BASKETBALL-QA: DIAGNOSING PERCEPTION AND REASONING IN MULTI-PLAYER BASKETBALL VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) show increasingly strong capabilities in sports video analysis, including question answering, commentary generation, and event recognition. However, existing evaluations rarely combine multiple interacting players, dense sequences of events, and fine-grained spatial and temporal relations within a single video understanding task. We introduce Basketball-QA, a basketball video understanding benchmark with over 8K questions covering player matching, spatial grounding, action recognition, temporal grounding, event relations, counting, hallucination and tactical analysis. Solving these questions requires models to jointly identify relevant players, determine their actions and locations, and associate events across time. In end-to-end evaluation of video-input MLLMs, the best multiple-choice accuracy is 76.0% and the best open-ended score is 2.87/5, revealing substantial limitations in complex multi-player basketball scenes. We then localize the source of these failures through a series of increasingly targeted experiments. Probing atomic perception tests show that errors already arise when recovering the underlying facts. When correct perceptual facts are instead provided in text, a small language model without any task-specific fine-tuning achieves 78.5% multiple-choice accuracy, surpassing the best video-input model we evaluate. Together, these experiments indicate that current MLLMs are primarily limited by the acquisition and association of fine-grained visual evidence on Basketball-QA, rather than by reasoning over correctly provided perceptual information. We hope Basketball-QA will support the systematic evaluation and improvement of MLLMs for complex sports video understanding.

1 INTRODUCTION

Sports video understanding is shifting from perception-first pipelines, in which detection, tracking, and action recognition provide structured signals for downstream analysis, toward direct end-to-end understanding with multimodal large language models (MLLMs) (Li et al., 2023; Liu et al., 2023; Bai et al., 2025), and video-language models extend these capabilities to video description and conversational question answering (Maaz et al., 2024; Lin et al., 2024; Cheng et al., 2024). Unlike general or long-video understanding, which centers on narration, plot summarization, and long-range temporal dependencies, sports understanding focuses on fine-grained and objectively verifiable facts. Understanding a play involving multiple interacting players requires identifying who did what, where, and when, and relating these observations across participants and time. A fluent description, however, does not guarantee that the underlying events were correctly recovered. A model may recognize a shot while confusing the shooter, the origin of the pass, or the defender's position, and one such error is enough to distort the interpretation of the entire play. More importantly, the same incorrect description has multiple possible causes. The model may have failed to perceive the relevant facts, or it may have reasoned incorrectly over facts it had perceived. This ambiguity echoes broader concerns about unreliable fine-grained perception and visual hallucination (Fu et al., 2024; Rahmanzadehgervi et al., 2024; Guan et al., 2024), and raises a central question. **When a model fails to understand multi-player interactions, does the difficulty lie in acquir-**

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

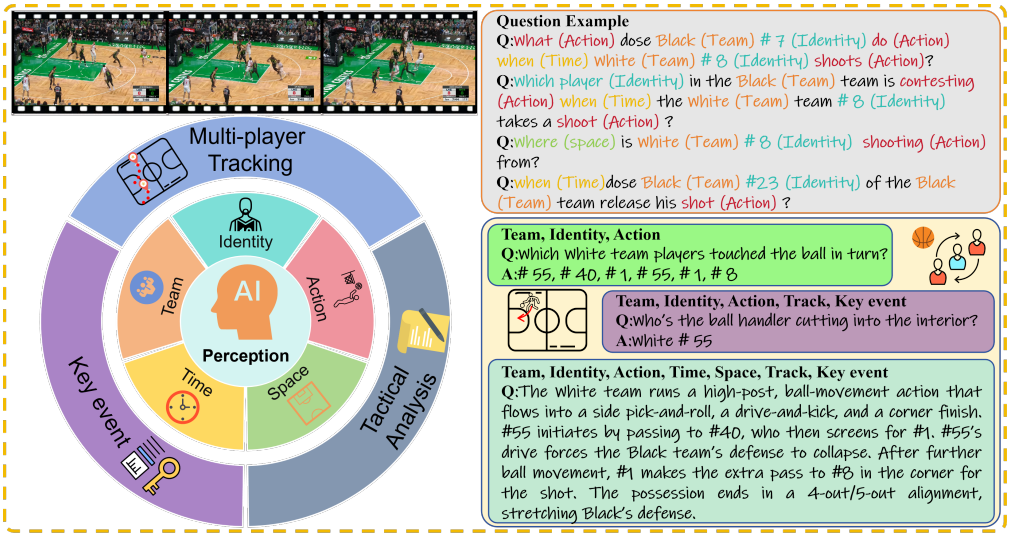


Figure 1: Hierarchical capability structure with example questions. The five inner dimensions represent atomic perceptual facts, composing these facts further supports multi-player tracking and key-event understanding. The questions on the right progressively extend from atomic facts to combinations such as team, identity, and action, and finally integrate trajectories and key events for tactical analysis, requiring evidence that binds who, where, what, and when across players and across time.

ing and associating visual facts, or in reasoning over facts that have already been correctly established?

To investigate this question, we introduce Basketball-QA, a fine-grained video understanding benchmark with over 8K questions grounded in 209 clips from professional basketball broadcasts spanning multiple seasons. Figure 1 organizes the capabilities around five basic evidence types, namely *team*, *identity*, *action*, *time*, and *space*. Their combinations support questions about player matching, spatial and temporal grounding, event relations, and trajectories, extending to tactical analysis that integrates multiple players' behavior. Multiple-choice and open-ended questions assess both answer selection and detailed factual reconstruction, and false-premise questions additionally test whether models reject unsupported claims about players, actions, or locations.

In end-to-end evaluation, the best of 20 video-capable models reaches 76.0% multiple-choice accuracy, while the highest open-ended score is 2.87/5, revealing substantial limitations on Basketball-QA. To localize the source of these failures, we further run a direct perception diagnostic that asks models to report players, key events, and trajectories. The best mean F1 scores are 0.63 for event timing, 0.85 for event content, and 0.48 for trajectories, showing that errors already arise when recovering the underlying facts. Finally, in place of raw video inputs, we provide the small language model with human-corrected, structured textual descriptions that capture the fine-grained details of the clip. Under structured-facts input, Qwen3.5-4B improves over its video-only results from 28.8% to 78.5% on multiple-choice accuracy, and from 1.62/5 to 4.55/5 on open-ended questions, surpassing every evaluated video-capable model. These results show that verified evidence supports strong question answering. Across the three experiments, the dominant failure in end-to-end video QA is the recovery and association of fine-grained visual evidence.

Our contributions are twofold. First, we introduce Basketball-QA, a benchmark that evaluates fine-grained multi-player video understanding through combinations of player, action, spatial, and temporal evidence, including false-premise tests and tactical questions. Second, we combine a systematic video QA evaluation, a direct perception diagnostic, and structured-facts input experiments to examine the sources of failure. These complementary analyses reveal substantial weaknesses in recovering and associating visual facts and demonstrate how explicitly available evidence supports question answering by small language models.

Table 1: Coverage of dimensions relevant to fine-grained basketball video understanding. A checkmark indicates that the dimension is explicitly and repeatedly evaluated, a cross indicates that it is not a central evaluation target.

Dimension	Sports-QA	SPORTU	SoccerChat	PL-NBA	Basketball-QA
Professional broadcast video	✓	×	✓	✓	✓
Multiple sports	✓	✓	×	×	×
Action/event recognition	✓	✓	✓	✓	✓
Temporal or causal event relations	✓	×	✓	✓	✓
Rule/foul explanation	×	✓	✓	×	×
Jersey-number recognition	×	×	×	×	✓
Player matching	×	×	×	✓	✓
Court spatial grounding	×	×	×	×	✓
Temporal action localization	×	×	×	✓	✓
Player trajectory	×	×	×	×	✓
Multi-player simultaneous state	×	×	×	×	✓
Tactical analysis	×	×	×	×	✓
Hallucination	×	×	×	×	✓

2 Related Work

Video QA is a standard way to evaluate video-capable MLLMs on temporal localization, event order, causality, and situated reasoning. Early datasets established open-ended and spatio-temporal question answering (Xu et al., 2017; Jang et al., 2017; Xiao et al., 2021; Wu et al., 2024a; Yu et al., 2019). Recent suites such as EgoSchema (Mangalam et al., 2023), MVBench (Li et al., 2024), Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), LongVideoBench (Wu et al., 2024b), TempCompass (Liu et al., 2024), and Perception Test (Patraucean et al., 2023) extend coverage to long-form, long-context, and diagnostic settings. While these benchmarks characterize broad capabilities, end-to-end answers emerge from the interaction of visual perception, question interpretation, reasoning, and answer composition. Answer accuracy alone is therefore insufficient either to establish reliable recall of persistent, entity-specific facts or to attribute failures to evidence retrieval rather than downstream reasoning. Fine-grained perception diagnostics corroborate this concern. Work such as BLINK, Blind, HallusionBench, and PerceptionBench (Fu et al., 2024; Rahmanzadehgervi et al., 2024; Guan et al., 2024; Lin et al., 2026) shows that MLLMs can produce fluent answers while misreading visual details or hallucinating unsupported content. These findings motivate evaluating whether model responses remain grounded in visual evidence, not merely whether the final answer is correct.

