Breaking Down Video LLM Benchmarks: Knowledge, Spatial Perception, or True Temporal Understanding?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Existing video understanding benchmarks often conflate knowledge-based and purely image-based questions, rather than clearly isolating a model's temporal reasoning ability, which is the key aspect that distinguishes video understanding from other modalities. We identify two major limitations that obscure whether higher scores truly indicate stronger understanding of the dynamic content in videos: (1) strong language priors, where models can answer questions without watching the video; and (2) shuffling invariance, where models maintain similar performance on certain questions even when video frames are temporally shuffled. To alleviate these issues, we propose VBenchComp, an automated pipeline that categorizes questions into different domains: LLM-Answerable, Semantic, and Temporal. Specifically, LLM-Answerable questions can be answered without viewing the video; Semantic questions remain answerable even when the video frames are shuffled; and Temporal questions require understanding the correct temporal order of frames. The rest of the questions are labeled as Others. This can enable fine-grained evaluation of different capabilities of a video LLM. Our analysis reveals nuanced model weaknesses that are hidden by traditional overall scores, and we offer insights and recommendations for designing future benchmarks that more accurately assess video LLMs.

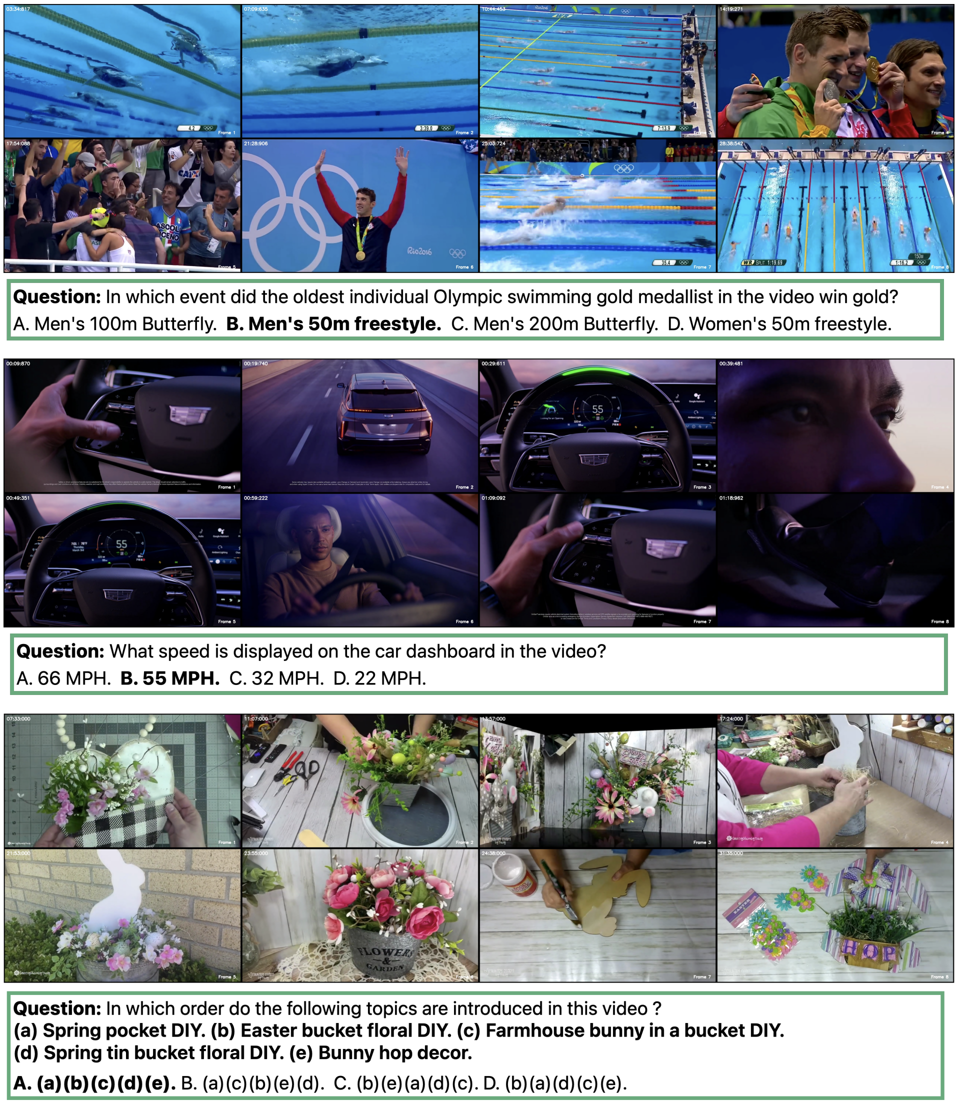
1 Introduction

The rapid progress of video Large Language Models (video LLMs) has led to the emergence of a wide range of video understanding benchmarks, such as VideoMME [4], MLVU [5], LongVideoBench [1], EgoSchema [2], and others. While this surge of benchmarks offers broader coverage for evaluating different capabilities, it also introduces considerable computational cost and redundancy. As shown in Table 1, evaluating a 2B-parameter model (e.g., Qwen2-VL) across existing video QA benchmarks requires 190.6 A100 GPU hours. This computational cost escalates dramatically to 491.9 hours for a 72B model, raising serious concerns about the computational burden of benchmarking video LLMs given the growing number of video understanding datasets.

Table 1: A100 GPU hours needed for evaluating video LLMs (Qwen2-VL) across different benchmarks.

Benchmark	Number of Questions	Model Size		
		2B	7B	72B
LongVideoBench [1]	1337	14.3	15.0	26.4
Egoschema [2]	500	2.7	2.2	6.7
NexTQA [3]	4996	16.0	20.0	58.0
