BASKETVISION: BENCHMARKING MLLMS' GRASP OF COMPLEX DYNAMIC SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

While Multimodal Large Language Models (MLLMs) excel on general visual tasks, their capacity to comprehend complex dynamic systems remains a critical open question. Such systems, governed by physical laws, explicit rules, and multi-agent interactions, form the fabric of the real world. To facilitate a systematic diagnosis of current MLLM limitations, we introduce BasketVision, a new benchmark that leverages professional basketball as a microcosm for these dynamic environments. BasketVision probes model capabilities across seven dimensions—spanning perception, reasoning, and prediction—through 6,000 curated, bilingual questions from professional game data. An automated data generation pipeline underpins the benchmark, ensuring both scalability and fine-grained precision. Our evaluation of 23 leading models reveals a chasm between machine and human cognition: human experts attain 96.34% accuracy, while the premier model, GPT-4o, achieves only 63.15%. The analysis pinpoints spatial reasoning as a persistent bottleneck and uncovers specific patterns of task specialization. BasketVision thus serves as a crucial apparatus for charting the frontiers of MLLMs and steering future work toward more robust reasoning in dynamic visual worlds.

1 INTRODUCTION

The impressive strides made by Multimodal Large Language Models (MLLMs) have reshaped the landscape of visual understanding, with leading models attaining human-competitive performance on canonical benchmarks for tasks like VQA and image captioning (Yin et al., 2023; Antol et al., 2015; Goyal et al., 2017). Yet, these achievements largely stem from tasks centered on static images or isolated video clips. A more profound challenge lies in assessing and enhancing their capacity to interpret *complex dynamic systems*—environments defined by the interplay of multiple agents governed by shared rules and physical principles (Liu et al., 2022). From urban traffic to team sports, the real world abounds with such systems. Making sense of them demands more than object recognition; it requires a deep grasp of evolving spatial configurations, temporal dependencies, and the underlying goals and strategies that drive actions.

Current evaluation paradigms, however, prove inadequate for this task. Many benchmarks are confined to static imagery (Fang et al., 2024; Yu et al., 2024) or feature generic video content drawn from daily life (Li et al., 2024; Ning et al., 2023). This leaves them ill-equipped to probe model reasoning in the high-density, structured interactions characteristic of rule-governed domains. They seldom necessitate the fine-grained spatial and temporal precision vital for true comprehension, such as verifying if a player is out of bounds or forecasting the success of a coordinated maneuver. Furthermore, the prohibitive manual effort of annotating such intricate events has historically constrained the scale and diversity of evaluation datasets (Li & Lu, 2024), creating a persistent disconnect between MLLM evaluation and the exigencies of real-world dynamic scene analysis.

We address this gap by positing that professional basketball serves as an ideal *microcosm* for the systematic diagnosis of MLLM capabilities. A basketball game is a high-tempo, multi-agent setting with a clear objective function, constrained by explicit rules like shot clocks and fouls and rich with emergent strategies. True understanding demands a hierarchy of skills, from low-level perception to high-level tactical inference.

To this end, we introduce BasketVision, a multimodal benchmark engineered to evaluate MLLMs in this demanding context. This paper's contributions are three-fold. First, we present a new bench-

mark of 6,000 curated questions in both image and video formats, designed to systematically assess MLLM proficiency across seven distinct dimensions of spatio-temporal reasoning and strategic analysis. Second, we detail an automated data generation pipeline that integrates court recognition, perspective transformation, and player tracking, enabling the creation of spatially-grounded questions at scale and ensuring the benchmark's extensibility. Third, our extensive evaluation of 23 state-of-the-art MLLMs on BasketVision uncovers a profound 33-point performance deficit between human experts (96.34%) and the most capable model (GPT-4o at 63.15%), pinpointing spatial reasoning as a systemic weakness and revealing intriguing patterns of task specialization.

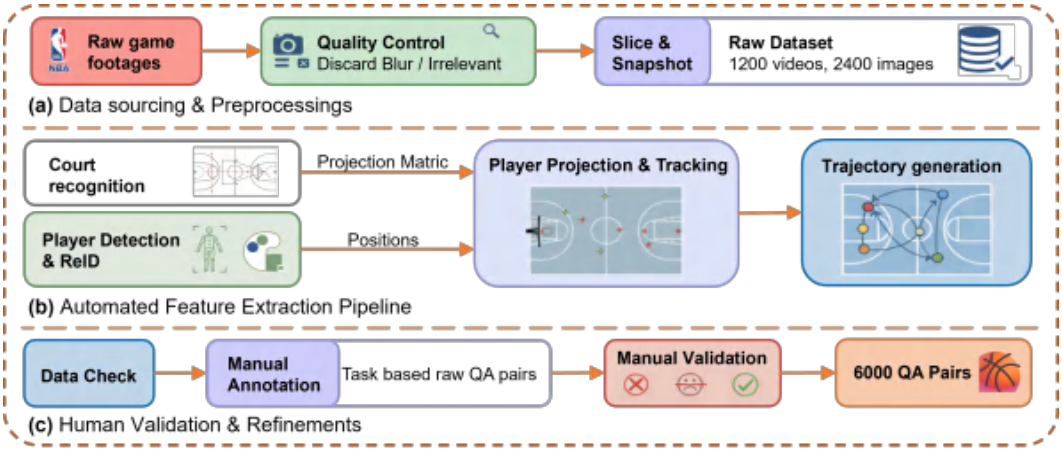

Figure 1: Overall data generation pipeline of BasketVision.