Sports video provides a demanding testbed for studying these perceptual facts. Foundational resources (Li et al., 2021; Shao et al., 2020; Xu et al., 2024; Cui et al., 2023; Giancola et al., 2018; Deliege et al., 2021) evaluate multi-athlete action localization, fine-grained action hierarchies, multi-object tracking, and broadcast-camera understanding. These tasks expose the visual primitives needed for language-based sports understanding, such as player identity, action, location, and time.

Building on these perceptual tasks, recent benchmarks evaluate sports understanding through language-based questions and explanations. Sports-QA (Li et al., 2026) poses descriptive, temporal, causal, and counterfactual questions over multi-sport professional footage, and a typical question asks "What do the players perform before performing straight jump?". SPORTU (Xia et al., 2025) targets rule violation recognition and subtle foul explanation, with questions such as "What kind of foul does this video show?". SoccerChat (Gautam et al., 2025) grounds multimodal conversation in soccer refereeing and game events, asking questions such as "How did the referee and linesman signal the offside decision against the red/blue-jerseyed team?". PL-NBA (Zhao et al., 2026) labels every sub-event of an offensive possession with its type, outcome, and acting player, and evaluates recognizing the full event sequence, localizing events in time, and anticipating offense tendency. These examples reflect distinct emphases, sequential actions and event relations in Sports-QA, rule violations in SPORTU, refereeing decisions in SoccerChat, and possession-level event sequences in PL-NBA. Across all four, the central task is identifying which events, rules, or decisions apply to a single play. Basketball-QA instead makes the joint binding of a specific player's identity to court positions, action times, and simultaneous multi-player states the central evaluation target, it further

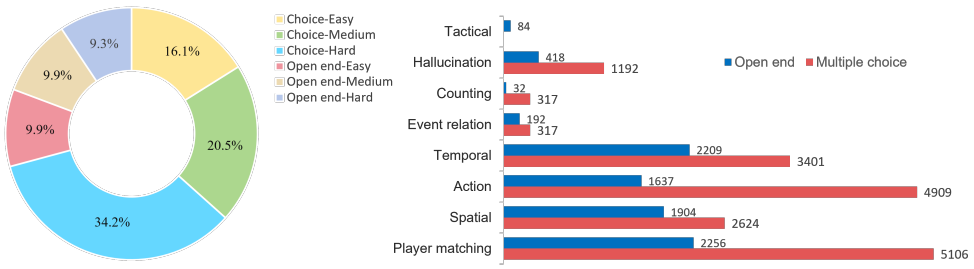


Figure 2: As illustrated by the donut chart on the left, multiple-choice questions are predominantly concentrated in the hard split, whereas open-ended questions are distributed more evenly across difficulty levels. The bar chart on the right further indicates that capability labels are non-exclusive, and therefore counts may exceed the total number of questions.

extends questions to tactical analysis over multiple players and adds false-premise tests that assess whether MLLMs reject claims about players, actions, or locations that are unsupported by the video. Table 1 summarizes the differences in benchmark coverage.

3 Basketball-QA Benchmark

3.1 Benchmark overview

We introduce Basketball-QA to investigate the limitations of atomic event perception and compositional reasoning in basketball offensive-defensive possessions. Each question is grounded in a possession-level clip and probes whether a model can recover individual facts about players and events and combine them to reconstruct what happened during the play. The benchmark contains over 8K question-answer pairs grounded in 209 clips from professional basketball broadcasts spanning multiple seasons. The clips are approximately 3-30 seconds long and depict complete possessions or compact offensive-defensive sequences. As summarized in Figure 2, the benchmark comprises approximately 5.6K multiple-choice questions and 2.4K open-ended questions, distributed over capability dimensions.

The questions cover *player matching*, *spatial grounding*, *action recognition*, *temporal grounding*, *event relations*, *counting*, and *hallucination detection*. Player matching associates the actions or locations queried in a question with the correct player using team membership and jersey number. Spatial grounding identifies court regions. Action recognition distinguishes specific actions. Temporal grounding determines when an action or state occurs. Event relations connect events and their participants, such as who receives the ball after a pass. Counting aggregates relevant players or events, while hallucination detection requires rejecting false premises involving nonexistent players, actions that did not occur, or locations a player never reached. Open-ended questions additionally cover tactical analysis, which combines players, locations, actions, and event relations to explain a play. Labels are non-exclusive, so one question may test several capabilities.

Questions are divided into easy, medium, and hard levels according to whether they require two, three, or at least four core evidence types, respectively. A representative hard multiple-choice item illustrates why some questions require the integration of several evidence types. *"Which player completed the first drive into the paint in this clip?"*. Answering this question requires several perception and reasoning steps. A drive is not a single visually explicit action. It involves rapid movement toward the paint while dribbling. The model must trace each candidate player's trajectory and determine whether that trajectory reaches the semantically defined paint region. It must then order the qualifying drives in time and identify the earliest player by reading the jersey number. Same-team distractors make this identity binding especially important. Errors in trajectory tracking, dribble detection, temporal ordering, jersey-number recognition, or court-region interpretation may each lead to an error answer. Figure 6 in Appendix A shows representative benchmark questions together with their clip frames.

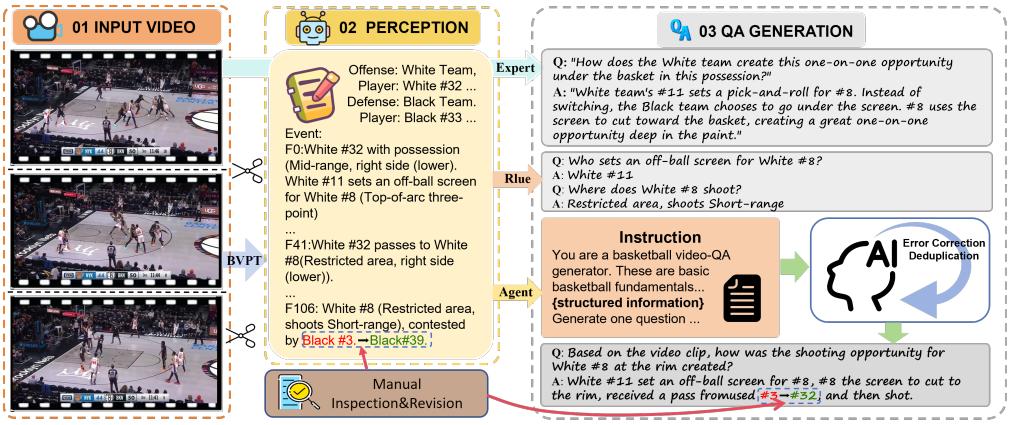


Figure 3: Construction pipeline for Basketball-QA, with four steps: (1) clip selection, (2) structured facts extraction by BVPT, (3) automated and expert-written QA generation, (4) human verification of the structured facts and LLM-generated QA.

3.2 Benchmark construction and quality control

Figure 3 presents the construction pipeline. Broadcast videos are first segmented into the short possessions. The Basketball Video Perception Pipeline (BVPT) extracts player tracks, team assignments, jersey numbers, and actions, and consolidates them, together with their timestamps, into structured match facts (details in Appendix B). These records are candidates rather than ground-truth annotations. Question generation proceeds along two paths, automated generation from the structured facts, and manual authoring by domain experts. The automated path comprises two complementary branches covering all difficulty levels. The rule-based generator instantiates QA pairs deterministically in six categories, temporal localization, possession flow and passing chains, counting, individual play patterns, hallucination and negation detection, and cross-entity states. The LLM generator composes more varied multi-fact questions per clip under per-difficulty quotas, with semantic deduplication against previously generated questions. Because the correct answer of a hallucination question is an explicit rejection, such questions invite a shortcut, when only one option is phrased as a negation, a model can select it by surface form without examining the video. Therefore, generated hallucination multiple-choice questions pass through a distractor enhancement step, in which an isolated LLM reads the question together with its correct answer and produces a distractor of the same negation style, replacing an existing option. After enhancement, a question contains two negation-style options of which only one contradicts the verified record, so the negation form no longer uniquely identifies the correct answer. A complete instance is shown in Figure 6. The manual path covers tactical analysis, a domain expert first reviews the clips, selects those that indeed contain tactical plays, and writes the questions and reference answers directly from the original video. A second expert then independently reviews them against the original video. This subset therefore bypasses the structured facts and the automatic generation pipeline, and its reference answers reflect the tactical interpretation that domain experts form from the raw video.