VideoMME [4]	2700	18.0	20.7	50.7
MLVU [5]	2174	14.1	15.9	31.7
LVBench [6]	1549	7.5	8.7	22.2
PerceptionTest [7]	19140	118.1	131.2	296.3
Total	32396	190.6	213.7	491.9

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.



Question: In which event did the oldest individual Olympic swimming gold medallist in the video win gold?
A. Men's 100m Butterfly. **B. Men's 50m freestyle.** C. Men's 200m Butterfly. D. Women's 50m freestyle.

Question: What speed is displayed on the car dashboard in the video?
A. 66 MPH. **B. 55 MPH.** C. 32 MPH. D. 22 MPH.

Question: In which order do the following topics are introduced in this video ?
(a) Spring pocket DIY. (b) Easter bucket floral DIY. (c) Farmhouse bunny in a bucket DIY.
(d) Spring tin bucket floral DIY. (e) Bunny hop decor.
A. (a)(b)(c)(d)(e). B. (a)(c)(b)(e)(d). C. (b)(e)(a)(d)(c). D. (b)(a)(d)(c)(e).

Figure 1: Examples of LLM-Answerable, Semantic and Temporal questions in VideoMME [4]: (Top) The model uses LLM's prior knowledge to answer correctly without the need of video; (Middle) The model relies on semantic understanding to answer without requiring temporal comprehension; (Bottom) The model relies on comprehensive temporal understanding to answer.

Beyond the computational cost, current video understanding benchmarks often conflate different skills and fail to truly evaluate the video understanding capability. We identify two key limitations that undermine meaningful evaluation. First, some questions can be answered correctly without access to the video, since models rely on their pretrained language priors rather than visual evidence, as shown in Figure 1. These questions primarily test the underlying LLM's factual knowledge and reasoning skills, rather than evaluating the model's ability to process and understand visual content. As a result, high performance on these questions can misleadingly inflate benchmark scores, giving the false impression of strong video understanding when, in fact, the model may not be attending to the visual input at all. Second, some questions primarily assess static semantic understanding and do not require comprehension of the video's temporal structure. For example, models often achieve similar performance even when the video frames are randomly shuffled, indicating that their predictions rely heavily on spatial or frame-level cues rather than temporal reasoning. This shuffling invariance exposes a critical flaw: current benchmarks may significantly overestimate a model's true temporal understanding, conflating static visual recognition with dynamic sequence reasoning.