2 Related Works

Early benchmarks focused on static image understanding in general domains. The initial wave of MLLM evaluation centered on static images. Foundational benchmarks like MME (Fu et al., 2023), MMBench (Fang et al., 2024), and MM-Vet (Yu et al., 2024) established holistic assessments of core skills, but their general domains offer limited purchase on performance within structured, dynamic settings.

Subsequent efforts delved into more specialized yet still static perceptual abilities. Later works targeted specific skills. Benchmarks like Flickr30k Entities (Plummer et al., 2017) addressed fine-grained grounding, while ViewSpatialBench (Li et al., 2025) began exploring spatial reasoning. However, these static evaluations inherently cannot capture the temporal evolution and multi-agent dynamics defining interactive systems.

The focus then shifted to video benchmarks for general temporal understanding. Recognizing the constraints of images, the community turned to video. A new generation of benchmarks, including MVBench (Li et al., 2024) and comprehensive suites like Video-Bench (Ning et al., 2023) and Video-MME (Fu et al., 2025), emerged to evaluate temporal understanding across a broad spectrum of tasks.

However, existing video benchmarks still fall short in evaluating complex, rule-governed systems. A critical limitation is that these benchmarks seldom feature the highly structured, rule-governed nature of our target domain. Even specialized benchmarks for phenomena like video hallucination, such as VideoHallucer (Wang et al., 2024) and EventHallusion (Zhang et al., 2024), focus on generic events rather than domain-specific, strategic actions. While the aforementioned benchmarks laid critical groundwork, recent research highlights a growing consensus on the persistent shortcomings of MLLMs in spatiotemporal reasoning. Works like VLM4D (Zhou et al., 2025) and OST-Bench (Lin et al., 2025) have introduced benchmarks specifically to probe these deficiencies in video, finding that models struggle to maintain temporal coherence and integrate visual inputs with historical memory. Concurrently, other studies have systematically analyzed the root causes of these failures, identifying architectural limitations and data deficiencies as key culprits, particu-

larly for tasks requiring spatial imagination across different views (Zhang et al., 2025) or complex 2D-to-3D visualization (Wang et al., 2025). Even in highly structured domains like basketball, specialized models based on traditional architectures still encounter difficulties in capturing long-range spatial-temporal interactions (Xu et al., 2025).

3 METHODS

3.1 DATASET OVERVIEW

3.1.1 OBJECTIVE

Our primary goal in creating the BasketVision dataset is to establish a rigorous and comprehensive framework for evaluating the visual understanding of Multimodal Large Language Models (MLLMs) within the specific context of basketball. The design is guided by several key objectives: to probe a wide spectrum of visual competencies, from foundational object recognition to sophisticated spatial and strategic analysis; to present models with authentic, high-complexity scenarios that mirror the dynamics and tactical depth of real games; to ensure that evaluation tasks are directly relevant to practical application requirements; and to establish a reproducible methodology for comparing model capabilities.

3.1.2 EVALUATION STRATEGY

It has been well-documented that probabilistic generative models often struggle with numerical precision, tending to produce plausible yet factually incorrect numerical outputs (Chelli et al., 2024), particularly for tasks demanding multi-step reasoning (Adel & Alani, 2025). To mitigate this known issue while retaining evaluative depth, BasketVision adopts a hybrid assessment strategy. For tasks requiring precise numerical answers, we employ a multiple-choice format, which constrains the output space and reduces the likelihood of numerical hallucination. For more nuanced tasks involving complex reasoning, we preserve an open-ended, natural language generation format. Every question-answer pair has been meticulously reviewed by human annotators to ensure ground-truth accuracy and evaluative authenticity (Maleki et al., 2024).

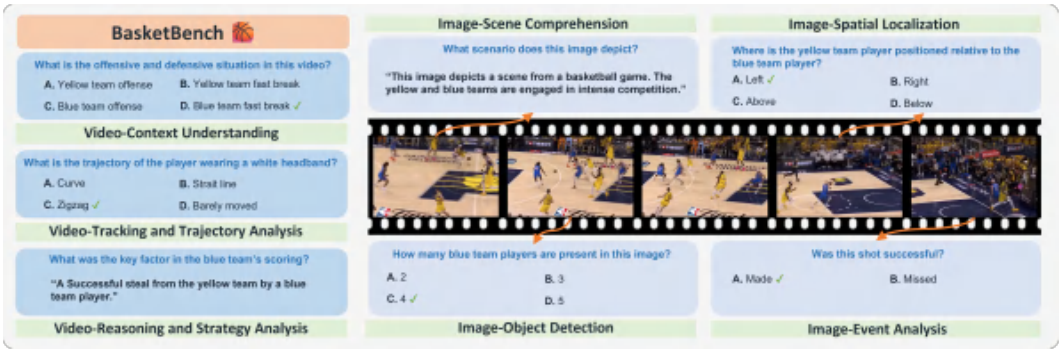

Figure 2: Exemplars illustrating the capability dimensions and corresponding sub-tasks within the BasketVision benchmark.

3.2 EVALUATION DIMENSIONS

As detailed in Table 1, BasketVision is structured around seven core capability dimensions shown in figure 2, each comprising three distinct sub-tasks. This hierarchical design ensures a thorough examination of model abilities, covering visual perception, spatial reasoning, temporal analysis, and higher-order game understanding across both static images and video sequences. Detailed descriptions of sub-tasks are available in Appendix A.

Table 1: The seven tasks in BasketVision with corresponding input and output formats.

