., He, T., Chen, Y., Gan, C., Ma, X., Zhong, C., Zhang, Y., Wang, Y., Lin, H., et al.: Mecd: Unlocking multi-event causal discovery in video reasoning. Advances in neural information processing systems **37**, 92554–92580 (2024)
14. Chen, Y., Ge, Y., Ge, Y., Ding, M., Li, B., Wang, R., Xu, R., Shan, Y., Liu, X.: Egoplan-bench: Benchmarking multimodal large language models for human-level planning. International Journal of Computer Vision **134**(3), 118 (2026)
15. Cioppa, A., Giancola, S., Deliege, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Van Droogenbroeck, M.: Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3491–3502 (2022)
16. Clark, C., Zhang, J., Ma, Z., Park, J.S., Tripathi, R., Lee, S., Salehi, M., Ren, J., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28652–28668 (2026)

17. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9921–9931 (2023)
18. Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Van Droogenbroeck, M.: Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4508–4519 (2021)
19. Felsen, P., Lucey, P., Ganguly, S.: Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders. In: Proceedings of the European conference on computer vision (ECCV). pp. 732–747 (2018)
20. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
21. Gemini Team: Gemini: A family of highly capable multimodal models (2025), `https://arxiv.org/abs/2312.11805`, accessed 30 June 2026
22. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: A scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1711–1721 (2018)
23. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
24. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2255–2264 (2018)
25. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)
26. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering atari with discrete world models. arXiv preprint arXiv:2010.02193 (2020)
27. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104 (2023)
28. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5804–5813 (2017)
29. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
30. Idrees, H., Zamir, A.R., Jiang, Y.G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M.: The thumos challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding **155**, 1–23 (2017)
31. Islam, M.M., Nagarajan, T., Wang, H., Bertasius, G., Torresani, L.: Bimba: Selective-scan compression for long-range video question answering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29096–29107 (2025)
32. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)

33. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
34. Lee, Y.C., Lu, E., Rumbley, S., Geyer, M., Huang, J.B., Dekel, T., Cole, F.: Generative omnimatte: Learning to decompose video into layers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12522–12532 (2025)
35. Li, B., Zhao, K., Zhang, C., Mitra, C., Nyandwi, J.d.D., Bertasius, G.: Timeblind: A spatio-temporal compositionality benchmark for video llms. arXiv preprint arXiv:2602.00288 (2026)
36. Li, J., Niu, L., Zhang, L.: 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. pp. 21273–21282 (2022)
37. Li, M., Zhao, Y., Yu, B., Song, F., Li, H., Yu, H., Li, Z., Huang, F., Li, Y.: Api-bank: A comprehensive benchmark for tool-augmented llms. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 3102–3116 (2023)
38. Li, Q., Xing, Z., Wang, R., Zhang, H., Dai, Q., Wu, Z.: Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12112–12123 (2025)
39. Liang, J., Jiang, L., Murphy, K., Yu, T., Hauptmann, A.: The garden of forking paths: Towards multi-future trajectory prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10508–10518 (2020)
40. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
41. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
42. Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., Hou, L.: Tempcompass: Do video llms really understand videos? In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 8731–8772 (2024)
43. Ma, W.D.K., Lewis, J.P., Kleijn, W.B.: Trailblazer: Trajectory control for diffusion-based video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
44. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
45. Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., Gaidon, A.: It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: European conference on computer vision. pp. 759–776. Springer (2020)
46. Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., Scialom, T.: Gaia: a benchmark for general ai assistants. In: International Conference on Learning Representations. vol. 2024, pp. 9025–9049 (2024)
47. Micheli, V., Alonso, E., Fleuret, F.: Transformers are sample-efficient world models. In: International Conference on Learning Representations (2023)
48. Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.t., Koh, P., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.: Factscore: Fine-grained atomic evaluation of factual precision in