While many existing benchmarks claim to be comprehensive, there is currently no standardized protocol for assessing their effectiveness. Each dataset emphasizes different aspects of video com-

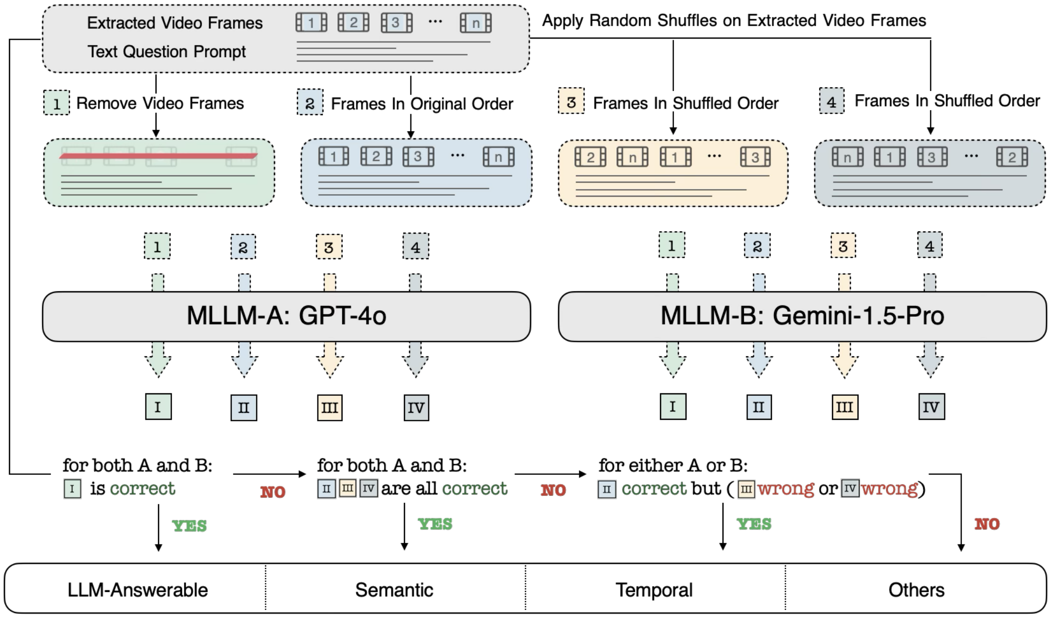

Extracted Video Frames
Text Question Prompt
Apply Random Shuffles on Extracted Video Frames
1 Remove Video Frames
2 Frames In Original Order
3 Frames In Shuffled Order
4 Frames In Shuffled Order
MLLM-A: GPT-4o
MLLM-B: Gemini-1.5-Pro
for both A and B: I is correct
NO
YES
for both A and B: II III IV are all correct
for either A or B: II correct but (III wrong or IV wrong)
LLM-Answerable
Semantic
Temporal
Others

Figure 2: An overview of our standardized protocol: benchmark questions are categorized into four groups. Questions answerable by both GPT-4o and Gemini without video are classified as LLM-Answerable. For the remaining questions, we apply random shuffles to the extracted frames twice: if both models answer correctly before and after shuffling, the question is classified as Semantic. If one model answers correctly before but fails after shuffling, the question is classified as Temporal. All other questions are categorized as Others.

prehension, yet lacks a clear metric for how well it captures temporal reasoning, which is the core capability that distinguishes video from static images. We introduce VBenchComp, an automated evaluation pipeline that categorizes questions into four distinct domains: LLM-Answerable, Semantic, Temporal, and Other. This structured categorization disentangles the contributions of language priors, static visual understanding, and genuine temporal reasoning, enabling a more diagnostic and interpretable evaluation of video models. Based on this, we curate a core benchmark subset that emphasizes both semantic and temporal understanding, and introduce a dedicated metric, the VBenchComp Score, which provides a more focused and light-weighted evaluation protocol to better guide model development and comparison. Importantly, we find that results obtained from this core set are consistent with those from the full benchmark suite, while reducing computational cost significantly.