Task Name	Input		Output	
	Image	Video	MC	SA
Scene Comprehension	✓			✓
Object Detection	✓		✓	
Spatial Localization	✓		✓	
Event Analysis	✓		✓	
Context Understanding		✓	✓	
Tracking and Trajectory Analysis		✓	✓	
Reasoning and Strategy Analysis		✓		✓

MC: Multiple-choice, SA: Short answer

3.3 Data Source and Curation

To ensure richness and diversity, we sourced game footage from three premier basketball competitions: the NBA 2024-25 Season Playoffs (84 games), the CBA 2024-25 Season Playoffs (37 games), and six Olympic Games from 2004 to 2024 (9 games). This multi-source strategy guarantees a wide array of playing styles, camera work, court configurations, and visual contexts.

A systematic quality control process was implemented to curate the raw footage. We discarded irrelevant clips and camera angles that offered minimal spatial information or were compromised by excessive motion blur. Following this curation, the dataset comprises 6,000 high-quality images and 1,200 video segments, amounting to over 2,300 minutes of gameplay, all maintaining a consistent visual standard suitable for precise spatial and temporal analysis.

3.4 Data Generation Pipeline

As shown in figure 1, the creation of our dataset is underpinned by an automated data generation pipeline composed of four primary stages. Each stage is tailored to extract specific information pertinent to our seven evaluation dimensions.

3.4.1 Court Recognition and Keypoint Detection

This initial stage automatically detects the boundaries and key geometric features of the basketball court from a given frame. (Details in Algorithm 1)

Algorithm 1 Court Recognition Algorithm

Input: Raw basketball game frame $I \in \mathbb{R}^{H \times W \times 3}$
Output: Court keypoints $P_{court} \in \mathbb{R}^{N \times 2}$, homography matrix $H \in \mathbb{R}^{3 \times 3}$
$I_{HSV} \leftarrow \text{HSV}(I)$ {Convert to HSV color space}
$(h_{lower}, h_{upper}) \leftarrow \text{HistogramAnalysis}(I_{HSV}[\frac{H}{3} :, :])$ {Detect court color range}
$M \leftarrow \text{inRange}(I_{HSV}, (h_{lower}, s_{min}, v_{min}), (h_{upper}, s_{max}, v_{max}))$ {Create binary mask}
$M \leftarrow \text{MorphologyClose}(\text{MorphologyOpen}(M, K_{3\times3}), K_{3\times3})$ {Refine mask}
$\mathcal{C} \leftarrow \text{findContours}(M)$ {Extract contours}
$C_{court} \leftarrow \arg\max_{c \in \mathcal{C}} \text{Area}(c)$ {Select largest contour}
$\mathcal{L} \leftarrow \text{HoughLinesP}(C_{court}, \rho = 1, \theta = \frac{\pi}{360}, \text{threshold} = 50)$ {Line detection}
$\mathcal{L}_{merged} \leftarrow \text{MergeLines}(\mathcal{L}, \theta_{tol} = \frac{\pi}{90})$ {clustering}
$P_{intersect} \leftarrow \text{FindIntersections}(\mathcal{L}_{merged})$ {Line intersection points}
$P_{court} \leftarrow \text{MatchToTemplate}(P_{intersect}, P_{template})$
$H \leftarrow \text{DLT}(P_{court}, P_{template})$ {Direct Linear Transform}

3.4.2 Court Projection and Perspective Transformation

This stage employs perspective transformation to map the 3D court view into a standardized 2D top-down representation. The mathematical basis is a homography matrix $H \in \mathbb{R}^{3\times3}$, which relates

corresponding points between the image plane and the target court plane:

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

where (x, y) denotes a point in the source image, and its corresponding point on the projected plane is (x_{proj}, y_{proj}) after normalization:

$$x_{proj} = \frac{x'}{w'}, \quad y_{proj} = \frac{y'}{w'} \tag{2}$$

The matrix H is computed using the Direct Linear Transform (DLT) algorithm, based on the court keypoints detected in the previous stage and their known locations on a standard court template. (Details in Algorithm 2)

Algorithm 2 Court Projection Algorithm

1: **Input:** Original frame $I \in \mathbb{R}^{H \times W \times 3}$, homography matrix $H \in \mathbb{R}^{3 \times 3}$, court config $C = (w_{court}, h_{court})$
2: **Output:** 2D projected court view $I_{2D} \in \mathbb{R}^{h_{court} \times w_{court} \times 3}$
3: $(w_{target}, h_{target}) \leftarrow (C.w, C.h)$ {Target court dimensions}
4: $I_{2D} \leftarrow$ warpPerspective$(I, H, (w_{target}, h_{target}))$ {Perspective transformation}
5: valid $\leftarrow$ ValidateGeometry(I_{2D}, H) {Geometric constraint validation}
6: Store H for tracking pipeline {Matrix persistence}

Figure 3: An example of automated court recognition and keypoint marking.

3.4.3 Player Detection and Re-identification

This stage is responsible for identifying and tracking individual players across frames. It combines pose estimation with color-based feature extraction for robust re-identification (ReID). (Details in Algorithm 3)