long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12076–12100 (2023)
49. Namekata, K., Bahmani, S., Wu, Z., Kant, Y., Gilitschenski, I., Lindell, D.: Sg-i2v: Self-guided trajectory control in image-to-video generation. In: International Conference on Learning Representations. vol. 2025, pp. 38483–38505 (2025)
50. NVIDIA: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
51. OpenAI: GPT-4o System Card (2024), https://arxiv.org/abs/2410.21276, accessed 30 June 2026
52. Pan, J., Zhang, Q., Zhang, R., Lu, M., Wan, X., Zhang, Y., Liu, C., She, Q.: Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. arXiv preprint arXiv:2511.05489 (2025)
53. Pan, Y., Zhang, C., Bertasius, G.: Basket: A large-scale video dataset for fine-grained skill estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28952–28962 (2025)
54. Papalampidi, P., Keller, F., Frermann, L., Lapata, M.: Screenplay summarization using latent narrative structure. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1920–1933 (2020)
55. Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al.: Toolllm: Facilitating large language models to master 16000+ real-world apis. In: International Conference on Learning Representations. vol. 2024, pp. 9695–9717 (2024)
56. Qwen Team: Qwen3 Technical Report. arXiv preprint arXiv:2505.09388 (2025), https://arxiv.org/abs/2505.09388, accessed 30 June 2026
57. Qwen Team: Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025), https://arxiv.org/abs/2511.21631, accessed 30 June 2026
58. Rasheed, H., Zumri, M., Maaz, M., Yang, M.H., Khan, F.S., Khan, S.: Videocom: Interactive video reasoning via chain of manipulations. arXiv preprint arXiv:2511.23477 (2025)
59. Robine, J., Höftmann, M., Uelwer, T., Harmeling, S.: Transformer-based world models are happy with 100k interactions. arXiv preprint arXiv:2303.07109 (2023)
60. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3202–3212 (2015)
61. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: European conference on computer vision. pp. 683–700. Springer (2020)
62. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. Advances in neural information processing systems **36**, 68539–68551 (2023)
63. Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al.: Mastering atari, go, chess and shogi by planning with a learned model. Nature **588**(7839), 604–609 (2020)
64. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.: Mastering the game of go without human knowledge. nature **550**(7676), 354–359 (2017)
65. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)

66. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
67. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016)
68. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
69. Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., et al.: Grandmaster level in starcraft ii using multi-agent reinforcement learning. nature **575**(7782), 350–354 (2019)
70. Wan Team: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
71. Wang, A., Huang, H., Fang, J.Z., Yang, Y., Ma, C.: Ati: Any trajectory instruction for controllable video generation. arXiv preprint arXiv:2505.22944 (2025)
72. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A.: Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291 (2023)
73. Wang, J., Zhang, Y., Zou, J., Zeng, Y., Wei, G., Yuan, L., Li, H.: Boximator: Generating rich and controllable motions for video synthesis. arXiv preprint arXiv:2402.01566 (2024)
74. Wang, S., Jin, J., Wang, X., Song, L., Fu, R., Wang, H., Ge, Z., Lu, Y., Cheng, X.: Video-thinker: Sparking "thinking with videos" via reinforcement learning. arXiv preprint arXiv:2510.23473 (2025)
75. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4581–4591 (2019)
76. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European conference on computer vision. pp. 396–416. Springer (2024)
77. Wei, J., Sun, Z., Papay, S., McKinney, S., Han, J., Fulford, I., Chung, H.W., Passos, A.T., Fedus, W., Glaese, A.: Browsecomp: A simple yet challenging benchmark for browsing agents (2025), `https://arxiv.org/abs/2504.12516`, accessed 30 June 2026
78. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
79. Xia, H., Ge, H., Zou, J., Choi, H.W., Zhang, X., Suradja, D., Rui, B., Tran, E., Jin, W., Ye, Z., et al.: Sportr: A benchmark for multimodal large language model reasoning in sports. arXiv preprint arXiv:2511.06499 (2025)
80. Xia, H., Yang, Z., Zou, J., Tracy, R., Wang, Y., Lu, C., Lai, C., He, Y., Shao, X., Xie, Z., et al.: Sportu: A comprehensive sports understanding benchmark for multimodal large language models. In: International Conference on Learning Representations. vol. 2025, pp. 29892–29950 (2025)

81. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
82. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1645–1653 (2017)
83. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025)
84. Xu, J., Zhao, G., Yin, S., Zhou, W., Peng, Y.: Finesports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21773–21782 (2024)
85. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
86. Yang, Z., Wang, S., Zhang, K., Wu, K., Leng, S., Zhang, Y., Li, B., Qin, C., Lu, S., Li, X., et al.: LongVT: Incentivizing "thinking with long videos" via native tool calling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 33816–33826 (2026)
87. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)
88. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning. In: International Conference on Learning Representations (2020)
89. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
90. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9127–9134 (2019)
91. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations. pp. 543–553 (2023)
92. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 32903–32914 (2026)
93. Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Hassan, A., Celikyilmaz, A., Liu, Y., Qiu, X., et al.: Qmsum: A new benchmark for query-based multi-domain meeting summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5905–5921 (2021)
94. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025)

95. Zhou, L., Xu, C., Corso, J.: Towards automatic learning of procedures from web instructional videos. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
96. Zhou, S., Xu, F.F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., et al.: Webarena: A realistic web environment for building autonomous agents. In: International Conference on Learning Representations. vol. 2024, pp. 15585–15606 (2024)