2 A Standardized Protocol for Breaking Down Video LLM Benchmarks

In this section, we propose a standardized protocol (as shown in Figure 2) for decomposing video LLM benchmarks into four distinct domains: (1) LLM-answerable questions to focus on the prior language capabilities of the LLM backbone, (2) semantic understanding questions to evaluate the model's ability to understand semantic content, (3) temporal understanding questions to measure the model's capacity to capture temporal dependencies and dynamic changes, and (4) Other questions that may either require overly advanced comprehensive reasoning or are poorly constructed and thus lack sufficient distinctiveness. Our goal is to disentangle these question types to provide a more precise and comprehensive evaluation of video LLMs. It will also assist future benchmarks in refining their question design strategies and focusing more on authentic video-understanding questions.

LLM-Answerable Questions. Answer leakage is a critical issue in image-QA benchmarks, where MLLMs can often generate correct answers without relying on the image itself. Instead of genuinely integrating visual and textual information, these models leverage their pre-trained knowledge from LLM to infer answers based solely on the text [8]. This undermines the intended goal of evaluating a model's multimodal understanding capabilities. Multimodal answer leakage can be summarized into two categories: 1) text-answerable questions, where the question itself provides sufficient information for the model to answer, rendering the associated visual input unnecessary; 2) memorized questions, where the MLLM has previously encountered the same question during training and recalls the corresponding answer from memory rather than reasoning from the given image. As a result, certain questions can be answered solely by a text-based LLM without requiring visual input. To assess this, we perform a text-only evaluation using both GPT-4o and Gemini-1.5-Pro. As shown in Figure 2, if both models correctly answer a given question without the video, we classify the corresponding QA pair as an LLM-Answerable question. We then analyze the entire benchmark and compute the proportion of such questions relative to the total, denoted as α.

Semantic Questions: Shuffling Frames but Consistent Answer. After filtering for LLM-answerable questions, we further identify a subset of questions that focus specifically on semantic understanding. To achieve this, we introduce a diagnostic procedure: for each video-question pair, we first generate answers using Gemini-1.5-Pro and GPT-4o. We then shuffle the extracted frames and query the models again - repeating this process twice. If both models consistently provide correct answers despite the disrupted temporal order (before and after shuffling the extracted frames), we classify the question as semantic, indicating that static visual information from a single or a certain group of frames alone are sufficient for answering. By applying this procedure across the benchmark, we compute the proportion of such questions, denoted as β, to quantify the prevalence of questions relying solely on semantic understanding. A high β suggests that the benchmark may be biased toward spatial or appearance-based cues, potentially inflating a model's perceived temporal reasoning capability. This highlights the need to construct more temporal-related questions that explicitly require sequential understanding to ensure a more rigorous and targeted evaluation of video LLMs.

Temporal Questions & Others. After classifying questions into LLM-Answerable and Semantic categories, the remaining questions are further divided into Temporal and Others. To identify Temporal questions, we apply the following criterion: if GPT-4o or Gemini-1.5-Pro answers the question correctly when provided with frames in their original order but fails to do so after the frames are shuffled, we classify the question as Temporal, indicating that the right sequential information is crucial for the answering process. Unlike semantic or frame-independent tasks, these questions assess whether the model can correctly infer event progression and temporal consistency over time. By introducing a controlled perturbation—shuffling the frame order, we isolate the questions for temporal understanding capacity, distinguishing them from purely visual or semantic understanding.

Lastly, the remaining questions will be labeled as Others. This category includes questions that are either too difficult to answer for all SOTA models or are so comprehensive that they may require additional modalities, such as audio, to resolve. Questions may depend on recognizing spoken dialogue, distinguishing between environmental sounds, or interpreting non-visual context cues like tone or timing. For example, in VideoMME [4], answering certain questions may depend on recognizing spoken dialogue, distinguishing between environmental sounds, or interpreting non-visual context cues like tone or timing.

VBenchComp: Quantifying Video Benchmark Composition. To systematically analyze and quantify the composition of video LLM benchmarks, we introduce **VBenchComp**, a diagnostic tool that applies our standardized protocol (Figure 2) to decompose the benchmark into its four key domains. VBenchComp computes the ratios of LLM-Answerable, Semantic, Temporal, and Others questions, denoted as α, β, γ, and δ respectively.

