MVPBench: A Multi-Video Perception Evaluation Benchmark for Multi-Modal Video Understanding

Purui Bai[1], Tao Wu[2], Jiayang Sun[1], Xinyue Liu[1], Huaibo Huang[1], Ran He[1*]
[1]MAIS & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]SIST, ShanghaiTech University, Shanghai, China
Email: {purui.bai, huaibo.huang, ran.he}@cripac.ia.ac.cn, wutao2022@shanghaitech.edu.cn, {sunjiayang2025, liuxinyue2025}@ia.ac.cn

***Abstract*—The rapid progress of Large Language Models (LLMs) has spurred growing interest in Multi-modal LLMs (MLLMs) and motivated the development of benchmarks to evaluate their perceptual and comprehension abilities. Existing benchmarks, however, are limited to static images or single videos, overlooking the complex interactions across multiple videos. To address this gap, we introduce the Multi-Video Perception Evaluation Benchmark (MVPBench), a new benchmark featuring 14 subtasks across diverse visual domains designed to evaluate models on extracting relevant information from video sequences to make informed decisions. MVPBench includes 5K question-answering tests involving 2.7K video clips sourced from existing datasets and manually annotated clips. Extensive evaluations reveal that current models struggle to process multi-video inputs effectively, underscoring substantial limitations in their multi-video comprehension. We anticipate MVPBench will drive advancements in multi-video perception. The evaluation code and dataset will be available at *https://github.com/MVPBench/MVPBench*.**

***Index Terms*—Large Language Models, Interpretable and Explainable AI, Representation & Reasoning, Generative AI Models, Neural Network Applications, Perceptual Neural Networks**

I. Introduction

With the advancement of Large Language Models (LLMs) [1], [2], there is a growing interest in the application within the domain of visual modalities, further leading to the development of Multi-modal Large Language Models (MLLMs) [3], [4]. As multi-modal application technologies have progressed, corresponding evaluation benchmarks have also been developed, which evolved from single images [5], [6], to multiple images [7], [8], and eventually to video modalities [7], [9], [10]. However, a significant gap exists in current MLLMs' evaluation benchmarks, in that they do not assess the models' capabilities in handling multiple dynamic video inputs. As illustrated in Fig. 1, these evaluation frameworks fall short in measuring the processing abilities of models with real-world scenarios input.

To address the aforementioned challenges, we propose the Multi-Video Perception Evaluation Benchmark (MVPBench) for Multi-modal Video Understanding. This benchmark provides a comprehensive assessment of MLLMs' capabilities in processing and understanding multi-video inputs. To the best of our knowledge, MVPBench represents the first benchmark in the MLLM research domain explicitly designed to evaluate multi-video comprehension.

MVPBench includes a diverse set of **14** subtasks that evaluate the model's capabilities across various dimensions. These tasks range from basic to advanced levels, covering a variety of question-answering formats from low-level pattern recognition to high-level semantic interpretation, thereby imposing rigorous demands on model performance across perceptual and cognitive dimensions. During the design process, these tasks adhere to the design principles outlined in Fig. 2, which emphasize the consideration of both the multiplicity of evaluation inputs and the temporal aspects of evaluation videos. As benchmark designers, our core principle is to enable a comprehensive evaluation of MLLMs. Accordingly, our framework is designed to encompass both multi-video processing capabilities and a set of complementary tasks that do not strictly require multiple video inputs, the task effectiveness of which has been demonstrated through repeated validation by prior works.

To validate the rationality of MVPBench, we conducted evaluations on **15** state-of-the-art MLLMs of different scales. Results indicate that even the most advanced models achieve an average accuracy of only **31.10%**, significantly below human performance at **88.89%**. The variation in model performance across different task types highlights MVPBench's capability in pinpointing the limitations of current models in perception tasks. We expect MVPBench to drive the development of MLLMs with enhanced multi-video comprehension capabilities. The complete evaluation code and dataset will be available upon acceptance.

II. Related Work

Video MLLM. In recent years, researchers have expanded visual modalities from static images to dynamic videos, employing innovative technologies such as VideoChat [18], and Valley [19] to exploit the substantial potential of LLMs in video understanding tasks. Recent studies, including Qwen2.5-VL [20] and LLaVA-Video [21], have elevated MLLMs'

This work was supported by National Natural Science Foundation of China (Grant Nos. 62576342, 62425606, 62550062, 32341009), Beijing Natural Science Foundation (4252054, L257008), Beijing Nova Program (20230484276, 20240484601).
* Corresponding author.