Human verification operates before and after generation processes. Before generation, annotators verify every structured fact against the source video, correcting player identities, actions, locations, timestamps, and event relations. The LLM-generated questions pass an isolated-LLM filter that rejects items answerable from common basketball knowledge alone, inconsistent with the verified records, or ambiguous. Rule-based questions are exempt because their templates instantiate the verified records directly. Surviving questions must also satisfy format-specific requirements, a unique correct answer with locally plausible distractors for multiple choice, and explicit rejection of an unsupported player, action, or location for hallucination questions. Finally, annotators review a sampled subset of the filtered questions for factual consistency, dependence on clip-specific evidence, reference-answer completeness, ambiguity, and duplication. Systematic issues found at any stage are fed back into the extraction, generation, or filtering procedures.In our benchmark, the LLM generator, the hallucination distractor generator, and the deduplication LLM all use Qwen-Plus(Qwen Team, 2025).

4 EXPERIMENTS

We use three experiments to distinguish evidence acquisition from evidence use on the benchmark summarized in Figure 2. First, end-to-end video QA evaluates 20 models to establish the overall failure pattern in Section 4.2. Second, direct atomic-perception evaluation tests whether models can recover atomic facts including player-, event-, and trajectory-level facts from the video in Section 4.3. Third, structured facts evaluation tests how effectively small language models answer benchmark questions when human-verified video facts are supplied explicitly, and how model capacity and supervised fine-tuning (SFT) affect evidence use in Section 4.4.

4.1 EXPERIMENTAL SETUP

We evaluate 20 video-capable models, 12 proprietary APIs, Gemini-3.1-Pro (Google DeepMind, 2026a), Gemini-3.5-Flash (Google DeepMind, 2026b), and Gemini-3.8-Flash (Google DeepMind, 2026c) from Google DeepMind. GPT-4o (Hurst et al., 2024), GPT-5.5 (OpenAI, 2026a), and GPT-6-Astra (OpenAI, 2026b) from OpenAI. Claude-Opus-4.8 (Anthropic, 2026a) and Claude-Opus-5 (Anthropic, 2026b) from Anthropic. Qwen3.5-Omni-Plus (Qwen Team, 2026b), Qwen3-VL-Plus (Bai et al., 2025), and Qwen-VL-Plus (Bai et al., 2023) from Alibaba. GLM-4.6V (Hong et al., 2025) from Zhipu AI. 8 open-weight MLLMs of varied scales run locally, LLaVA-NeXT-Video-7B (Zhang et al., 2024), Qwen3-VL-8B-Instruct (Bai et al., 2025), InternVL3.5-248B-A28B and InternVL3.5-8B (Wang et al., 2025a), InternVideo2.5-Chat-8B (Wang et al., 2025b), VideoLLaMA3-7B (Zhang et al., 2025), Qwen3.5-4B (Qwen Team, 2026a), and VideoChat-Flash-Qwen2.5-7B (Li et al., 2025). In the main QA evaluation, to control visual exposure, every evaluated model receives the same sampled input, uniform 2 fps sampling capped at 50 frames per clip at native resolution, provided as native video or as the corresponding ordered frames, depending on the interface. For main QA, all models share the same evaluation set, prompt template, and judging protocol. The system prompts are provided in Appendix C.

This paper adopts an LLM-as-judge paradigm, whereby all open-ended responses are scored by GLM-5.1 on a 1–5 factual-grounding rubric. A response counts as correct when it receives a 4 or a 5. Three human experts audit this scoring, the rubric is contained within the open-ended judge prompts in Appendix C, and the audit protocol and audit results are given in Appendix D.

4.2 MAIN RESULTS

Table 2 reports overall and per-capability results. The strongest multiple-choice model, Gemini-3.8-Flash, reaches 76.0%, with GPT-6-Astra close behind at 73.4%. Open-ended performance remains low even for the leading models, GPT-6-Astra attains the best mean judge score of only 2.87, indicating persistent difficulty in generating factually grounded answers. Three human experts auditing the judge agree with it on whether a response earns a 4 or a 5 on 85.0% to 100.0% of responses across models, as detailed in Appendix D, so the low open-ended scores reflect response quality rather than the judging protocol. Because aggregate accuracy obscures the source of failure, the per-capability columns report multiple-choice accuracy and open-ended judge scores for each capability label. Player matching is frequent because most questions identify or query a particular player, so label frequency does not itself indicate difficulty. The three strongest models by overall multiple-choice accuracy, Gemini-3.8-Flash, GPT-6-Astra, and Gemini-3.5-Flash, share a consistent capability profile. Averaged across the three, hallucination detection at 0.796 and event relations at 0.790 score highest, followed by action at 0.737, player matching at 0.711, and temporal reasoning at 0.686. Spatial grounding is weaker at 0.615, and counting is lowest at 0.516. This profile shows that even leading models struggle to recover precise spatial and numerical evidence despite stronger performance on event relations and false-premise detection.

Open-ended scores remain low for the same three models, per-capability means range from 2.31 to 2.70 across the seven capabilities shared with multiple-choice QA. Counting has the highest mean at 2.70, followed by temporal grounding at 2.60 and event relations at 2.58. Spatial grounding is lowest at 2.31. The tactical-analysis questions form a dedicated open-ended subset with a mean of 2.22. They demand a compositional account of multiple players, locations, actions, and their temporal or event relations, rather than recognition of one atomic fact.

Table 2: Overall and per-capability performance of video MLLMs. PM, Spa., Act., Tem., ER, Cnt., Hal., and TA denote *player matching*, *spatial grounding*, *action recognition*, *temporal grounding*, *event relation*, *counting*, *hallucination detection*, and *tactical analysis*, respectively. Capability labels follow Section 3 and overlap.

Model	Overall	PM	Spa.	Act.	Tem.	ER	Cnt.	Hal.	TA
Gemini-3.8-Flash	**.760**/2.48	**.762**/2.48	**.662**/2.34	**.795**/2.63	**.726**/2.73	.826/2.76	.581/2.97	**.916**/2.75	2.23
Gemini-3.5-Flash	.629/1.84	.631/1.84	.547/1.80	.658/1.86	.620/1.94	.683/1.98	.431/2.03	.764/1.65	1.62
Gemini-3.1-Pro	.619/2.05	.622/2.06	.576/1.97	.632/2.15	.616/2.65	.661/2.00	.428/2.03	.773/2.86	1.68
GPT-6-Astra	.734/**2.87**	.739/**2.88**	.637/**2.79**	.758/**3.02**	.713/**3.13**	**.861**/**2.99**	.537/**3.09**	.709/2.92	**2.80**
GPT-5.5	.531/1.86	.522/1.86	.451/1.86	.540/1.90	.541/2.09	.595/1.73	.495/1.94	.447/1.57	1.80
GPT-4o	.329/1.40	.305/1.40	.224/1.41	.349/1.38	.316/1.29	.480/1.32	.521/1.84	.207/1.03	1.30
Claude-Opus-5	.621/2.28	.611/2.28	.515/2.18	.632/2.32	.595/2.41	.756/2.44	.527/2.91	.516/2.03	2.16
Claude-Opus-4.8	.500/1.96	.497/1.97	.449/1.88	.504/2.04	.481/2.54	.632/1.87	.383/2.03	.511/**3.07**	1.59
Qwen3.5-Omni-Plus	.521/1.81	.508/1.81	.413/1.75	.547/1.85	.491/1.67	.729/2.07	.530/2.66	.459/1.13	1.86
Qwen3-VL-Plus	.423/1.59	.412/1.59	.330/1.52	.442/1.61	.417/1.90	.574/1.60	.495/1.75	.389/1.95	1.61
Qwen-VL-Plus	.389/1.36	.383/1.35	.316/1.34	.393/1.31	.369/1.21	.554/1.54	.505/1.81	.284/1.01	1.42
GLM-4.6V	.404/1.56	.398/1.56	.367/1.56	.410/1.52	.436/1.39	.489/1.50	.335/1.50	.470/1.05	1.54
InternVL3.5-248B-A28B	.376/1.34	.368/1.34	.303/1.34	.393/1.31	.366/1.32	.494/1.45	**.671**/1.53	.403/1.11	1.42
InternVL3.5-8B	.273/1.51	.265/1.51	.239/1.54	.286/1.44	.298/1.31	.438/1.49	.428/1.41	.112/1.03	1.54
InternVideo2.5-Chat-8B	.334/1.54	.304/1.54	.240/1.59	.340/1.47	.337/1.37	.410/1.56	.581/1.50	.212/1.00	1.64
VideoLLaMA3-7B	.340/1.49	.344/1.49	.269/1.52	.352/1.47	.352/1.47	.428/1.56	.406/1.66	.378/1.01	1.66
LLaVA-NeXT-Video-7B	.334/1.31	.339/1.32	.434/1.34	.282/1.30	.359/1.24	.346/1.29	.332/1.38	.028/1.03	1.15
Qwen3.5-4B	.288/1.62	.282/1.62	.201/1.66	.299/1.53	.283/1.29	.385/1.52	.313/1.44	.240/1.00	1.62
Qwen3-VL-8B-Instruct	.294/1.52	.314/1.53	.246/1.50	.310/1.49	.313/1.45	.405/1.44	.137/1.28	.426/1.38	1.62
VideoChat-Flash-Qwen2.5-7B	.289/1.20	.285/1.21	.246/1.23	.290/1.23	.291/1.34	.434/1.05	.332/1.00	.136/1.03	1.01