Benchmark profiling and skill gap identification. VBenchComp not only quantifies benchmark composition but also identifies potential gaps in coverage. For instance, an overrepresentation of LLM-Answerable questions (α) suggests that the benchmark may underestimate the need for genuine multimodal understanding. Conversely, an excess of Semantic questions (β) could create an illusion of strong temporal understanding, when in reality, the model might rely primarily on static frame information. A low proportion of Temporal questions (γ) may indicate inadequate assessment of dynamic event comprehension.

3 Discussion

VBenchComp provides a structured and interpretable framework for dissecting the capabilities of video LLMs, highlighting whether models rely on language priors, static semantics, or genuine temporal reasoning. This diagnostic lens not only clarifies what current benchmarks actually measure, but also helps researchers identify blind spots in model behavior.

References

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

[2] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 2024.

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

[4] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[5] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[6] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.

[7] 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. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023.

[8] King Zhu, Qianbo Zang, Shian Jia, Siwei Wu, Feiteng Fang, Yizhi Li, Shawn Gavin, Tuney Zheng, Jiawei Guo, Bo Li, et al. Lime: Less is more for mllm evaluation. *arXiv preprint arXiv:2409.06851*, 2024.

[9] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[11] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv:2311.10122*, 2023.

A Experimental Results

A.1 An Overview of VBenchComp

We apply the standardized categorization protocol described in Section 2 to seven widely-used video question answering benchmarks, quantifying their distributions across four diagnostic categories: LLM-Answerable, Semantic, Temporal, and Others. Table 2 summarizes the raw counts and their corresponding percentages relative to the total number of questions in each benchmark. Across all benchmarks, we observe a considerable variation in the proportion of question types, which reflects their differing emphases on language, semantic, and temporal capabilities. For instance, *NextQA* [3], *LongVideoBench* [1], *MLVU* [5], *Egoschema* [2], and *VideoMME* [4] contain a significant portion of LLM-Answerable questions, which indicates potential answer leakage and reliance on language priors. In contrast, benchmarks like *LVBench* [6] contains relatively fewer LLM-Answerable questions. On the other hand, with the exception of *LongVideoBench* and *LVBench*, all other benchmarks have more than 30% of Semantic questions, where frame shuffling has minimal impact on the model's ability to produce correct answers.

Table 2: Compositions of question types across different video understanding benchmarks. Each cell (except Total) shows the count and its percentage of the total.

Dataset	**Total**	**Text**	**Semantic**	**Temporal**	**Others**
LongVideoBench [1]	1337	363 / 27.15%	308 / 23.03%	235 / 17.58%	431 / 32.24%
Egoschema [2]	500	133 / 26.60%	182 / 36.40%	45 / 9.00%	140 / 28.00%
NextQA [3]	4996	1738 / 34.79%	1880 / 37.63%	437 / 8.75%	941 / 18.83%
VideoMME [4]	2700	841 / 31.15%	810 / 30.00%	371 / 13.74%	678 / 25.11%
MLVU [5]	2174	621 / 28.57%	643 / 29.57%	383 / 17.62%	527 / 24.23%
LVBench [6]	1549	140 / 9.04%	321 / 20.72%	355 / 22.92%	733 / 47.32%
PerceptionTest [7]	19140	3642 / 19.03%	6283 / 32.82%	3117 / 16.29%	6098 / 31.86%

Table 3: Benchmarking public models under VBenchComp categorization. (All settings use 64 frames, except for VideoMME-long, which uses 128.)

(a) Egoschema [2]

Size	Model	Overall	LLM	Semantic	Temporal	Others
7B	Qwen2-VL [9]	65.8	85.0	83.5	37.8	33.6
	LLaVA-OV [10]	66.2	75.2	83.5	57.8	37.9
	LLaVA-Video [11]	61.8	72.2	82.4	46.7	30.0
72B	Qwen2-VL [9]	77.4	87.2	95.1	64.4	49.3
	LLaVA-OV [10]	65.2	78.9	84.6	40.0	35.0
	LLaVA-Video [11]	70.4	81.2	90.7	53.3	39.3

(b) NextQA [3]