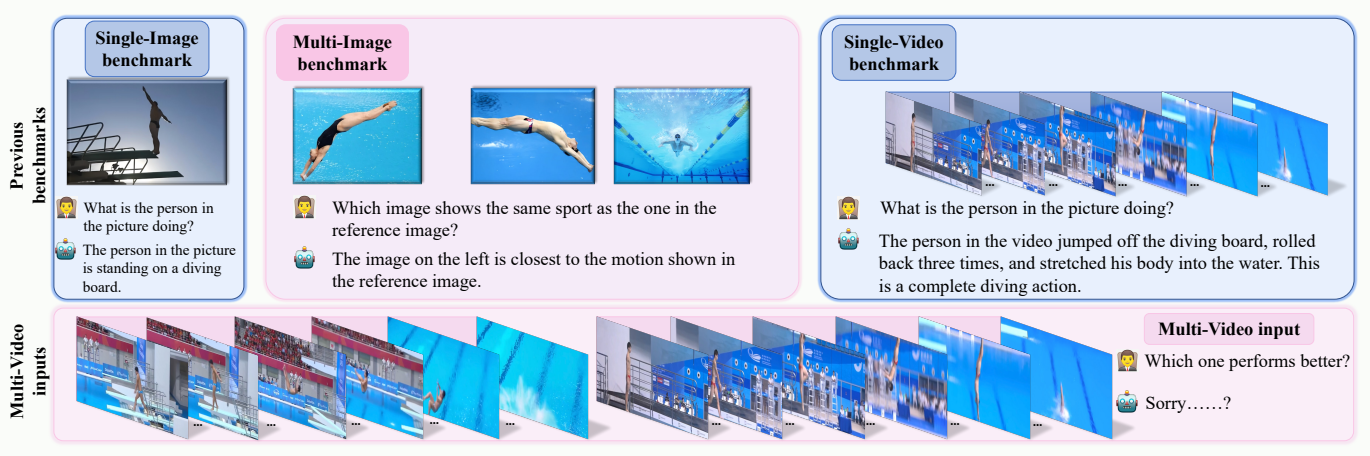

Fig. 1: **Challenges in managing Multi-Video inputs.** Previous benchmarks are primarily designed to address straightforward visual text problems, such as describing the content and associations of static images. While video description provides additional information to enhance the detail of content descriptions, it remains challenging to effectively manage comparisons between multiple related videos.

TABLE I: **Prompt examples of MVPBench.**

Subdomain	Subtask	Source	Prompt
Temporal Segment	Tem w/ Cap	YouCook2 [11]	*Sort these video clips according to the recipe's textual description. Recipe Description: step0: add water to a pan. step1: stir in the pack of sauce. step2: stir and cook until the soup boils*
Splicing	Tem w/o Cap	Manual Collection	*Sort the given video clips. According to their content or logical sequence.*
Content Assessment	Daily life action evaluation	EPIC-Skills [12]	*Your task is to compare these two clips. Determine which one has better task and action completion.*
	Professional action evaluation	AQA-7 [13]	*Your task is to compare these two clips. Determine which one has better task and action completion.*
	Weather condition evaluation	Manual Collection	*Your task is to examine the weather conditions in each video. Determine which video shows the most severe weather.*
Video Quality Assessment	Video clarity evaluation	CVQAD [14]	*Compare these videos based on their clarity and smoothness (video quality). Rank them from highest to lowest quality.*
	Video brightness evaluation	SDSD [15]	*Identify the video clip among the three given clips. With the lowest brightness and the darkest shooting conditions.*
	Video forensic detection	Manual Collection	*Compare the three video clips. Determine which one is the real video.*
Video Logic Inference	Multiview visual perception pairing	PEV [16]	*The reference video and the correct candidate video are the corresponding perspectives. Identify which of the three candidate videos corresponds to the reference video.*
	Common sense judgment of physical laws	Manual Collection	*Please compare the following three video clips. Identify which one adheres to normal physical laws.*
Similar Video Pairing	Gait recognition matching	CASIA Gait Database [17]	*The reference video and the correct candidate video depict the same person walking. Determine which of the two candidate videos matches the reference video.*
	Cinematographic style matching	Manual Collection	*The reference video and the correct candidate video share the same filming style. Choose the video that matches the reference video.*
	Olympic sports matching	Manual Collection	*The reference video and the correct candidate video depict the same sports activity. Compare the two candidate videos and select the one that matches the reference video.*
	Dance style matching	Manual Collection	*The reference video and the correct candidate video depict the same dance type. Compare the two candidate videos and select the one that matches the reference video.*

capabilities in the video modality through extensive video-text training data pairs.

However, the majority of training data for the video modality remain limited to single video-text pairs, restricting their ability to process multiple videos simultaneously. Currently, there are no established evaluation standards for MLLM's multi-video understanding capabilities in the research community.