Algorithm 3 Player Detection and ReID Algorithm

Input: Frame $I \in \mathbb{R}^{H \times W \times 3}$, previous tracks $T_{prev} = \{t_i\}_{i=1}^{M}$
Output: Player detections $D = \{d_j\}_{j=1}^{N}$, updated tracks T_{new}
Pose Detection:
$\mathcal{K} \leftarrow$ YOLOPose(I) {Extract keypoints $\mathcal{K} = \{(x_k, y_k, c_k)\}_{k=1}^{17}$}
$\mathcal{B} \leftarrow$ DeriveBBoxes$(\mathcal{K})$ {Bounding boxes $\mathcal{B} = \{(x_1, y_1, x_2, y_2)_j\}_{j=1}^{N}$}
$P_{center} \leftarrow$ EstimateCenter$(\mathcal{K}_{ankle})$ {Center points from ankle midpoints}
Color Feature Extraction:
for each detection $d_j \in D$ **do**
    $R_j \leftarrow$ CropRegion$(I, \mathcal{B}_j)$ {Crop player region}
    $\mathcal{C}_j \leftarrow$ KMeans(HSV$(R_j), k = 2)$ {Dominant colors}
    $F_j \leftarrow$ ExtractFeatures$(\mathcal{C}_j)$ {HSV feature vector}
end for
ReID Tracking:
$\hat{P}_i \leftarrow$ KalmanPredict$(T_{prev}[i])$ {Predict track positions}
$C_{ij} \leftarrow \alpha \cdot d_{color}(F_i, F_j) + \beta \cdot d_{pos}(\hat{P}_i, P_j)$ {Cost matrix}
$\mathcal{M} \leftarrow$ Hungarian(C) {Optimal assignment}
$T_{new} \leftarrow$ UpdateTracks$(T_{prev}, D, \mathcal{M})$ {Update/create/prune tracks}

3.4.4 Player Position Calibration and Projection

The final stage projects the detected player positions onto the standard 2D court representation derived earlier. (Details in Algorithm 4)

Algorithm 4 Player Position Projection Algorithm

Input: Player tracks $T = \{t_i\}_{i=1}^{M}$, homography matrix $H \in \mathbb{R}^{3 \times 3}$
Output: Projected 2D player positions $P_{2D} = \{(x_{2D}, y_{2D})_i\}_{i=1}^{M}$
for each track $t_i \in T$ **do**
    $p_i \leftarrow t_i$.center_point {Extract center point (x_i, y_i)}
    $\begin{bmatrix} x_i' \\ y_i' \\ w_i' \end{bmatrix} \leftarrow H \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$ {Homography transformation}
    $p_{2D,i} \leftarrow (\frac{x_i'}{w_i'}, \frac{y_i'}{w_i'})$ {Normalize coordinates}
    valid$_i \leftarrow$ InBoundary$(p_{2D,i}, \mathcal{B}_{court})$ {Boundary validation}
    Store $(p_{2D,i}, \text{ID}_i, t)$ {Position with track ID and timestamp}
end for
$\mathcal{T} \leftarrow$ GenerateTrajectories(P_{2D}) {Trajectory data generation}

3.4.5 Pipeline Integration and Quality Control

The end-to-end pipeline integrates these four stages, operating on a per-frame basis with built-in validation checks to ensure data consistency. Following the automated processing, question-answer pairs are systematically generated using domain-specific templates. A final layer of manual validation by sports analytics professionals ensures the quality and correctness of the ground-truth data.

3.5 Dataset Statistics and Comparison

3.5.1 Dataset Statistics

The automated pipeline yields BasketVision, a dataset of 6,000 curated, multimodal questions in both Chinese and English, spanning both image-based and video-based reasoning tasks. Table 2 provides a detailed statistical breakdown. Given the continuous nature of professional sports, the dataset is designed for regular expansion, ensuring its long-term relevance and capacity to introduce novel challenges.

Figure 4: Examples of court projection from four Olympic Games (2004, 2008, 2020, 2024). Each pair displays the original 3D view (top) and the corresponding 2D projection (bottom).

3.5.2 Comparison with Existing Benchmarks

To situate our contribution, we compare BasketVision against prior benchmarks targeting visual understanding. While datasets such as EventHallusion, MM-Vet, and ViewSpatialBench have been pivotal in advancing MLLM evaluation, they are often circumscribed by modality (image- or video-only), domain (generic scenes), or task scope. As shown in Table 3, BasketVision offers a unique combination of attributes. It is the first to integrate both image and video modalities within a specialized, rule-governed sports context, targeting a broad suite of spatial, temporal, and event-centric reasoning tasks. This design facilitates a more rigorous diagnosis of model capabilities in scenarios demanding complex localization, multi-agent temporal understanding, and fine-grained action recognition. BasketVision thus fills demonstrable gaps in scale, modality, and domain specificity, establishing a new foundation for research into MLLMs for sports analytics and dynamic scene understanding.