Multiple-choice accuracy shows consistent difficulty degradation, 19 of the 20 models lose accuracy from easier to harder questions, with relative reductions reaching 43% for Qwen3.5-4B, 34% for InternVideo2.5-Chat-8B, and 32% for Qwen-VL-Plus. The exception, InternVL3.5-8B, remains between 0.23 and 0.29 at every level, which is close to the 0.25 random-guessing level and is not meaningful for the difficulty analysis. Open-ended scores show a different pattern, they do not degrade with difficulty. Averaged over the same three strongest models, mean judge scores are 2.29 on easy, 2.40 on medium, and 2.50 on hard questions, and only 8 of the 20 models score lower on hard than on easy open-ended questions. The strongest models hold or improve their open-ended scores as difficulty increases. Hard open-ended questions typically anchor the players, moment, and location in the question stem, offloading the event localization and trajectory recovery that Section 4.3 identifies as the weakest capabilities, so stronger models can answer from recognition within the anchored moment. Easy questions instead require error-free multi-step descriptions. Table 6 in Appendix E reports the full multiple-choice and open-ended results at all three difficulty levels.

4.3 Direct atomic-perception evaluation

The results in Section 4.2 reveal substantial limitations in MLLMs' fine-grained understanding of basketball video. However, question-answering performance alone does not establish whether these limitations arise from inaccurate perception of the video or from subsequent reasoning, question interpretation, and answer composition. To investigate this distinction, we design a direct fine-grained perception experiment. Given a video, each model produces one report per clip in the structured format illustrated in Appendix B. The rule-based scorer parses these reports and compares them with the human-verified records. We compare these observations against human-verified facts along four dimensions, player matching, event timing, event content, and player trajectory, each computed as an F1 score that credits correct perception and penalizes hallucinated perception (more details in Appendix F).

Figure 4 summarizes the diagnostic results for the 20 models in Table 2, with each dimension averaged over the same 209 clips. Player matching is the strongest dimension, 15 of the 20 models exceed an F1 of 0.5, led by GPT-6-Astra at 0.96, and the five models below 0.5 are all open-weight systems. The other three dimensions are much weaker. On event content, only 7 models exceed 0.5. On event timing, only GPT-6-Astra exceeds 0.5 with 0.63, and every other system scores at most 0.31. On player trajectory, no system reaches 0.5, and the best score is 0.48. Averaged over the 20 systems, the mean F1 is 0.66 on player matching, 0.36 on event content, and only 0.14 on

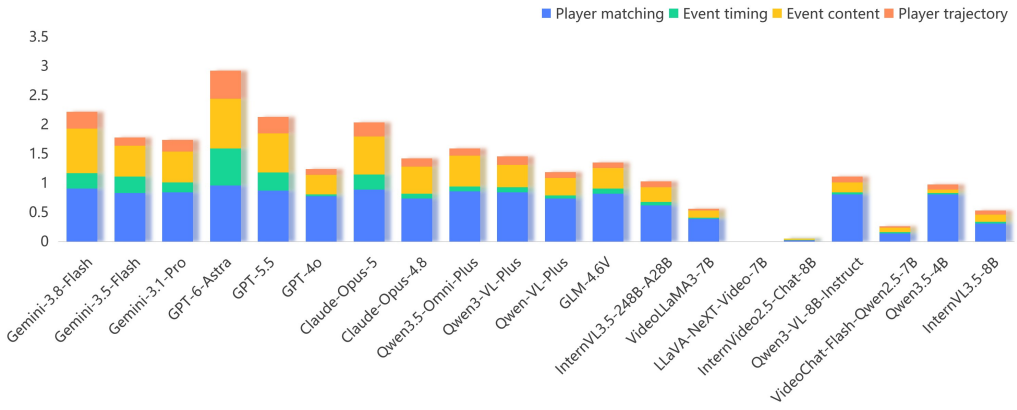


Figure 4: Direct fine-grained perception of basketball video by MLLMs.

player trajectory and 0.13 on event timing, showing that substantial errors are already present in direct reports of the video evidence. Detailed per-model scores are reported in Appendix Table 7.

Comparison with the end-to-end results in Section 4.2 shows a positive association between perceptual accuracy and benchmark performance. GPT-6-Astra leads all four dimensions and ranks second in overall multiple-choice accuracy, and Gemini-3.8-Flash, the strongest multiple-choice model in Table 2, ranks second on player matching and event content. Claude-Opus-5 ranks fourth in overall multiple-choice accuracy and fourth on event content. Open-ended results show the same association. However, excellence in a single dimension does not necessarily translate to strong overall benchmark performance. For example, Qwen3-VL-Plus scores 0.84 on atomic player matching while Gemini-3.5-Flash scores 0.83, yet Qwen3-VL-Plus achieves only 42.3% overall multiple-choice accuracy compared with 62.9% for Gemini-3.5-Flash. Likewise, Qwen3-VL-8B-Instruct reaches 0.81 on player matching but only 29.4% overall accuracy, with event-timing, event-content, and trajectory scores of just 0.03, 0.17, and 0.10. This discrepancy is consistent with the compositional design discussed in Section 3. Basketball-QA contains no single-dimension player-matching questions, and all questions combine multiple evidence dimensions. Answering even a player-focused question requires associating the correct identity with additional evidence about actions, locations, times, or event relations. Thus, proficiency in a single perceptual dimension is insufficient when the other required evidence cannot be reliably recovered and associated.

Overall, this section shows that current MLLMs still make substantial errors in directly perceiving players, events, and trajectories, with event timing and player trajectory the weakest dimensions. Perceptual accuracy is positively associated with benchmark performance across both multiple-choice and open-ended evaluation.

4.4 Evaluation from structured facts

Section 4.3 locates substantial errors in the direct recovery of visual facts. We therefore decouple perception from question answering, instead of video, models receive human-corrected structured facts. We provide this text to Qwen3.5 text models at 0.8B and 4B, each evaluated in its base version and after SFT. Training, validation, and test sets are drawn from Basketball-QA and partitioned at the clip level in a 7:1:7 ratio, keeping clips and their shared evidence reports disjoint across splits, so base and SFT models are evaluated on the same held-out test set. Open-ended responses are scored by the same judge as in the main evaluation. The supervised fine-tuning configuration, including the construction of the training data and all hyperparameters, is provided in Appendix G.

As shown in Table 3, with structured facts as the only input, the 4B base model already surpasses the best video-input models of Table 2 on both multiple-choice accuracy 78.5% and open-ended score of 4.55. After SFT, the 0.8B model likewise attains a high level on the same test set. These results indicate that the limiting factor is the reliability of the evidence rather than model scale or

Table 3: Overall and per-capability performance with structured facts input, using the format of Table 2. All results are computed on the clip-disjoint held-out test set (2,847 multiple-choice and 1,128 open-ended questions).

Model	Overall	PM	Spa.	Act.	Tem.	ER	Cnt.	Hal.	TA
Qwen3.5-0.8B (base)	.388/2.38	.393/2.44	.316/2.55	.399/2.20	.373/2.12	.507/1.58	.223/1.09	.133/1.44	1.38
Qwen3.5-0.8B (SFT)	.907/4.67	.899/4.70	.834/4.76	.921/4.63	.875/4.72	.919/3.47	.878/3.36	.970/4.86	1.80
Qwen3.5-4B (base)	.785/4.55	.773/4.65	.722/4.65	.802/4.58	.775/4.60	.817/4.05	.655/4.27	.753/4.56	2.24
Qwen3.5-4B (SFT)	.960/4.76	.956/4.79	.924/4.85	.965/4.73	.944/4.81	.964/4.16	.950/3.91	.991/4.86	1.80

direct video understanding, given accurate facts, a small-sized base model exploits them effectively without training, and task-specific training extends the same competence to very small models.