Video understanding benchmark. Recent advancements in video understanding benchmarks have expanded the evaluation of Video-Language Models (VLMs) across diverse domains and temporal scales. Early benchmarks [22], [23] focused on general video-language understanding, while more recent benchmarks address specialized capabilities, such as MVBench [9] and Video-MME [10], featuring videos with multimodal elements (e.g., *subtitles, audio*), EgoSchema [24] and LongVideoBench [25] evaluate event- or story-level understanding across extended temporal horizons.

Notably, VideoVista [26] incorporates a multi-video comprehension category, namely the Video-Video Relation Rea-

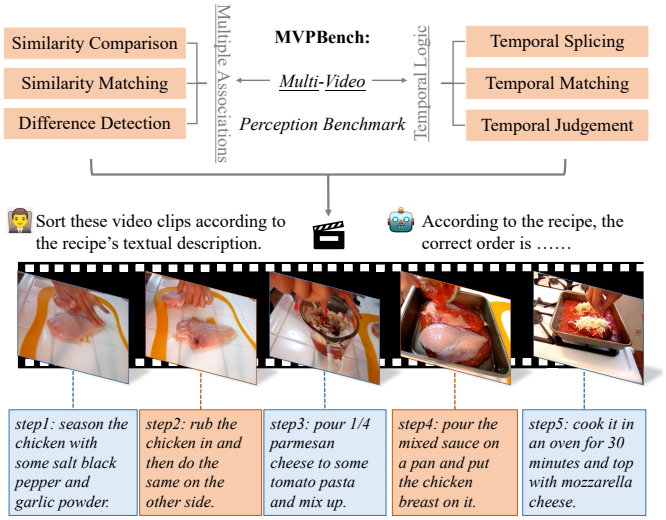

Fig. 2: **Design principles of MVPBench.** A key feature of MVPBench is its ability to consider both the multiplicity of evaluation inputs and the temporal characteristics of evaluation videos.

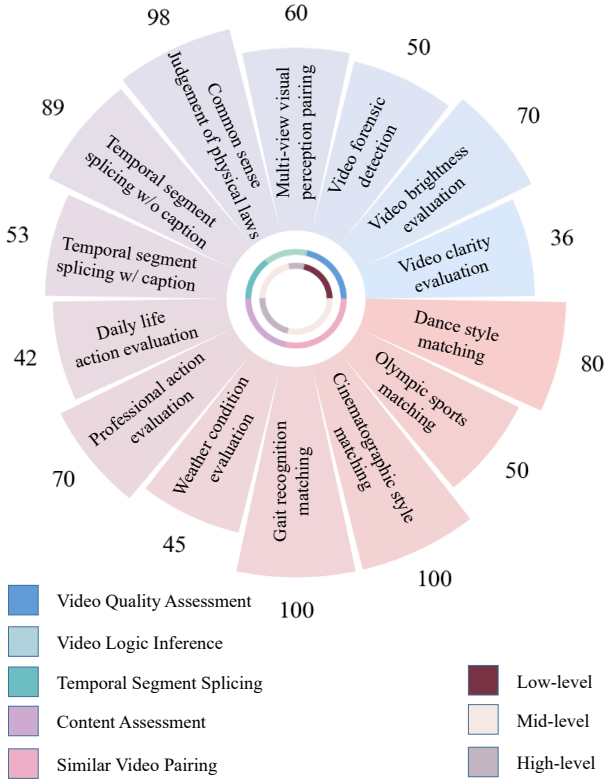

Fig. 3: **Statistics of MVPBench.** The benchmark includes 14 tasks in 5 domains, ranging from low-level pattern comparison to mid-level temporal logic reasoning, and extending to high-level visual content understanding.

soning mentioned in their work. However, VideoVista is fundamentally still a conventional benchmark primarily focused on single-video understanding and reasoning. In contrast, our proposed MVPBench is a specialized, holistic, and rigorously designed benchmark which aims to evaluate the diverse perception and reasoning capabilities of MLLMs across multiple video inputs by encompassing various task areas that require joint comprehension and extended reasoning across multiple videos.

III. MVPBench

MVPBench encompasses a range of visual tasks across different dimensions and levels, with dataset partly derived from existing datasets related to the task, supplemented by a manually curated substantial collection of high-quality video materials. As shown in Tab. I, we standardized the prompt design and processing workflow to ensure clear task definitions and to mitigate the influence of extraneous factors such as ambiguous textual information. In total, we contributed 5K visual question answering questions and 2.7K videos. Fig. 3 illustrates the distribution of each task.

A. Dataset Collection Process

Caption-guided Temporal Segment Splicing. Assesses MLLMs' ability to reconstruct shuffled video sequences leveraging text-visual alignment. Using YouCook2 [11], we segment videos into recipe-step clips. Given full textual step sequences, models reorder clips based on temporal ordering prompts.