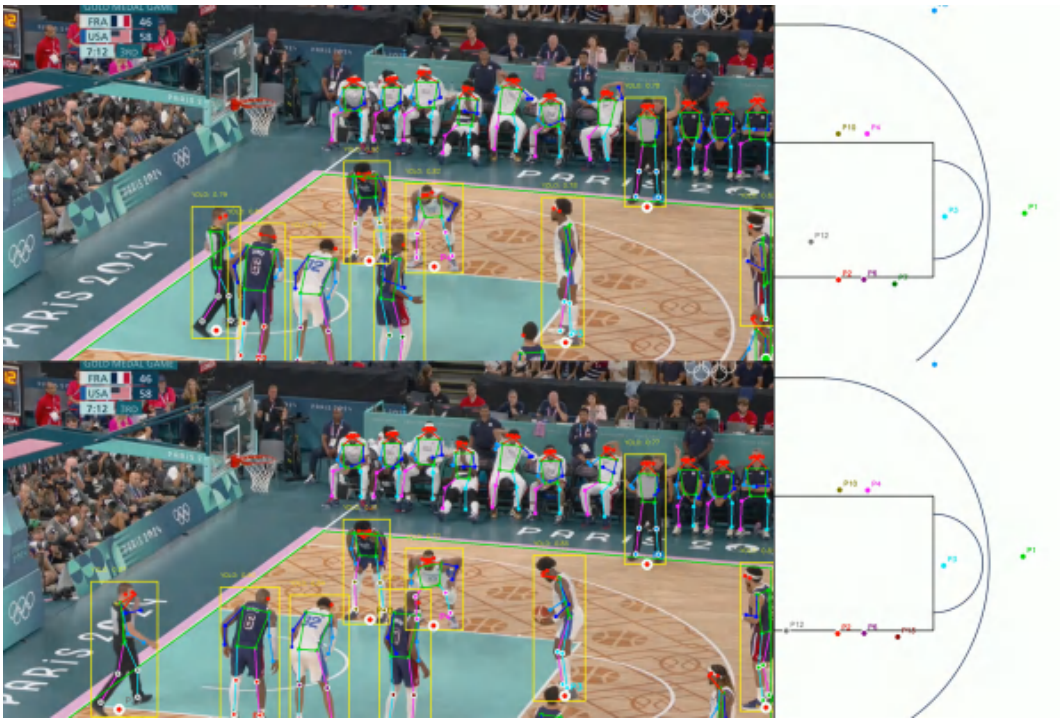

Figure 5: Examples of player re-identification and subsequent projection onto the 2D court view.

Table 2: Statistical breakdown of the BasketVision Dataset.

Category	Count	Percentage
Total Samples	6000	100%
By Task Type:		
Scene Comprehension	863	14.4%
Object Detection	954	15.9%
Spatial Localization	882	14.7%
Event Analysis	771	12.9%
Context Understanding	846	14.1%
Tracking & Trajectory Analysis	947	15.8%
Reasoning & Strategy Analysis	737	12.3%
By Language:		
Chinese	2000	33.34%
English	4000	66.67%
By Data Source:		
NBA 2024-25 Playoffs	3800	63.34%
CBA 2024-25 Playoffs	1700	28.34%
Olympic Games	500	8.34%

4 Experiments

4.1 Experimental Setup

Our experimental protocol was designed for a comprehensive evaluation of MLLM capabilities, incorporating a diverse suite of models, a standardized evaluation framework, and a carefully established human performance baseline.

Table 3: A comparison of BasketVision with other prominent visual understanding benchmarks.

Benchmark	Data Types	Task Types	Image	Video	QA	Lang.
BasketVision	**img+vid**	**7**	**2400**	**1200**	**6000**	**mul**
MM-Vet Yu et al. (2024)	images	6	200	-	218	mul
ViewSpatialBench Li et al. (2025)	images	5	5700+	-	5700+	EN
MMBench-Video Fang et al. (2024)	videos	26	-	609	1998	EN
MVBench Li et al. (2024)	videos	20	-	3641	4000	EN
VSI-Bench Yang et al. (2025)	videos	8	-	288	5000+	EN
Video-MME Fu et al. (2025)	videos	12	-	900	2700	mul
VidHal Choong et al. (2024)	videos	5	-	1000	3000	EN
EventHallusion Zhang et al. (2024)	videos	2	-	400	711	EN

We selected 23 prominent MLLMs for evaluation, including 16 proprietary and 7 open-source models. The proprietary group includes offerings from OpenAI (GPT-4o series (OpenAI, 2024)), Anthropic (Claude series), Google (Gemini series (Team, 2024)), Amazon (Nova series (AGI, 2025)), Baidu (ERNIE-4.5-VL), ByteDance (UI-TARS (Qin et al., 2025)), Perplexity (Sonar-Pro), and Minimax (minimax-01 (MiniMax, 2025)). The open-source cohort comprises models from Meta (Llama-4 series), Mistral AI (Mistral-Medium-3.1, Pixtral series (Agrawal et al., 2024)), and Qwen (Qwen2.5-VL-32B-Instruct (Bai et al., 2025)).

For models lacking native video processing capabilities, we employed a standardized frame sampling strategy, extracting one frame per second from each video clip. This sequence of key frames was provided as input, ensuring that all models received equivalent visual-temporal information and enabling an equitable comparison across architectures.

We employed accuracy as the primary evaluation metric. For multiple-choice questions, accuracy was computed directly. For open-ended questions requiring qualitative assessment (Scene Comprehension, Reasoning & Strategy Analysis), a more nuanced evaluation was conducted. Two trained annotators, with expertise in sports analytics, evaluated each generated response against a pre-defined rubric. This rubric assessed responses based on factual correctness, semantic completeness, and the presence of critical keywords derived from the ground-truth answer. A response was marked manually by these annotators following the rubric. In cases of disagreement, a third senior annotator made the final decision.

To contextualize model performance, we established a human baseline by recruiting 3 sports science professors and 13 undergraduate students. The professors served as our expert baseline, achieving an average accuracy of 96.34% across all tasks. The undergraduate cohort represented a non-expert, general population baseline. The participants were presented with the exact same questions and visual stimuli as the MLLMs, and their responses were graded using the same accuracy metrics to ensure a direct and fair comparison.

4.2 Results and Analysis

The complete results, presented in Table 4, reveal several critical insights into the current capabilities and limitations of MLLMs in the context of complex dynamic scenes. Scores in the table represent accuracy percentages, the highest score in each column is bolded.

4.2.1 A Persistent Gap Between Machine and Human Performance

Perhaps our most salient finding is the stark performance delta between human experts and all evaluated models. The expert baseline of 96.34% showcases robust human comprehension, whereas the top-performing MLLM, GPT-4o, caps out at 63.15%. This significant disparity, which echoes findings from other spatiotemporal reasoning benchmarks (Lin et al., 2025), suggests the chasm between correlational pattern matching and genuine causal understanding. Humans leverage rich "intuitive physics" and "theory of mind" to build predictive world models from limited data (Tenenbaum, 2018). In contrast, MLLMs, despite their scale, primarily learn statistical co-occurrences, excelling at generating plausible sequences without a deeper model of the world they describe (Bender et al.,

Table 4: Detailed performance of 23 evaluated models across the seven tasks in BasketVision.