Tactical analysis tests both perception and basketball expertise, since a high-quality answer must reason about the relations among multiple players, such as the pick-and-roll, help defense, defensive rotations, and even the players' relative positioning and coordinated action timing. The tactical-analysis results of the small models above confirm the difficulty of this capability, even with structured facts as input, the 0.8B and 4B models score only between 1.38 and 2.24. We therefore select the 12 frontier models of Table 2 and answer the same tactical analysis questions with the structured facts. Figure 5 compares the two input conditions. Replacing the video input with structured facts improves every model, the average score rises from 1.80 to 3.09. The largest improvement is achieved by Gemini-3.5-Flash, increasing from 1.62 to 3.13.

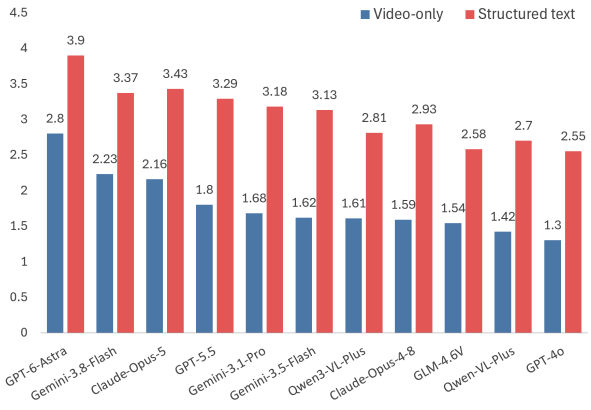


Figure 5: Tactical-analysis scores under video input and structured facts input for 12 frontier models from Table 2.

When accurate evidence is supplied, small models improve dramatically on QA, after SFT, the 0.8B model rises from 38.8% to 90.7% accuracy and from 2.38 to 4.67 in open-ended score, indicating that unreliable perception, rather than an inability to use provided facts, is the primary defect limiting their basketball understanding. Frontier models exhibit the same benefit, as verified evidence improves tactical analysis for all models. These results indicate that providing accurate perception substantially improves basketball understanding, while reasoning over multiple players remains a further challenge beyond perception.

5 Conclusion

We introduce Basketball-QA, a benchmark of over 8K questions built from 209 possession-level clips of professional basketball broadcasts. It evaluates fine-grained multi-player understanding across player, action, spatial, and temporal evidence. Our experiments proceed in three progressively stronger stages. End-to-end evaluation reveals that current MLLMs handle complex multi-player relations poorly. Direct atomic-perception tests then locate the weakness in perception. Supplying verified structured facts in place of the video shows the gain obtained once the perceptual evidence is correct. Together they indicate that the primary bottleneck for current MLLMs is the acquisition and association of fine-grained visual evidence rather than reasoning over facts that are already given. We hope Basketball-QA will contribute to progress in fine-grained multi-player video understanding.

AI USE STATEMENT

We used generative AI tools in two parts of this work. In benchmark construction, large language models assisted in drafting candidate questions and distractors from the BVPT-derived, human-verified structured facts, the generated QA pairs were subject to the sample-based human quality review described in Section 3. In paper preparation, AI tools assisted with drafting and polishing the manuscript text. AI tools did not design the experiments, run the evaluations, or perform the analyses, all technical content, claims, and conclusions are the authors' own. The LLM judge used for open-ended scoring is a disclosed component of our evaluation protocol (Section 4) and is audited by three human experts (Appendix D). We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Anthropic. Introducing claude opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, 2026a.

Anthropic. Introducing claude opus 5. https://www.anthropic.com/news/claude-opus-5, 2026b.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. doi: 10.48550/arXiv.2406.07476. URL https://doi.org/10.48550/arXiv.2406.07476.

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 (ICCV)*, pp. 9887–9897. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00910. URL https://doi.org/10.1109/ICCV51070.2023.00910.

Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4503–4514. IEEE, 2021. doi: 10.1109/CVPRW53098.2021.00508. URL https://doi.org/10.1109/CVPRW53098.2021.00508.

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. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://doi.org/10.1109/CVPR52734.2025.02245.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Sushant Gautam, Cise Midoglu, Vajira Thambawita, Michael A. Riegler, Pål Halvorsen, and Mubarak Shah. Soccerchat: Integrating multimodal data for enhanced soccer game understanding, 2025. URL https://arxiv.org/abs/2505.16630.

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1792–179210. IEEE, 2018.

Google DeepMind. Gemini 3.1 Pro. https://deepmind.google/models/gemini/pro/, 2026a.

Google DeepMind. Gemini 3.5: Frontier intelligence with action. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/, 2026b.

Google DeepMind. Gemini 3.8 Flash. https://deepmind.google/models/gemini/flash/, 2026c.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1359–1367. IEEE, 2017.

Haopeng Li, Andong Deng, Jun Liu, Hossein Rahmani, Yulan Guo, Bernt Schiele, Mohammed Bennamoun, and Qiuhong Ke. Sports-qa: A large-scale video question answering benchmark for complex and professional sports. *International Journal of Computer Vision*, 134 (5), 2026. doi: 10.1007/s11263-026-02734-1. URL https://doi.org/10.1007/s11263-026-02734-1.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, pp. 19730–19742, 2023. URL https://arxiv.org/abs/2301.12597.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024. doi: 10.1109/CVPR52733.2024.02095. URL https://doi.org/10.1109/CVPR52733.2024.02095.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2025.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13516–13525, October 2021. doi: 10.1109/ICCV48922.2021.01328. URL https://doi.org/10.1109/ICCV48922.2021.01328.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024. doi: 10.18653/v1/2024.emnlp-main.342. URL https://doi.org/10.18653/v1/2024.emnlp-main.342.

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. Perceptionbench: Evaluating atomic visual perception in multimodal large language models, 2026. URL https://arxiv.org/abs/2607.24957.

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

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? In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.517. URL https://doi.org/10.18653/v1/2024.findings-acl.517.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2024. URL https://arxiv.org/abs/2306.05424.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. volume 36, pp. 46212–46244, 2023.

OpenAI. GPT-5.5 system card. https://deploymentsafety.openai.com/gpt-5-5, 2026a.

OpenAI. GPT-6 Astra system card. https://deploymentsafety.openai.com/gpt-6-astra, 2026b.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5-4B. https://huggingface.co/Qwen/Qwen3.5-4B, 2026a.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026b.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Asian Conference on Computer Vision*, pp. 293–309. Springer, 2024.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622. IEEE, 2020.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025b.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. 2024a.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. 2024b.

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

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

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. doi: 10.1145/3123266.3123427. URL https://doi.org/10.1145/3123266.3123427.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21773–21782, June 2024. doi: 10.1109/CVPR52733.2024.02057. URL https://doi.org/10.1109/CVPR52733.2024.02057.

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. In *AAAI*, 2019. doi: 10.1609/aaai.v33i01.33019127. URL https://doi.org/10.1609/aaai.v33i01.33019127.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. LLaVA project blog, 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yunhao Zhao, Haoying Sun, Jiarui Li, Zhuming Wang, Ya Jing, Xiangbo Shu, Lifang Wu, and Changwen Chen. Pl-nba: A possession-level universal basketball video dataset supporting multiple visual understanding tasks, 2026. URL https://arxiv.org/abs/2608.19646.

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. In *CVPR*, 2025. doi: 10.1109/CVPR52734.2025.01278. URL https://doi.org/10.1109/CVPR52734.2025.01278.

A BENCHMARK EXAMPLE QUESTION

In this section, we illustrate several QA examples, as shown in Figure 6.