Size	Model	Overall	LLM	Semantic	Temporal	Others
7B	Qwen2-VL [9]	81.3	88.7	90.9	70.0	54.1
	LLaVA-OV [10]	80.3	89.8	91.1	65.7	48.2
	LLaVA-Video [11]	84.4	92.8	92.3	73.7	56.7
72B	Qwen2-VL [9]	84.0	91.1	92.6	70.9	60.0
	LLaVA-OV [10]	83.2	93.4	93.9	66.6	50.6
	LLaVA-Video [11]	85.4	94.0	94.7	73.7	56.6

(c) VideoMME [4]

Size	Model	Overall	LLM	Semantic	Temporal	Others
7B	Qwen2-VL [9]	60.6	77.8	78.4	36.7	31.1
	LLaVA-OV [10]	59.0	76.3	76.8	37.2	28.2
	LLaVA-Video [11]	63.9	79.3	82.0	42.6	34.7
72B	Qwen2-VL [9]	68.2	86.8	86.3	49.6	33.8
	LLaVA-OV [10]	68.7	87.2	86.3	52.6	33.6
	LLaVA-Video [11]	70.8	88.1	88.9	51.8	38.1

(d) MLVU [5]

Size	Model	Overall	LLM	Semantic	Temporal	Others
7B	Qwen2-VL [9]	62.5	77.8	79.5	43.6	37.4
	LLaVA-OV [10]	65.2	77.1	88.0	47.5	36.1
	LLaVA-Video [11]	63.7	77.8	83.1	49.6	33.6
72B	Qwen2-VL [9]	67.9	81.8	85.4	52.5	41.4
	LLaVA-OV [10]	74.2	88.1	92.5	62.7	44.0
	LLaVA-Video [11]	74.2	87.4	92.5	64.0	43.8

(e) LongVideoBench [6]

Size	Model	Overall	LLM	Semantic	Temporal	Others
7B	Qwen2-VL [9]	52.8	74.4	70.5	42.6	27.6
	LLaVA-OV [10]	58.9	79.1	82.1	49.8	30.2
	LLaVA-Video [11]	59.8	81.3	84.4	49.8	29.7
72B	Qwen2-VL [9]	58.0	82.4	76.0	46.4	30.9
	LLaVA-OV [10]	59.8	87.3	84.4	49.8	32.9
	LLaVA-Video [11]	62.8	87.3	85.7	52.8	31.3

(f) PerceptionTest [7]

Size	Model	Overall	LLM	Semantic	Temporal	Others
7B	Qwen2-VL [9]	60.7	71.9	84.2	49.7	35.3
	LLaVA-OV [10]	58.0	66.0	84.9	45.9	31.9
	LLaVA-Video [11]	68.3	75.4	87.9	60.8	47.8
72B	Qwen2-VL [9]	68.1	77.7	92.1	62.7	40.5
	LLaVA-OV [10]	62.5	75.6	89.8	50.6	32.8
	LLaVA-Video [11]	69.6	76.0	92.1	61.2	47.0

A.2 Benchmarking Public Models Under VBenchComp Categorization

Table 3 benchmarks recent public video-language models under our proposed VBenchComp framework, which categorizes questions into LLM-answerable, Semantic, and Temporal types. This fine-grained categorization provides a more diagnostic view of model capabilities compared to a single overall score. As shown in Table 3(a), Qwen2-VL-7B slightly outperforms LLaVA-Video-7B in terms of the traditional overall score on Egoschema. However, this superficial advantage is misleading. A breakdown of the scores shows that the performance gain is almost entirely due to LLM-answerable questions that do not require visual or temporal understanding. However, the two models perform similarly on Semantic questions, and Qwen2-VL-7B even lags behind on Temporal questions, which indicates a weaker grasp of fine-grained video temporal understanding. These findings suggest that Qwen2-VL-7B's advantage is largely attributable to its stronger language model backbone, rather than superior visual or temporal reasoning. In contrast, LLaVA-Video-7B, though slightly behind overall, demonstrates more balanced capabilities across semantic and temporal dimensions.