Models	SC	OD	SL	EA	CU	TTA	RSA	Overall
Proprietary MLLM								
openai/gpt-4o	64.72%	56.88%	35.73%	**88.35%**	61.70%	61.49%	73.17%	**63.15%**
openai/gpt-4o-mini	63.18%	38.24%	27.63%	48.71%	54.90%	4.08%	71.50%	44.03%
anthropic/claude-3-opus-20240229	55.51%	49.06%	**53.86%**	47.65%	53.94%	16.61%	49.32%	46.56%
anthropic/claude-3.5-haiku	75.46%	6.32%	25.56%	27.81%	**97.96%**	21.51%	65.28%	45.70%
anthropic/claude-opus-4.1	81.60%	71.38%	22.81%	44.55%	67.31%	15.46%	**87.09%**	55.89%
anthropic/claude-sonnet-4	84.67%	67.41%	7.48%	40.39%	63.08%	14.82%	60.23%	48.30%
anthropic/claude-3-5-sonnet-20241022	83.13%	38.14%	41.22%	22.60%	50.39%	16.21%	73.55%	46.46%
google/gemini-2.5-flash	78.53%	46.11%	11.83%	29.19%	73.98%	49.10%	65.69%	50.63%
google/gemini-2.5-flash-lite	66.25%	28.85%	10.44%	48.61%	62.85%	19.18%	52.43%	41.23%
google/gemini-flash-1.5	53.97%	25.98%	41.12%	44.25%	50.29%	16.01%	46.90%	39.79%
amazon/nova-lite-v1	60.11%	29.33%	39.63%	22.50%	97.86%	2.56%	34.56%	40.94%
amazon/nova-pro-v1	73.93%	26.78%	30.21%	29.09%	28.04%	20.97%	44.29%	36.19%
baidu/ernie-4.5-vl-424b-a47b	61.65%	46.08%	51.72%	22.40%	62.78%	17.81%	30.44%	41.84%
bytedance/ui-tars-1.5-7b	**87.74%**	45.47%	13.47%	28.21%	62.68%	15.71%	43.63%	42.42%
perplexity/sonar-pro	72.39%	56.95%	16.24%	48.41%	97.52%	49.00%	79.08%	59.94%
minimax/minimax-01	69.32%	**81.97%**	29.14%	76.85%	62.56%	70.77%	44.94%	62.22%
Generative Vision Model								
google/gemini-2.5-flash-image-preview	58.72%	48.39%	24.30%	76.81%	92.23%	67.94%	53.15%	60.22%
Open-source MLLM								
meta-llama/llama-4-maverick	58.58%	26.68%	10.04%	28.18%	27.94%	**74.67%**	62.00%	41.16%
meta-llama/llama-4-scout	57.04%	45.37%	5.32%	22.30%	26.30%	15.61%	54.61%	32.36%
mistralai/mistral-medium-3.1	67.79%	78.73%	9.24%	48.31%	50.19%	72.85%	70.24%	56.76%
mistralai/pixtral-12b	80.06%	72.32%	28.95%	28.01%	26.20%	12.43%	9.21%	36.74%
mistralai/pixtral-large-2411	70.86%	38.34%	25.16%	45.67%	50.09%	63.47%	39.98%	47.65%
qwen/qwen2.5-vl-32b-instruct	77.00%	76.53%	22.75%	27.94%	62.46%	65.44%	42.50%	53.52%
Human								
Human Expert Avg.	93.78%	99.50%	96.53%	98.86%	96.85%	99.14%	89.75%	96.34%

(All values in %). **SC**: Scene Comprehension, **OD**: Object Detection, **SL**: Spatial Localization, **EA**: Event Analysis, **CU**: Context Understanding, **TTA**: Tracking & Trajectory Analysis, **RSA**: Reasoning & Strategy Analysis.

2021). Their failure to consistently reason about the game state implies they are not yet simulating the underlying dynamics.

4.2.2 Proprietary Models Exhibit a Performance Advantage

A clear performance hierarchy between proprietary and open-source models is also evident. The leading proprietary systems consistently outperform their open-source counterparts. This gap is likely attributable not just to model scale, but to a compounding advantage in data curation and alignment techniques. Proprietary models benefit from a ”data flywheel,” where massive-scale,

high-quality proprietary data and extensive human feedback from Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022) creates a significant performance moat. This vast and diverse data allows them to achieve better generalization and alignment than models trained on publicly available, and often noisier, datasets.

4.2.3 Pronounced Task Specialization Over General Competence

Rather than exhibiting broad competence, models display pronounced patterns of specialization. For instance, Minimax-01 excels at Object Detection (81.97%) but falters elsewhere. This phenomenon may be explained by the well-documented problem of "negative transfer" in multitask learning, where the optimization signals from disparate tasks interfere with each other. Research has shown that when a single set of weights is optimized, the gradients for different tasks can be conflicting (Yu et al., 2020). Optimizing for perceptual acuity and abstract reasoning simultaneously can thus lead to a "seesaw effect", where updates that improve one capability harm another, preventing the model from achieving holistic competence.

4.2.4 Advanced Reasoning Capabilities on Specific Tasks

Notably, this specialization extends to high-level reasoning. Claude-Opus-4.1 achieves the highest score on Task 7, Reasoning & Strategy Analysis (87.09%), significantly outperforming many competitors on tasks requiring tactical and causal inference. This likely stems from the powerful Language Model (LM) backbone, where abstract reasoning has been shown to be an emergent capability of scale (Wei et al., 2022). However, this excellence is isolated; the same model performs poorly on Spatial Localization (22.81%). This demonstrates a critical failure in "modality grounding"—the model's powerful abstract reasoning module is not effectively conditioned on the precise geometric information from the visual encoder. The abstract concepts remain detached from their physical manifestation in the scene.