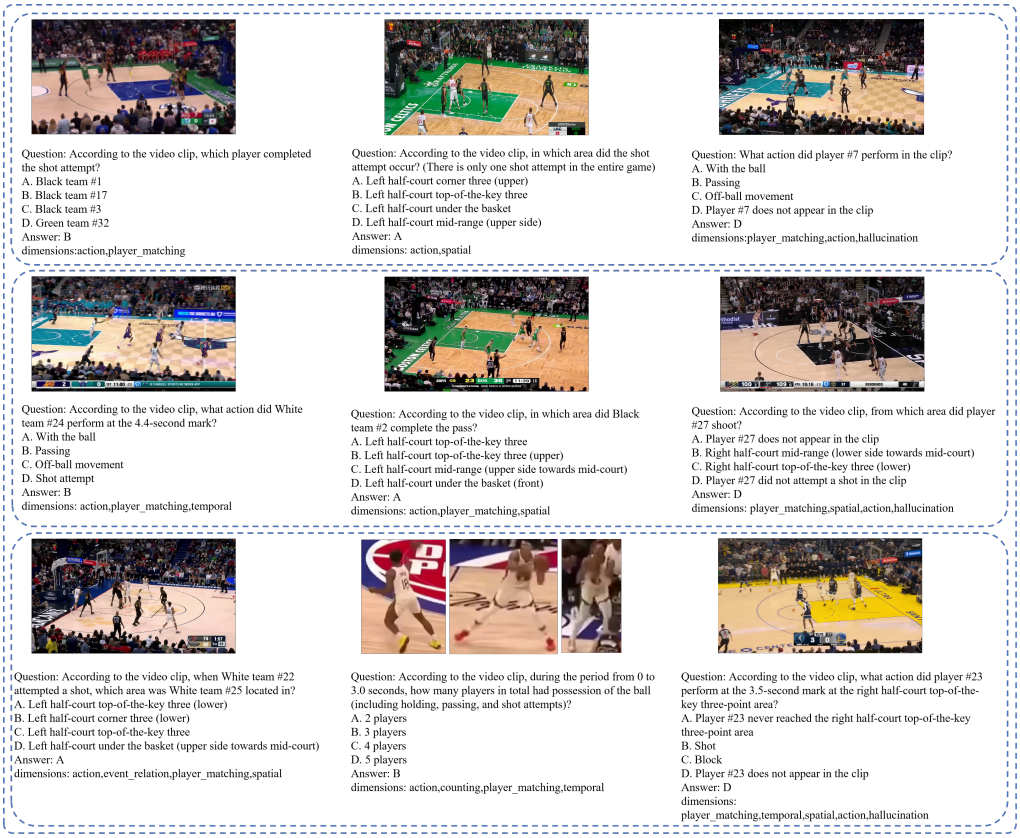


Figure 6: Representative benchmark QA instances. Rows 1–3 correspond to questions of easy, medium, and hard difficulty levels, respectively.

B BASKETBALL VIDEO PERCEPTION PIPELINE

The Basketball Video Perception Pipeline (BVPT) serves two supporting roles in this work. First, it is a *benchmark-construction and verification substrate*, it proposes structured facts from which candidate questions can be authored. The proposed facts are not accepted as ground truth automatically, and annotators inspect the records against the source clips and correct identities, trajectories, court regions, event timing, actions, and relations before question authoring. Second, these human-corrected records serve as reference facts for evaluation, they are the scoring reference of the direct-perception diagnostic in Section 4.3 and the structured facts input in Section 4.4.

BVPT converts basketball broadcasts into structured player- and event-level information through six stages. Player detection localizes players in each frame; cross-frame identity association links detections into stable tracks; team assignment separates these tracks into the two teams using uniforms and player features. Action recognition and jersey-number recognition then label each track with its jersey number and an action timeline that assigns to each time interval one of the standard action categories (holding the ball, off-ball movement, shot, pass, dribble, rebound, screen, and defensive contest). Finally, court calibration uses court markings and standard court geometry to estimate the mapping from the broadcast view to a canonical court plane, placing player locations in a common coordinate system. The zone labels in Figure 7 are canonical annotation names, not a requirement that models reproduce a particular string. We explicitly map physically equivalent basketball terminology to the same region, for example, “paint” or “restricted area” can match the basket-area label or the frontal short-range label it contains, “elbow” or “wing” can match the ap-

propriate toward-middle directional label, and "top of the key," "45-degree wing," and "corner" map to their corresponding three-point sectors. An example of structured facts is shown in Table 4. For brevity, only two players from each team are presented.

Table 4: An example of the structured facts extracted by BVPT.

```
Teams:  red vs.  black
Duration:  13.9s

Roster:
 red:  #13, #24, #44, #30, #1
 black:  #2, #1, #26, #25, #13

[red #30]
 [0.0s-2.6s] off-ball movement | right half-court, mid-range, left side (lower)
  movement:  right half-court, mid-range, left side (lower)
   -> right half-court, short-range, left side (lower)
 [2.7s-7.8s] off-ball movement | right half-court, short-range, left side (lower)
  movement:  right half-court, short-range, left side (lower)
   -> right half-court, restricted area, left side (lower)
 [7.9s-8.3s] off-ball movement | right half-court, restricted area, low baseline side toward middle
 [8.4s-9.0s] contest | right half-court, restricted area, low baseline side toward middle
 [9.1s-10.8s] off-ball movement | right half-court, restricted area, low baseline side toward middle
 [10.9s-11.0s] off-ball movement | right half-court, restricted area, left side (lower)
 [11.1s-11.8s] off-ball movement | right half-court, restricted area, low baseline side toward middle
 [11.9s-13.9s] off-ball movement | right half-court, restricted area, low baseline side

[red #44]
 [0.0s-1.5s] off-ball movement | right half-court, mid-range, high side toward middle
  movement:  right half-court, mid-range, high side toward middle
   -> right half-court, mid-range, high side (center-top)
 [1.6s-7.1s] off-ball movement | right half-court, mid-range, high side (center-top)
  movement:  right half-court, mid-range, high side (center-top)
   -> right half-court, short-range, high side toward middle
 [7.2s-8.5s] off-ball movement | right half-court, short-range, high side toward middle
  movement:  right half-court, short-range, high side toward middle
   -> right half-court, restricted area, high side toward middle
 [8.6s-9.0s] off-ball movement | right half-court, restricted area, high side toward middle
 [9.1s-10.1s] off-ball movement | right half-court, restricted area, high side (center-top)
 [10.2s-11.0s] off-ball movement | right half-court, restricted area, high side toward middle
 [11.1s-11.5s] off-ball movement | right half-court, short-range, high side toward middle
 [11.6s-11.7s] off-ball movement | right half-court, mid-range, high side toward middle
 [11.8s-12.2s] off-ball movement | right half-court, mid-range, left side (upper)
 [12.3s-12.8s] contest | right half-court, mid-range, left side (lower)
  movement:  right half-court, mid-range, left side (upper)
   -> right half-court, mid-range, left side (lower)
 [12.9s-13.1s] off-ball movement | right half-court, mid-range, left side (lower)
 [13.2s-13.9s] off-ball movement | right half-court, top-of-arc three-point (center)

[black #1]
 [0.0s-2.4s] holding the ball [possession] | left half-court, top-of-arc three-point (center)
  movement:  left half-court, top-of-arc three-point (center)
   -> right half-court, top-of-arc three-point (center)
  defender:  red #1
 [2.5s-2.8s] pass [possession] | right half-court, top-of-arc three-point (center)
  defender:  red #1
 [2.9s-4.3s] off-ball movement | right half-court, top-of-arc three-point (center)
  movement:  right half-court, top-of-arc three-point (center)
   -> right half-court, mid-range, left side (upper)
  defender:  red #1
 [4.4s-6.3s] off-ball movement | right half-court, mid-range, left side (upper)
  movement:  right half-court, mid-range, left side (upper)
   -> right half-court, top-of-arc three-point (center)
 [6.4s-8.4s] off-ball movement | right half-court, top-of-arc three-point (center)
 [8.5s-11.0s] out of frame
 [11.1s-11.7s] off-ball movement | right half-court, top-of-arc three-point (center)
 [11.8s-12.2s] holding the ball [possession] | right half-court, top-of-arc three-point (center)
 [12.3s-12.8s] shot attempt [possession] | right half-court, top-of-arc three-point (center)
 [12.9s-13.9s] off-ball movement | right half-court, top-of-arc three-point (center)

[black #2]
 [0.0s-0.5s] off-ball movement | right half-court, mid-range, baseline side toward middle
 [0.6s-3.2s] off-ball movement | right half-court, mid-range, left side (lower)
 [3.3s-4.8s] holding the ball [possession] | right half-court, mid-range, left side (lower)
 [4.9s-5.3s] pass [possession] | right half-court, top-of-arc three-point (center)
 [5.4s-5.8s] off-ball movement | right half-court, top-of-arc three-point (center)
 [5.9s-6.6s] screen | right half-court, top-of-arc three-point (center)
 [6.7s-7.1s] off-ball movement | right half-court, top-of-arc three-point (center)
  defender:  red #24
 [7.2s-8.1s] off-ball movement | right half-court, mid-range, left side (lower)
```

```
 movement:  right half-court, mid-range, left side (lower)
  -> right half-court, short-range, left side (lower)
[8.2s-8.5s] off-ball movement | right half-court, short-range, left side (lower)
 movement:  right half-court, short-range, left side (lower)
  -> right half-court, short-range, left side (upper)
[8.6s-9.0s] off-ball movement | right half-court, restricted area, left side (upper)
 movement:  right half-court, restricted area, left side (upper)
  -> right half-court, restricted area, high side toward middle
 defender:  red #24
[9.1s-9.4s] off-ball movement | right half-court, restricted area, high side toward middle
 defender:  red #24
[9.5s-9.7s] rebound [possession] | right half-court, restricted area, left side (upper)
 defender:  red #24
[9.8s-10.2s] holding the ball [possession] | right half-court, restricted area, left side (upper)
 defender:  red #24
[10.3s-10.6s] pass [possession] | right half-court, restricted area, left side (upper)
 defender:  red #24
[10.7s-10.9s] off-ball movement | right half-court, restricted area, left side (lower)
 defender:  red #24
[11.0s-12.8s] off-ball movement | right half-court, short-range, left side (lower)
 defender:  red #24
[12.9s-13.9s] off-ball movement | right half-court, short-range, left side (upper)
 defender:  red #24
```