Caption-free Temporal Segment Splicing. Evaluates MLLMs' ability to reconstruct shuffled video sequences using visual cues alone. We curated videos with inherent temporal logic (e.g., plant growth), segmented them into randomized clips, and tasked models with reordering via temporal reasoning.

Daily Life Action Evaluation. Assesses MLLMs' ability to comparatively evaluate action quality in similar tasks. Using EPIC-Skills [12] videos with quality annotations, we reorganize clips into subtask pairs. Models identify the clip demonstrating superior execution quality through pairwise comparison.

Professional Action Evaluation. Assesses MLLMs' capability for pairwise quality comparison of professional actions. Using AQA-7 [13], we reorganize clips into pairs with significant referee-scored quality differences, requiring binary selection of the superior-quality video.

Weather Condition Evaluation. Assesses MLLMs' ability to classify weather intensity through temporal dynamics and physical cues. We curated "rainy", "snowy", and "windy" videos at three intensity levels. Models perform comparative analysis on triplets per weather type to identify specified conditions.

Video Clarity Evaluation. Assesses MLLMs' capability to evaluate perceptual video quality incorporating temporal artifacts from encoding/bitrates. Using CVQAD [14], we curate triplet sequences with shuffled quality levels (high/medium/low) per content type. Models analyze temporal

TABLE II: **The comparison of various benchmarks:** whether the benchmark incorporates videos sourced from diverse open domains (Open.), whether the benchmark incorporates egocentric videos(Ego.), whether the benchmark incorporates complex tasks spanning multiple hierarchical levels of perception(Com.), whether the benchmark incorporates multi-video comprehension category (Multi.) and whether the benchmark incorporates videos synthesized by the generative model(Gen.)

Benchmarks	#Videos	#QA Pairs	Open.	Ego.	Com.	Multi.	Gen.	Data source
MVBench [9]	3,641	4,000	✓	✓	✓	✗	✗	existing datasets
Video-Bench [27]	5,917	17,036	✓	✗	✓	✗	✗	existing datasets
EgoSchema [24]	5,063	5,063	✗	✓	✗	✗	✗	ego-centric video
AutoEval-Video [28]	327	327	✓	✗	✗	✗	✗	Youtube
TempCompass [29]	410	7,540	✓	✗	✓	✗	✗	ShutterStock
Video-MME [10]	900	2,700	✓	✗	✓	✗	✗	Youtube
LongVideoBench [25]	3,763	6,678	✓	✗	✓	✗	✗	web channels
VideoVista [26]	894	24,906	✓	✗	✓	✓	✗	YouTube
MVPBench	2,774	5,050	✓	✓	✓	✓	✓	web videos, synthetic videos, datasets

degradation patterns to restore original quality rankings.

Video Brightness Evaluation. Assesses MLLMs' capability to evaluate scene illumination conditions. Using SDSD [15] with controlled lighting variations, we generate triplet queries (1 dark + 2 bright or 2 dark + 1 bright) per trial. Models identify videos matching specified lighting conditions through multiple-choice selection.

Video Forensic Detection. Evaluates MLLMs' ability to distinguish authentic versus AI-generated videos. Real videos exhibiting physical plausibility are paired with synthetic counterparts: captions generated via GPT-4o [30] drive video synthesis using Kling [31] and Ying [32]. Models identify authentic videos by detecting spatiotemporal inconsistencies through physical law reasoning.

Multiview Visual Perception Pairing. Assesses MLLMs' capability for viewpoint-invariant object matching. Using paired clips from [16], we present a reference video alongside three candidates (one correct match, two distractors). Models identify matching clips by analyzing cross-view correspondence through global object feature detection.

Common sense judgment of physical laws. Probes MLLMs' understanding of physical causality in temporal dynamics. We curated natural phenomena videos (e.g., glacier melting, sand flow) with physically implausible reversed versions. Models discern plausible temporal logic from triplets containing one normal sequence and two inverted clips.

Gait Recognition Matching. Evaluates MLLMs' cross-condition subject matching using CASIA Gait Database [17]. Given a reference gait sequence, models identify the same subject in candidate videos under varying conditions. Subjects with similar visual characteristics are deliberately paired to minimize confounding factors.

Cinematographic Style Matching. Evaluates MLLMs' recognition of temporal editing techniques beyond static frames. We curated styles with distinctive temporal signatures (e.g., slow motion, time-lapse, frame extraction). Given a reference video and three candidates, models identify the optimal stylistic match through temporal pattern analysis.

Olympic Sports Matching. Evaluates MLLMs' fine-grained motion discrimination within similar action categories. We curated sports sequences with high kinematic similarity (e.g., swimming strokes: butterfly, freestyle, breaststroke, backstroke), which static frames alone cannot reliably distinguish. Given a reference video and two same-category candidates, models discern subtle temporal variations.