4.2.5 A Trade-off Between Perceptual Acuity and General Reasoning

The performance trade-off between google/gemini-2.5-flash and its vision-optimized variant (-image-preview) perfectly illustrates the specialization-generalization dilemma. The specialized model's gains on perceptual tasks (e.g., Event Analysis, +47.62 pp) come at a direct cost to its general reasoning abilities (e.g., Scene Comprehension, -19.81 pp). This aligns with the concept of the "representation degeneration problem," where fine-tuning on a narrow task distribution can cause the model's internal representations to become less rich and expressive, collapsing into a lower-dimensional space that is optimal for the fine-tuning task but detrimental to general capabilities (Gao et al., 2019).

4.2.6 Spatial Reasoning as a Systemic Bottleneck

Across all models, Spatial Localization emerges as the most significant bottleneck, with the top score being a mere 53.86%. This systemic weakness, identified as a core challenge for current MLLMs in recent systematic analyses (Zhang et al., 2025), can be traced to fundamental architectural limitations. Most MLLMs employ Vision Transformers (ViTs) that process images as a "bag of patches" (Dosovitskiy et al., 2021). This architecture excels at identifying objects *within* patches but is not explicitly designed to encode the precise metric and topological relationships *between* them. As other research has demonstrated, this leads to significant performance cliffs when moving from 2D to 3D reasoning and a tendency to default to formulaic derivations rather than true spatial visualization (Wang et al., 2025). The cross-attention mechanism, while effective for coarse semantic alignment between text and image regions, lacks the strong inductive bias required for fine-grained geometric reasoning, leading to consistent failures on tasks demanding a precise coordinate system. This fundamental architectural limitation explains why even models with powerful abstract reasoning capabilities fail to "ground" their logic in the physical layout of the scene.

4.2.7 Diminishing Returns of Model Scaling

Finally, while scale correlates positively with performance, the persistence of fundamental flaws, particularly in spatial reasoning, suggests that simply scaling current architectures may yield diminishing returns. Current scaling laws largely predict performance on tasks that are well-represented

in the web-scale pre-training data (Kaplan et al., 2020). However, as recent studies suggest, performance on specialized spatial tasks can plateau at a relatively low upper bound, even as training data increases (Zhang et al., 2025). Abilities like precise spatio-temporal grounding may be "out-of-distribution" for the standard next-token prediction objective. Overcoming these deep-seated limitations will likely require more than just additional parameters; it will necessitate novel architectural inductive biases and training paradigms explicitly designed to foster robust spatio-temporal and causal world models.

4.3 Limitations

While BasketVision provides a rigorous framework for evaluating MLLMs on complex dynamic systems, we acknowledge several limitations that offer avenues for future work.

Domain Specificity. Our benchmark is intentionally focused on professional basketball to create a controlled yet complex environment. While this specialization is a key strength, the findings may not directly generalize to other dynamic systems with different rules, physics, or agent behaviors, such as autonomous driving or crowd simulation. Future work could adapt the BasketVision pipeline to other sports or structured domains.

Viewpoint and Data Bias. The dataset is curated from professional sports broadcasts, which predominantly feature a limited set of standardized camera angles. The performance of models may vary significantly when faced with different viewpoints, such as player-worn cameras or amateur footage, which are not represented in our current dataset.

5 Conclusion

In this work, we introduced BasketVision, a novel benchmark designed to move beyond conventional visual question answering and probe the ability of MLLMs to comprehend complex dynamic systems. By leveraging professional basketball as a structured, multi-agent microcosm, we offer a rigorous testbed for the nuanced capabilities required for real-world scene understanding: precise spatial localization, fine-grained temporal analysis, and high-level strategic reasoning.

Our extensive evaluation of 23 models reveals a sobering reality: a formidable 33-point performance gap separates today's most advanced MLLMs from human expert performance. The results systematically pinpoint spatial reasoning as a fundamental bottleneck across all models and uncover highly specialized performance profiles, where models excel in some areas while failing at others. This suggests that a truly general and robust understanding of dynamic environments remains an elusive goal.

BasketVision provides the community with a valuable tool to diagnose these deficiencies and guide subsequent research. The path forward may not lie in simply scaling existing architectures, but in developing new paradigms with explicit inductive biases for spatio-temporal grounding and causal inference. By providing a challenging and reproducible evaluation framework, we hope to accelerate progress toward models that can perceive and reason about the complex, dynamic world in which they are intended to operate.

Acknowledgments

The authors would like to thank the reviewers for their valuable feedback. This work was supported in part by [funding information, if applicable].

References

A. Adel and N. Alani. Can generative ai reliably synthesise literature? exploring hallucination issues in chatgpt. *AI & Society*, 2025. doi: 10.1007/s00146-025-02406-7. URL https://doi.org/10.1007/s00146-025-02406-7.

Amazon AGI. The amazon nova family of models: Technical report and model card, 2025. URL https://arxiv.org/abs/2506.12103.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, et al. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, et al. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2425–2433. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.279. URL https://doi.org/10.1109/ICCV.2015.279.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Madeleine Clare Elish, William Isaac, and Richard S. Zemel (eds.), *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pp. 610–623. ACM, 2021. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, et al. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis. *J Med Internet Res*, 26:e53164, May 2024. ISSN 1438-8871. doi: 10.2196/53164. URL https://www.jmir.org/2024/1/e53164.