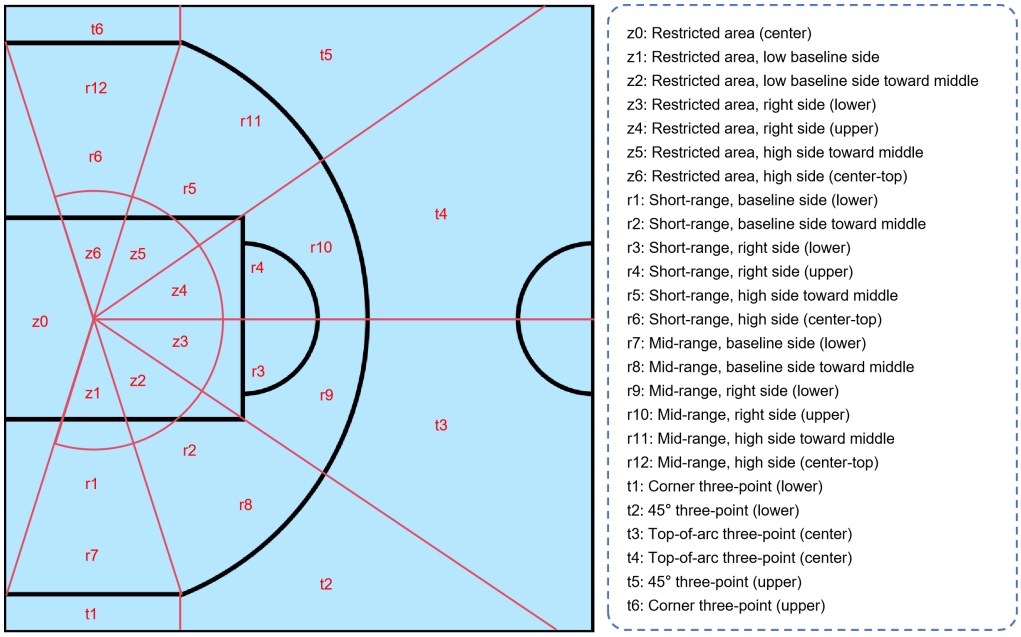


Figure 7: Default court-zone partition used by BVPT for spatial annotations. The half-court is divided by distance rings (`z*`, `r*`, and `t*`) and angular sectors, yielding stable near-range, mid-range, and three-point regions with directional qualifiers. These canonical labels support reproducible question generation and evaluation across camera views.

C Representative Prompt Examples

This section presents seven representative system prompts covering question authoring, compositional difficulty, hallucination distractor, model output constraints, and open-ended judging. Each prompt below is a system prompt, identified by its bold heading and operational role.

C.1 Question Authoring and Difficulty Prompts

Multi-choice QA authoring. "You are an expert NBA basketball-video benchmark question writer. Using the provided structured data, generate high-quality four-option multiple-choice questions. Every question must have exactly four substantive options (A–D), a unique answer, and an answer that

cannot be obtained from basketball common sense alone. Avoid yes/no wording: ask *who*, *which*, *what*, *where*, *when*, or *how many*."

Open-ended QA authoring. "You are an expert NBA basketball-video benchmark question writer. Using the provided structured data as the single source of truth, generate high-quality open-ended questions with no options. Each reference answer must be a concise, objective, and directly quoting the raw format of the structured data is forbidden. Every question must test recognition of behavior directly observable in the video: tactical-intent, causal, and possession-reasoning questions are forbidden, and long-temporal questions may describe only observed factual changes. Never use 'offense' or 'defense' as constraints, since possession may change within a clip; use team colors and jersey numbers instead. Ask exactly one question per item, name locations with the standard court-region labels, require trajectory evidence rather than basketball common sense, and label every dimension involved (player matching, temporal, spatial, action, plus temporal sequence, counting, or event relation when applicable)."

Difficulty control. "Difficulty equals the number of core evidence dimensions that must be jointly satisfied. The core dimensions are `action`, `id` (jersey identity), `temporal`, and `space`. `counting` and `event_relation` may be used only at hard difficulty; `hallucination` may occur at any difficulty. In the output schema, map `id`, `time`, and `space` to `player_matching`, `temporal`, and `spatial`. Do not use `temporal+action` or `space+action` when they leave the acting player ambiguous, except when the action is globally unique."

Hallucination-distractor enhancement. "Given a hallucination question and its correct answer option, generate a new distractor to replace the weakest option. Never negate the dimension the question asks about: the correct answer negates exactly one context dimension (player, action, or location), and the distractor must negate a different context dimension rather than the dimension already negated by the correct answer. Do not use temporal negations such as 'the clip is shorter than X seconds'."

C.2 Model-Evaluation and judging prompts

Video-model instruction. "You are an NBA game-video analysis expert. Answer only from the video content; do not guess or fabricate information absent from the video. Answer in concise and accurate English. Identify players by jersey number in the format '#number' (e.g., '#5', '#23') and teams by jersey color in the form 'the color team' (e.g., 'the white team'). Do not use player names. Use the standard action categories: holding the ball, off-ball movement, shot, pass, dribble, rebound, screen, and defensive contest. State times in seconds (e.g., 'at second 2.3') and describe locations by half court, left or right of the midline. Do not use the ambiguous labels 'offense' or 'defense' in place of a team or player, since possession may change within a clip; use the team color or jersey number instead. For multiple choice, output only the option letter (A, B, C, or D). For open-ended questions, answer strictly what is asked without off-topic content, and be concise and accurate."

Judge region-name mapping. When the judge must align tactical colloquialisms in a model response with the canonical zone names of the option text (e.g., "short range (lower side toward middle)"), the following mappings apply. (i) Distance rings: the restricted area (inner ring) equals directly under the basket and the no-charge semicircle; short range (middle ring) equals the upper-middle paint, the free-throw line and everything inside it, the elbows, and the short corners; mid range (outer ring) equals long two-pointers, the mid-range shooting area, and one step inside the three-point line; outside the three-point line equals beyond the arc, where the 45-degree three equals the 45-degree or wing three, the top-of-arc three equals top of the key, and the corner three equals the corner. (ii) Directional qualifiers (left half court; the right half court is symmetric): lower side toward middle and upper side toward middle equal the wing and the elbows; lower side and upper side equal the inside corner and the short corner; frontal (the "right side (upper/lower)" of the left half court and the "left side (upper/lower)" of the right half court) equals the middle of the free-throw line, inside the top of the arc, and the middle of the high post. (iii) Composite-zone containment: the paint, the key, and the restricted area contain "under the basket" and the frontal "short range"; the free-throw line and the elbows sit at the frontal or toward-middle boundary of "short range".

Open-ended judge. "Extract answer-specific facts from the reference answer and check whether the model supplies them, ignoring facts merely repeated from the question. Penalize vague directions and generic basketball-term lists. Treat physically equivalent court descriptions as matches

Table 5: Per-model expert audit of the LLM judge. Judge mean and Expert mean are average scores on the five-point rubric. Exact is the exact score agreement rate. Score-≥ 4 agreement is agreement after treating scores 4–5 as correct.