Dance Style Matching. Evaluates MLLMs' temporal pattern recognition across dance styles. We curated diverse styles with distinctive motion signatures (e.g., street, Latin, ballet), challenging to assess via static frames. Given a reference video and three candidates, models identify optimal stylistic matches through temporal dynamics analysis.

IV. Experiments

A. Experimental Setup

Evaluation setup: To mitigate bias, we standardize prompts to explicitly enumerate video roles as noted below. This design ensures equitable evaluation across heterogeneous model capabilities:

$$Prompt + Video1 : <video> + \ldots + VideoN : <video> + Prompt$$

In the prompt template, <video>serves as a placeholder for videos, with the model reserving these positions for subsequent input video visual tokens. To differentiate between various input videos, distinct video encodings are added before the placeholder. While the required form of model output varies across different tasks, we consistently require the model to produce strictly numerical outputs. This facilitates a straightforward horizontal comparison of the results.

We employ a set of predefined rules along with GPT-4-turbo [1] to extract the selected answer from the model's output. We perform manual verification on 15% of outputs with 98% MLLM-Human Agreement, validating the robustness of our automated extraction pipeline.

Human baseline: We adopted a stratified sampling approach to establish reliable human performance benchmarks for MVP-Bench, with 52 ± 3 questions per evaluator(20 evaluators in total). We calculated the inter-evaluator agreement (89.7%) to verify the rationality of the human baseline. Each evaluator

Fig. 4: **Task introduction of MVPBench**. Zoom in for better view.

received 10% randomly interspersed duplicate questions (104 total) from other evaluators' sets. For each duplicate pair (q_i, q_j) answered by evaluators (E_m, E_n), agreement was scored as:

$$\text{Agreement} = \frac{\sum_{k=1}^{N} \mathbb{I}(A_{m,k} = A_{n,k})}{N} \times 100\%, \qquad (1)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, N is the number of duplicate questions for consistency, and $A_{m,k}$ represents evaluator m's response to the k-th question.

Standardized Performance Assessment: To ensure a statistically rigorous framework for comparing model capabilities across heterogeneous subtasks with divergent chance performance levels, we adopt a standardized evaluation protocol grounded in information-theoretic principles. Let $\mathfrak{C}_t$ denote the *chance performance* for subtask t, defined as $\mathfrak{C}_t = 1/k_t$, where k_t represents the number of response options. For a model's observed accuracy A_{mt} on subtask t, we compute its

TABLE III: **Results of different models on MVPBench.** The best performance in each task is **in-bold**. The "Overall" metric denotes a weighted average across models, with weights proportional to the information gain potential of each respective task.

Model	Tem w/ Cap	Tem w/o Cap	Daily-Action	Pro-Action	Weather	Clarity	Brightness	Forensic
Human	95.16	80.01	80.96	85.72	83.34	90.00	95.71	97.00
Closed-source Multimodal LLMs								
Gemini-2.5-Flash [33]	**30.11**	**11.54**	43.08	28.74	**18.79**	35.45	13.00	40.00
GPT-4o [30]	28.47	10.21	**61.90**	**37.14**	12.51	40.00	10.00	**52.00**
Open-source Multimodal LLMs								
MiniCPM-V 4.5 [34]	4.23	-9.03	-9.52	-2.86	-15.37	-10.01	12.69	31.00
LLaVA-NeXT-Video-7B [21]	-7.90	-16.06	4.76	5.72	-7.14	-10.01	8.22	-1.99
Qwen2.5-VL-7B [20]	9.07	-5.12	19.04	5.72	-16.66	26.28	-13.33	22.00
Qwen2.5-VL-72B [20]	16.35	1.45	33.34	14.28	-8.32	34.85	0.00	34.00
InternVL3-8B [35]	9.07	-11.92	23.80	5.72	-9.99	-13.33	13.00	10.00
InternVL3-38B [35]	10.28	-7.30	38.10	5.72	-16.66	0.00	7.86	25.00
InternVL3-78B [35]	12.71	-2.92	38.10	14.28	-3.33	6.66	20.71	34.00
Internlm-xcomposer2-7B [36]	1.80	-12.55	-28.58	-14.28	-26.92	0.00	28.36	4.00
Internlm-xcomposer2.5-7B [37]	1.80	-11.01	-14.28	-8.58	-23.07	10.00	**37.32**	10.00
mPLUG-Owl3-7B [38]	11.50	-4.46	-19.04	-11.42	-19.99	6.66	29.29	13.00
Emu3-Chat [39]	-7.90	-16.06	-9.52	-2.86	-3.84	-20.00	-17.85	-7.99
LLaVA-OneVision-Qwen2-7B [40]	13.92	-0.73	33.34	8.58	-12.49	31.43	-6.66	31.00
LLaVA-OneVision-Qwen2-72B [40]	22.41	8.02	**61.90**	25.72	0.00	**41.71**	3.34	40.00