Wey Yeh Choong, Yangyang Guo, and Mohan S. Kankanhalli. Vidhal: Benchmarking temporal hallucinations in vision llms. *CoRR*, abs/2411.16771, 2024. doi: 10.48550/ARXIV.2411.16771. URL https://doi.org/10.48550/arXiv.2411.16771.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, et al. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a2326c9715a516c91174132e0170073a-Abstract-Datasets_and_Benchmarks_Track.html.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. doi: 10.48550/ARXIV.2306.13394. URL https://doi.org/10.48550/arXiv.2306.13394.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24108–24118. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Jun Gao, Di He, Xu Tan, Tao Qin, et al. Representation degeneration problem in training natural language generation models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SkEYojRqtm.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, et al. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *CoRR*, abs/2505.21500, 2025. doi: 10.48550/ARXIV.2505.21500. URL https://doi.org/10.48550/arXiv.2505.21500.

Jian Li and Weiheng Lu. A survey on benchmarks of multimodal large language models. *CoRR*, abs/2408.08632, 2024. doi: 10.48550/ARXIV.2408.08632. URL https://doi.org/10.48550/arXiv.2408.08632.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 22195–22206. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02095. URL https://doi.org/10.1109/CVPR52733.2024.02095.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, et al. Ost-bench: Evaluating the capabilities of mllms in online spatio-temporal scene understanding. *CoRR*, abs/2507.07984, 2025. doi: 10.48550/ARXIV.2507.07984. URL https://doi.org/10.48550/arXiv.2507.07984.

Siqi Liu, Guy Lever, Zhe Wang, Josh Merel, et al. From motor control to team play in simulated humanoid football. *Sci. Robotics*, 7(69), 2022. doi: 10.1126/SCIROBOTICS.ABO0235. URL https://doi.org/10.1126/scirobotics.abo0235.

Negar Maleki, Balaji Padmanabhan, and Kaushik Dutta. Ai hallucinations: A misnomer worth clarifying. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 133–138, Singapore, Singapore, 2024. doi: 10.1109/CAI59869.2024.00033.

MiniMax. Minimax-01: Scaling foundation models with lightning attention, 2025. URL https://arxiv.org/abs/2501.08313.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, et al. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *CoRR*, abs/2311.16103, 2023. doi: 10.48550/ARXIV.2311.16103. URL https://doi.org/10.48550/arXiv.2311.16103.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, et al. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vis.*, 123(1):74–93, 2017. doi: 10.1007/S11263-016-0965-7. URL https://doi.org/10.1007/s11263-016-0965-7.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, et al. Ui-tars: Pioneering automated gui interaction with native agents, 2025. URL https://arxiv.org/abs/2501.12326.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Josh Tenenbaum. Building machines that learn and think like people. In Elisabeth André, Sven Koenig, Mehdi Dastani, and Gita Sukthankar (eds.), *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, pp. 5. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018. URL http://dl.acm.org/citation.cfm?id=3237389.

Siting Wang, Luoyang Sun, Cheng Deng, Kun Shao, et al. Spatialviz-bench: Automatically generated spatial visualization reasoning tasks for mllms. *CoRR*, abs/2507.07610, 2025. doi: 10.48550/ARXIV.2507.07610. URL https://doi.org/10.48550/arXiv.2507.07610.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *CoRR*, abs/2406.16338, 2024. doi: 10.48550/ARXIV.2406.16338. URL https://doi.org/10.48550/arXiv.2406.16338.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, et al. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

Lingrui Xu, Mandi Liu, and Lei Zhang. Tacticexpert: Spatial-temporal graph language model for basketball tactics. *CoRR*, abs/2503.10722, 2025. doi: 10.48550/ARXIV.2503.10722. URL https://doi.org/10.48550/arXiv.2503.10722.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, et al. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 10632–10643. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00994. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, et al. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023. doi: 10.48550/ARXIV.2306.13549. URL https://doi.org/10.48550/arXiv.2306.13549.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, et al. Gradient surgery for multi-task learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, et al. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KOTutrSR2y.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *CoRR*, abs/2409.16597, 2024. doi: 10.48550/ARXIV.2409.16597. URL https://doi.org/10.48550/arXiv.2409.16597.

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, et al. Why do mllms struggle with spatial understanding? A systematic analysis from data to architecture. *CoRR*, abs/2509.02359, 2025. doi: 10.48550/ARXIV.2509.02359. URL https://doi.org/10.48550/arXiv.2509.02359.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, et al. VLM4D: towards spatiotemporal awareness in vision language models. *CoRR*, abs/2508.02095, 2025. doi: 10.48550/ARXIV.2508.02095. URL https://doi.org/10.48550/arXiv.2508.02095.

A Detailed Task and Sub-task Descriptions

This appendix provides detailed descriptions for the seven core tasks evaluated in the BasketVision benchmark.

- **Scene Comprehension:**

  - Global scene description
  - Court zone and context identification
  - Crowd and bench detection
- **Object Detection:**
  - Player count estimation
  - Player identity recognition
  - Basketball detection
- **Spatial Localization:**
  - Absolute player localization
  - Relative player positioning
  - Ball-player relationship assessment
- **Event Analysis:**
  - Situation detection
  - Shot recognition
  - Action recognition
- **Context Understanding:**
  - Context-aware video question answering
  - Key moment summarization
  - Temporal ordering verification
- **Tracking and Trajectory Analysis:**
  - Individual player trajectory reconstruction
  - Ball movement tracking
  - Team positional heatmap generation
- **Reasoning and Strategy Analysis:**
  - Team tactic/pattern recognition
  - Causal inference in events
  - Outcome prediction