Model	Judge mean	Expert mean	Exact	Score-≥ 4 agree.
Gemini-3.8-Flash	3.00	3.30	57.0%	92.0%
Gemini-3.5-Flash	3.00	3.10	78.0%	98.0%
Gemini-3.1-Pro	3.00	3.10	67.0%	94.0%
GPT-6-Astra	3.00	3.20	86.0%	89.0%
GPT-5.5	3.00	3.00	76.0%	100.0%
GPT-4o	3.00	3.25	60.0%	93.0%
Claude-Opus-5	3.00	3.50	56.0%	90.0%
Claude-Opus-4.8	3.00	2.95	68.0%	88.0%
Qwen3.5-Omni-Plus	3.00	2.60	63.0%	91.0%
Qwen3-VL-Plus	3.00	3.05	57.0%	97.0%
Qwen-VL-Plus	3.00	2.95	79.0%	92.0%
GLM-4.6V	3.00	2.70	54.0%	98.0%
InternVL3.5-248B-A28B	3.00	3.10	68.0%	88.0%
InternVL3.5-8B	2.75	2.65	76.0%	99.0%
InternVideo2.5-Chat-8B	2.65	2.35	55.0%	100.0%
VideoLLaMA3-7B	2.40	2.50	66.0%	97.0%
LLaVA-NeXT-Video-7B	2.40	2.15	63.0%	90.0%
Qwen3.5-4B	2.72	2.72	70.0%	85.0%
Qwen3-VL-8B-Instruct	2.85	2.70	59.0%	89.0%
VideoChat-Flash-Qwen2.5-7B	2.00	1.05	45.0%	90.0%

(e.g., paint/restricted area covering the basket area and frontal short range, elbow/wing and the corresponding toward-middle sectors). Assign a score on the following rubric. 5: all reference-answer key facts are accurately matched; 4: most key facts are matched, with only minor details omitted or imprecise; 3: approximately half of the key facts are matched, with substantial omissions or errors; 2: almost no key details are matched, the response is vague, largely repeats the question, or supplies mostly incorrect details; 1: the response is wholly incorrect, off-topic, or repeats the question without substantive answer content."

D Expert audit of LLM judging

For each model, we randomly sampled 20 responses for each score level from 1 to 5. If a model's score distribution did not contain enough responses at a given level, the remaining quota was filled by sampling uniformly from the other score levels. Three experts then assigned the same 1–5 rubric to every sampled response, and the expert score reported in Table 5 is the mean across the three ratings. The experts did not use the judge's rationale when scoring.

E Performance by difficulty level

Section 4.2 analyzes performance by rule-based compositional difficulty. Table 6 reports the underlying per-level results. Multiple-choice accuracy degrades from easy to hard for 19 of the 20 systems, while open-ended judge scores show no corresponding degradation, and the strongest systems hold or improve their open-ended scores as compositionality increases.

F Direct-perception diagnostic scoring

Each report is scored against the reference facts by a purely rule-based scorer, along four independent dimensions, each computed as an F1 score and reported as a mean over all clips. Because the same fact can be phrased in different words, we first manually review the model reports and compile

Table 6: Per-difficulty performance of video MLLMs.

	Multiple-choice accuracy			Open-ended score		
Model	Easy	Medium	Hard	Easy	Medium	Hard
Gemini-3.8-Flash	**.846**	.691	**.760**	2.35	2.55	2.55
Gemini-3.5-Flash	.716	.595	.607	1.83	1.77	1.91
Gemini-3.1-Pro	.696	.593	.598	1.84	2.07	2.25
GPT-6-Astra	.840	**.695**	.706	**2.69**	**2.89**	**3.05**
GPT-5.5	.611	.558	.476	1.79	1.81	1.99
GPT-4o	.391	.307	.312	1.44	1.39	1.35
Claude-Opus-5	.775	.585	.567	2.29	2.25	2.31
Claude-Opus-4.8	.602	.478	.464	1.76	2.01	2.11
Qwen3.5-Omni-Plus	.667	.463	.484	1.88	1.69	1.88
Qwen3-VL-Plus	.510	.417	.383	1.51	1.60	1.68
Qwen-VL-Plus	.511	.355	.348	1.44	1.28	1.34
GLM-4.6V	.487	.397	.367	1.68	1.53	1.46
InternVL3.5-248B-A28B	.416	.357	.369	1.36	1.31	1.36
VideoLLaMA3-7B	.389	.367	.299	1.48	1.41	1.59
LLaVA-NeXT-Video-7B	.365	.400	.279	1.35	1.32	1.26
InternVideo2.5-Chat-8B	.423	.351	.280	1.67	1.45	1.51
Qwen3-VL-8B-Instruct	.343	.307	.261	1.58	1.60	1.38
VideoChat-Flash-Qwen2.5-7B	.376	.258	.264	1.13	1.17	1.30
Qwen3.5-4B	.380	.330	.218	1.79	1.62	1.42
InternVL3.5-8B	.233	.289	.283	1.64	1.45	1.43

the terms that are semantically consistent but differently phrased into an equivalence list, and the rules then judge a match by this semantic equivalence rather than by exact string overlap.

Player matching scores the predicted roster against the reference roster as a set F1. Predicted players that match no reference player enter the precision denominator, penalizing player hallucination. Event content scores the recall of reference key events (who did what), while key events reported by the model that match no reference event enter the precision denominator, penalizing event hallucination. Event timing is identity-agnostic, same-type events whose start and end times both lie within 0.5 s of a reference event are paired one-to-one by greedy matching, and the dimension is scored as an F1. Player trajectory scores each player's where-and-when sequence, a segment matches when the half-court direction agrees and the two regions are equivalent. Matching segments are paired one-to-one by greedy matching at an IoU of at least 0.3, and the dimension is a duration-weighted F1. Pairing follows a two-level, identity-decoupled strategy, segments are first paired by identity, and remaining segments are then matched globally against unconsumed reference segments, so a wrong identity lowers pairing priority rather than discarding the segment.

Every report is generated with an output limit of 16,384 tokens, content lost to truncation receives no credit, and a refusal or an otherwise empty report is scored as an error on every dimension. Table 7 reports the detailed per-system scores.

G SUPERVISED FINE-TUNING DETAILS

Both Qwen3.5 models are fine-tuned with the LoRA implementation of ms-swift, and every hyperparameter in Table 8 is shared across the two scales. Training runs on 4 GPUs with distributed data parallelism for 4 epochs at an effective batch size of 32, in bfloat16 precision with gradient checkpointing. Each training instance is built from one clip: the human-verified structured facts are concatenated as context, followed by the question, and the model learns to generate the reference answer. This construction matches the structured-facts input format of the evaluation, keeping training and evaluation distributions aligned.

Table 7: Detailed fine-grained visual perception of basketball video by MLLMs.

Model	Player matching	Event timing	Event content	Player trajectory
Gemini-3.8-Flash	.91	.26	.76	.29
Gemini-3.5-Flash	.83	.28	.53	.14
Gemini-3.1-Pro	.84	.17	.53	.20
GPT-6-Astra	.96	.63	.85	.48
GPT-5.5	.87	.31	.67	.28
GPT-4o	.78	.03	.33	.10
Claude-Opus-5	.89	.26	.65	.24
Claude-Opus-4.8	.74	.08	.46	.14
Qwen3.5-Omni-Plus	.86	.08	.53	.12
Qwen3-VL-Plus	.84	.09	.38	.15
Qwen-VL-Plus	.74	.05	.30	.10
GLM-4.6V	.82	.09	.35	.09
InternVL3.5-248B-A28B	.62	.06	.25	.10
InternVL3.5-8B	.31	.03	.12	.07
InternVideo2.5-Chat-8B	.02	.01	.02	.00
VideoLLaMA3-7B	.39	.02	.12	.03
LLaVA-NeXT-Video-7B	.00	.00	.00	.00
Qwen3.5-4B	.81	.02	.06	.09
Qwen3-VL-8B-Instruct	.81	.03	.17	.10
VideoChat-Flash-Qwen2.5-7B	.14	.02	.08	.01

Table 8: Hyperparameters of supervised fine-tuning, shared by both model scales.

Hyperparameter	Value
Base models	Qwen3.5-0.8B / Qwen3.5-4B (fine-tuned separately)
Chat template	non-thinking mode
Fine-tuning method	LoRA on all linear layers
LoRA rank / α / dropout	16 / 64 / 0.05
Optimizer	AdamW
Epochs	4
Learning rate	1e-4, cosine schedule, 5% warmup
Weight decay / gradient clipping	0.1 / 1.0
Effective batch size	32 (1 per device $\times$ 8 accumulation $\times$ 4 GPUs)
Maximum sequence length	16,384
Precision	bfloat16
Attention implementation	SDPA
Evaluation and save interval	every 50 steps

H LIMITATIONS

Clip scale and diversity. The benchmark contains over 8K questions but grounds them in only 209 possession-level clips of 3–30 seconds from professional broadcasts, so the underlying video diversity is limited. Moreover, the clips are public game footage, and some evaluated models may have seen these games during pretraining.

BVPT scope and error accumulation. BVPT is specific to basketball and does not transfer to other sports without redesign. Its multi-stage design, from detection and tracking through action recognition and court calibration, accumulates errors across stages, so every structured facts must be verified and corrected against the source video by human annotators. This verification guarantees record quality but makes construction labor-intensive and bounds benchmark coverage by what BVPT can extract, including a fixed action vocabulary and a canonical court-zone set, which currently limits scaling to more clips, leagues, and sports.

Expert-written tactical reference. The tactical-analysis questions are authored and reviewed by two experts, and their reference answers encode one tactical interpretation of each play, so alternative but defensible readings may receive low judge scores.