Model	Multi-view	Common-Sense	Gait	Cinematographic	Olympic	Dance	**Overall**
Human	82.31	88.00	92.50	88.00	85.00	96.94	88.89
Closed-source Multimodal LLMs							
Gemini-2.5-Flash [33]	58.50	**38.17**	24.24	**37.00**	16.00	**19.38**	29.33
GPT-4o [30]	65.01	32.65	31.00	32.50	12.00	**19.38**	**31.10**
Open-source Multimodal LLMs							
MiniCPM-V 4.5 [34]	70.00	8.17	-3.49	-6.49	**37.14**	11.88	7.66
LLaVA-NeXT-Video-7B [21]	2.50	0.51	-3.49	-6.49	-5.56	-12.49	4.30
Qwen2.5-VL-7B [20]	55.00	11.23	24.24	19.00	-4.00	-1.24	10.90
Qwen2.5-VL-72B [20]	67.50	21.94	**31.83**	23.50	4.00	6.25	20.00
InternVL3-8B [35]	15.00	8.17	10.00	10.00	11.12	-10.62	3.99
InternVL3-38B [35]	20.01	12.76	14.50	17.50	-20.00	-4.99	6.97
InternVL3-78B [35]	35.01	18.88	17.50	23.50	-12.00	0.63	14.06
Internlm-xcomposer2-7B [36]	0.00	-4.08	-3.49	-6.49	-12.00	-10.62	-5.22
Internlm-xcomposer2.5-7B [37]	12.51	11.23	13.00	8.50	-12.00	-3.12	2.89
mPLUG-Owl3-7B [38]	35.01	-12.49	1.00	10.00	-16.66	9.28	3.41
Emu3-Chat [39]	-12.49	-8.67	-3.49	-18.49	-6.66	-10.62	-10.89
LLaVA-OneVision-Qwen2-7B [40]	**72.04**	11.23	17.50	11.50	4.00	-1.24	15.16
LLaVA-OneVision-Qwen2-72B [40]	**72.04**	23.47	25.00	20.50	16.00	10.00	25.79

normalized proficiency score $\mathscr{S}_{mt}$ as:

$$\mathscr{S}_{mt} = \frac{\mathsf{A}_{mt} - \mathfrak{C}_t}{1 - \mathfrak{C}_t} \tag{2}$$

Aggregate capability scores $\bar{\mathscr{S}_m}$ for model m across N subtasks are derived through weighted averaging of $\mathscr{S}_{mt}$ across tasks, with weights proportional to the task's *information gain potential* $(1 - \mathfrak{C}_t)$, thereby prioritizing tasks with higher discriminative power.

$$\bar{\mathscr{S}_m} = \frac{\sum_{t=1}^{N} w_t \cdot \mathscr{S}_{mt}}{\sum_{t=1}^{N} w_t}, \quad \text{where} \quad w_t = 1 - \mathfrak{C}_t \tag{3}$$

B. Analysis

- **Temporal Splicing (w/ & w/o Caption):** Lowest accuracy (10.39%; -4.40%), indicating core temporal reasoning limitations. Marginally better caption-guided performance suggests text partially mitigates challenges, but models still struggle with multi-video sequencing. Strong inter-task correlation confirms shared dependence on seriality—poorly handled by current architectures.
- **Action Evaluation (Daily/Professional):** Moderate accuracy (Daily: 18.43%; Professional: 7.44%) with positive cross-task correlation. The performance gap reveals scaling difficulty with temporal specificity, confirming context-dependent action comprehension. Inclusion of both subtasks prevents bias toward generic/domain-specific understanding.
- **Video Quality Assessment (Clarity/Brightness/Forensic):** Divergent trends: Clarity (11.98%) and Brightness (9.73%) rely on low-level features, while Forensic (22.40%) requires semantic reasoning. Weak inter-task correlation indicates quality assessment comprises distinct subskills—from pixel analysis to physical plausibility.
- **Logic Inference (Multiview/Common-Sense):** Highest accuracy in Multiview Pairing (37.84%, aligning with static matching), versus Common-Sense Judgment (11.54%, requiring temporal causality). This dichotomy validates MVP-Bench's coverage of static/dynamic reasoning.
- **Similarity Matching (Gait/Style/Sports/Dance):** Highly variable performance (Gait: 13.06%; Cinematographic: 11.70%; Olympic: 9.48%; Dance: 1.46%). Weak correlations confirm these probe distinct temporal dimensions: motion patterns (Gait/Dance) versus stylistic techniques (Cinematographic).