Interestingly, the comparison flips in VideoMME (Table 3(c)), where LLaVA-Video-7B outperforms Qwen2-VL-7B not just overall, but more meaningfully across both vision-dependent axes. While the two models perform similarly on LLM-answerable questions, LLaVA-Video-7B achieves notably higher scores on both Semantic (82.0 vs. 78.4) and Temporal (42.6 vs. 36.7) categories. This demonstrates that LLaVA-Video-7B possesses stronger visual and temporal understanding, reinforcing the claim that strong language knowledge alone are insufficient for robust video understanding.

These results collectively demonstrate a core limitation of traditional evaluation: a single overall score fails to capture specific model strengths and weaknesses. Only through our VBenchComp categorization can we identify crucial gaps in semantic or temporal understanding that would otherwise be masked. This insight is not only critical for fair benchmarking but also for guiding the development of next-generation video LLMs, where improvement must go beyond language modeling and target true temporal understanding.

A.3 VBenchComp Score: Fewer Questions, Deeper Video Understanding

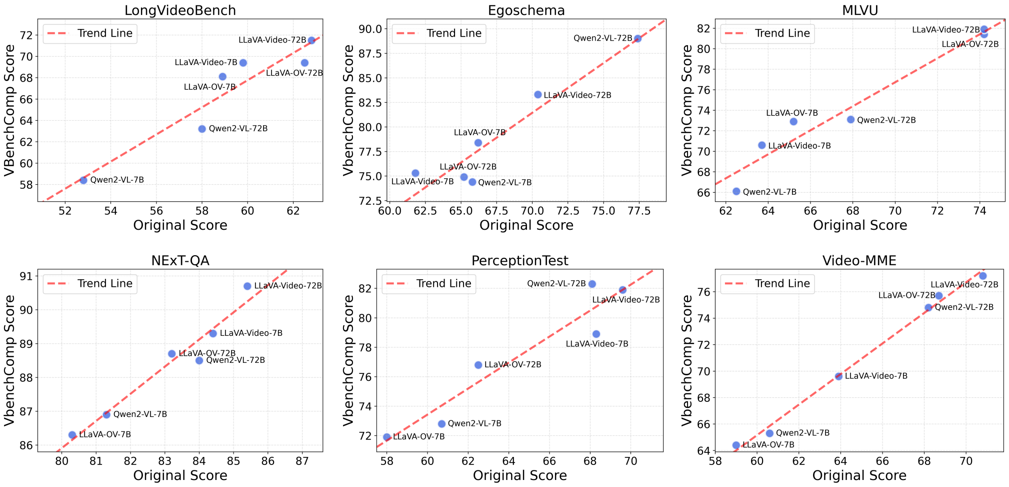


Figure 3: VBenchComp scores are aligned with the original scores but they can better evaluate the overall video LLM performance with less questions. The temporal video understanding capability of models under the trend line can be potentially over-estimated in the original benchmarks.

Based on the above analysis, we retain only the Semantic and Temporal questions from each benchmark to compute a focused evaluation score, denoted as the VBenchComp score. The results across models are shown in Figure 3. Despite removing nearly 50% of the original questions (as detailed in Table 2), the model rankings remain highly consistent with those based on the original scores. This strong correlation indicates that Semantic and Temporal questions alone are sufficient to preserve the discriminative power of the benchmark. It further suggests that many of the remaining questions may be redundant or less critical for evaluating core model capabilities, and that VBenchComp can serve as a more focused yet reliable metric for model comparison.

B Discussion

VBenchComp provides a structured and interpretable framework for dissecting the capabilities of video LLMs, highlighting whether models rely on language priors, static semantics, or genuine temporal reasoning. This diagnostic lens not only clarifies what current benchmarks actually measure, but also helps researchers identify blind spots in model behavior. However, our approach is not without limitations. First, while our categorization pipeline is automated and scalable, it heavily relies on GPT-4o and Gemini, which may introduce biases. Second, our core benchmark subset, while compute-efficient and representative in aggregate, may omit edge cases that appear in the full benchmark suite. Finally, VBenchComp focuses primarily on question-answering tasks; generalizing this framework to other video understanding tasks like captioning, retrieval, or grounding remains an important avenue for future work.