Collectively, our experimental findings suggest that current

TABLE IV: **The result of first-option frequency evaluated on Qwen2.5-VL-72B and its random expectation.** The numbers represent MLLM-Human Agreement [41].

Baseline	**Tem w/ Cap↑**	**Tem w/o Cap↑**	**Gait↑**	**Cinematographic↑**	**Olympic↑**	**Dance↑**
Model	21.13%	28.30%	49.00%	68.00%	78.00%	46.25%
Random Expectation	10.56%	16.67%	33.33%	33.33%	50.00%	33.33%

multimodal models exhibit significant limitations in processing and integrating multiple video inputs. As illustrated in Tab. III, some models [36], [39] not only underperform in complex multi-video tasks but also fail to surpass even random-chance baselines. We also found that current models exhibit systematic biases when processing multi-video inputs (e.g., *the model tends to respond 1, 2, 3, ... in the splicing task, and chooses the first video more frequently than other videos in the matching task*). These biases suggest that rather than genuinely analyzing and integrating visual information, the models rely on superficial heuristics or textual priors, leading to suboptimal and inconsistent reasoning. In terms of the model's positional bias toward the first video in candidate rankings, we have performed additional analyses, the results are shown in Tab. IV. When ground-truth answers are uniformly distributed, preferentially selecting the first option guarantees performance matching random chance. By intentionally reducing first-option frequency in our ground-truth, we avoid this situation—explaining why models frequently underperform random baselines in our results.

C. Ablation Study

To clarify the essential distinctions between single-video benchmarks, we conducted a comparative experiment to determine whether adopting the following alternative approach would result in a significant performance gap compared to the multi-video input method presented in our article.

To better showcase our experimental setup, we selected two diving videos exhibiting significant divergence in human-assigned quality ratings as examples. We processed both videos through the model and evaluated them using the two alternative methodologies. The experimental results and potential explanations for the inconsistency in performance are as below:

Single video input + Task-specific metrics comparison.

Video A (Human score: 98.4) → Caption: "Performed well" → Score: 7.0

Video B (Human score: 29.7) → Caption: "Lack of Liquidity", "Imperfect" → Score: 7.8

- Limited domain expertise makes MLLMs struggle to assign precise numeric ratings, whereas comparing multiple videos side by side lets the model make relative, ranking-like judgments—proven in InstructGPT that ranking-based fine-tuning outperform absolute scoring.
- Lack of unified scoring baseline. Each video is scored independently, so a "7" for Video A may represent very strong performance in that context, while a "7.8" for Video B may actually signal weaker performance. Absence of multi-video's visual information and comparability prevents direct judgement of these scores.

Single video input + Caption comparison.

Video A (Human score: 98.4) → Caption: "Forward 3½ Somersaults Pike (DD 3.2)", "Exceptional body control and stability", "Rushed takeoff and imperfect pike shape" → Score: 7.8

Video B (Human Score: 29.7) → Caption: "Forward 2½ Somersaults Pike (DD 2.8)", "Minor crown splash", "Good rhythm and slight reduction in fluidity" → Score: 8.1

- Models cannot reliably distill discriminative visual details into text (e.g., The extent of the splash when entering the water is not fully demonstrated).
- Subtle differences (e.g., degree of body tilt) become indistinguishable in text formats. Consequently, the model fails to discriminate salient actions when processing undifferentiated textual descriptions.
- Lack of visual information for quantitative scoring.

As illustrated in Tab. V, the model exhibits a significant performance degradation across the single-video style benchmark experiments on all evaluation tasks relative to our multi-video methodology. Our experiments demonstrate that single-video style evaluation methodologies invariably produce discrepancies between assessed performance and the model's core competencies.

TABLE V: **Ablation study across the single-video style benchmark experiments on all evaluation tasks.** All experimental values represent the mean MLLM-Human Agreement scores [41], averaged across models with native multi-video processing capabilities [30], [33], [40].

(a) The results of single-video input followed by scores comparison and direct multi-video comparison.

Subtasks	VQA	VLI	TSS	CA	SVP
Random	33.33%	33.33%	11.41%	45.22%	35.86%
Single-video	39.10%	37.97%	11.41%	51.59%	39.70%
Multi-video	**58.97%**	**65.19%**	**26.76%**	**61.15%**	**60.61%**

(b) The results of generating captions followed by textual comparison and direct multi-video comparison.

Subtasks	VQA	VLI	TSS	CA	SVP
Random	33.33%	33.33%	11.41%	45.22%	35.86%
Caption Comparison	34.62%	53.80%	14.08%	50.96%	46.36%
Video Comparison	**58.97%**	**65.19%**	**26.76%**	**61.15%**	**60.61%**

V. Conclusion

We introduce MVPBench, a benchmark designed to evaluate the perceptual understanding capabilities of MLLMs when dealing with multiple video inputs. Our experimental results indicate that these multi-video benchmarks present significant challenges to a wide range of current advanced MLLMs. Through our benchmark, we aim to stimulate further research in this area, encouraging exploration into enhancing the model's ability to comprehend multi-video inputs. MVPBench is positioned to serve as a comprehensive and integrated testing platform for evaluating related models, thereby supporting advancements in this field.

References

[1] A. J. OpenAI, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report, 2024," *URL https://arxiv. org/abs/2303.08774*, 2024.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NIPS*, 2024.

[4] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, p. 3, 2023.

[5] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo, "Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models," *arXiv preprint arXiv:2306.09265*, 2023.

[6] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *ECCV*, 2025, pp. 216–233.

[7] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan, "Seed-bench: Benchmarking multimodal large language models," in *CVPR*, 2024, pp. 13 299–13 308.

[8] X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna, "Blink: Multimodal large language models can see but not perceive," *arXiv preprint arXiv:2404.12390*, 2024.

[9] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *CVPR*, 2024, pp. 22 195–22 206.

[10] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang *et al.*, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," *arXiv preprint arXiv:2405.21075*, 2024.

[11] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in *AAAI*, 2018.

[12] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *CVPR*, 2018, pp. 6057–6066.

[13] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *WACV*, 2019, pp. 1468–1476.

[14] A. Antsiferova, S. Lavrushkin, M. Smirnov, A. Gushchin, D. Vatolin, and D. Kulikov, "Video compression dataset and benchmark of learning-based video-quality metrics," in *NIPS*, 2022, pp. 13 814–13 825.

[15] R. Wang, X. Xu, C.-W. Fu, J. Lu, B. Yu, and J. Jia, "Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment," in *ICCV*, 2021, pp. 9700–9709.

[16] R. Yonetani, K. M. Kitani, and Y. Sato, "Recognizing micro-actions and reactions from paired egocentric videos," in *CVPR*, 2016, pp. 2629–2638.

[17] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *ICPR*, 2006, pp. 441–444.

[18] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[19] R. Luo, Z. Zhao, M. Yang, J. Dong, D. Li, P. Lu, T. Wang, L. Hu, M. Qiu, and Z. Wei, "Valley: Video assistant with large language model enhanced ability," *arXiv preprint arXiv:2306.07207*, 2023.

[20] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[21] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024.

[22] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *CVPR*, 2017, pp. 2758–2766.

[23] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *ACM*, 2017, pp. 1645–1653.

[24] K. Mangalam, R. Akshulakov, and J. Malik, "Egoschema: A diagnostic benchmark for very long-form video language understanding," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 212–46 244, 2023.

[25] H. Wu, D. Li, B. Chen, and J. Li, "Longvideobench: A benchmark for long-context interleaved video-language understanding," in *NIPS*, 2024, pp. 28 828–28 857.

[26] Y. Li, X. Chen, B. Hu, L. Wang, H. Shi, and M. Zhang, "Videovista: A versatile benchmark for video understanding and reasoning," *arXiv preprint arXiv:2406.11303*, 2024.

[27] M. Ning, B. Zhu, Y. Xie, B. Lin, J. Cui, L. Yuan, D. Chen, and L. Yuan, "Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models," *arXiv preprint arXiv:2311.16103*, 2023.

[28] X. Chen, Y. Lin, Y. Zhang, and W. Huang, "Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering," in *ECCV*, 2024, pp. 179–195.

[29] Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, and L. Hou, "Tempcompass: Do video llms really understand videos?" *arXiv preprint arXiv:2403.00476*, 2024.

[30] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[31] Kwai, "https://klingai.kuaishou.com," 2024.

[32] Z. AI, "https://chatglm.cn/video," 2024.

[33] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[34] T. Yu, Z. Wang, C. Wang, F. Huang, W. Ma, Z. He, T. Cai, W. Chen, Y. Huang, Y. Zhao *et al.*, "Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe," *arXiv preprint arXiv:2509.18154*, 2025.

[35] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[36] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, X. Wei, S. Zhang, H. Duan, M. Cao *et al.*, "Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model," *arXiv preprint arXiv:2401.16420*, 2024.

[37] P. Zhang, X. Dong, Y. Zang, Y. Cao, R. Qian, L. Chen, Q. Guo, H. Duan, B. Wang, L. Ouyang *et al.*, "Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output," *arXiv preprint arXiv:2407.03320*, 2024.

[38] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou, "mplug-owl3: Towards long image-sequence understanding in multi-modal large language models," *arXiv preprint arXiv:2408.04840*, 2024.

[39] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[40] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[41] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu *et al.*, "A survey on llm-as-a-judge," *arXiv preprint arXiv:2411.15594*, 2024.