

VirtueBench: Evaluating Trustworthiness under Uncertainty in Long Video Understanding

Supplementary Material

1. Limitations of Existing Long-video Benchmark Evaluations

Existing benchmarks for long-video understanding, such as MLVU [4], LVBench [2], LongVideoBench [3], and VideoEval-pro [1], exhibit unreliable result when evaluating state-of-the-art VLMs. To demonstrate this, we conduct an evaluation using a uniform 64-frame sampling across all benchmarks, as shown in Table 1. We report results on both the original Multi-Choice Question (MCQ) and adapted Open-Ended (OE) formats to mitigate the impact of random guessing inherent in MCQs [1].

Counter-intuitively, we observe that models typically recognized for superior capabilities often yield lower scores. For instance, on MLVU, LLaVA-Video-7B and InternVideo2.5-8B significantly outperform more advanced models such as Qwen2.5-VL-7B, Mimo-VL-7B and Qwen3-VL-8B. Notably, Qwen2.5-VL-7B exhibits much lower performance, likely due to "honest refusal", where the model correctly declines to answer when critical visual information is missing from the sampled frames. Conversely, other models may achieve artificially high accuracy by hallucinating or guessing the correct answer despite missing evidence.

This reliability issue stems primarily from two structural flaws in existing evaluations:

- Existing benchmarks utilize ground truth answers derived from the full video. However, the correct answer may change or become invalid under different frame sampling rates. Comparing model outputs generated from sparse sampling against full-video ground truth leads to unfair comparisons.
- Due to context window limits, most models cannot process all video frames. Consequently, many questions

become unanswerable under incomplete sampling. Existing metrics treat honest refusals (indicating insufficient information) as incorrect, while potentially rewarding models that guess correct answers by chance.

In summary, by neglecting the impact of frame down-sampling, existing benchmarks fail to provide a trustworthy assessment of model capabilities. In contrast, VirtueBench ensures a fair evaluation by rigorously verifying answers manually against specific frame settings. Questions that become unanswerable or whose answers shift due to down-sampling are explicitly corrected, ensuring that models are evaluated on what is actually visible.

2. Source Distribution of VirtueBench

VirtueBench is constructed by aggregating data from 6 popular long-video benchmarks. Their proportional distribution within the final benchmark is shown in Fig 1.

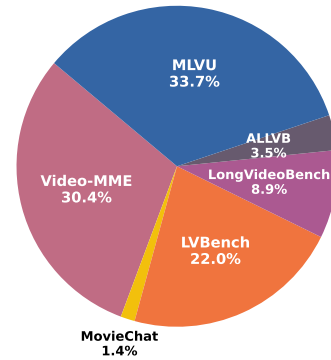


Figure 1. Distribution of Source.

Model	MLVU		LVBench		LongVideoBench		VideoEval-pro	VirtueBench
	OE	MCQ	OE	MCQ	OE	MCQ		Overall
Qwen2.5VL-7B	24.89	56.49	13.41	37.92	12.34	50.19	14.82	38.03
Qwen3VL-8B	34.73	65.73	16.76	41.27	17.35	57.67	24.13	38.78
Mimo-VL-7B	28.52	63.62	14.59	40.36	15.33	51.46	19.78	35.32
Keye-VL-8B	34.18	75.71	16.76	40.99	15.71	62.23	24.67	24.55
LLaVA-Video-7B	32.11	71.16	15.15	43.30	15.48	43.90	23.89	23.57
InternVideo2.5-8B	26.91	75.76	13.55	41.34	13.99	62.68	20.56	22.67

Table 1. Performance comparison of VLMs on existing long-video benchmarks versus VirtueBench. We evaluate popular VLMs on MLVU, LVBench, LongVideoBench and VideoEval-pro against our proposed VirtueBench, using a uniform 64-frame sampling strategy.

3. Prompt for Data Filtering

We use Gemini-2.5-Flash to perform data quality filtering on existing long-video benchmarks. The filtering consists of two stages: (1) removing questions that cannot be adapted to open-ended formats, and (2) excluding trivial questions answerable without visual reasoning. The exact prompts used in each stage are provided below.

(1) Filtering Questions Unsuitable for Open-Ended Adaptation

(2) Excluding Trivial Questions Without Video Understanding

Prompt for difficulty filtering

```
Answer the question using the provided image if possible.
If the image is insufficient but the answer can be inferred from common sense, give the most likely answer.
If neither is possible, reply: "I don't know."
Output only the answer. Do not include explanations.
Question: {question}
```

Prompt for answerability filtering

You are helping filter a dataset of multiple-choice video questions to determine whether each question can be converted into an open-ended format.

Task:

Evaluate whether the question can be reliably asked and judged as an open-ended question, based solely on watching the video. Assume you are familiar with its content.

Output:

Return exactly one word: "Keep" or "Discard". Do not provide explanations.

Keep if:

- The question can be answered objectively by someone who has watched the video, even if reasoning or summarizing visual evidence is required.
- The answer contains at most 3 distinct elements (objects, actions, events), except for counting questions where the answer is a single number.
- Or, the question provides a fixed set of elements and asks for ordering or ranking them.

Discard if any of the following apply:

- The question requires comparing or evaluating the answer options.
- The question is only meaningful when the options are provided.
- The question depends on specific timestamps (e.g., "What happens at 01:20?").
- The question relies on subtitles or precise subtitle timing.
- The answer requires listing more than 3 distinct elements, unless it explicitly asks for ordering.
- The question is subjective or interpretive (e.g., emotions, opinions).

Examples:

- What color is the car that drives past the house? → Keep
- Which of the following animals appears in the video? → Discard
- Which statement is true according to the video? → Discard
- What happens at 01:20 in the video? → Discard
- Summarize the events that happen in the video. → Discard
- How does the character feel after entering the room? → Discard
- Put these 4 actions in order: (a) man enters, (b) car leaves, (c) dog barks, (d) door closes → Keep
- Describe the sequence leading to Kaffee and Jo's discussion about poisoning. → Discard

Here is the question: {question}

Reference answer: {answer}

4. Prompt for Judgement

Since our benchmark consists entirely of open-ended questions, we adopt an LLM-as-a-judge evaluation paradigm and use GPT-4o as the assessment model. As described in the main paper, the evaluation procedure contains two stages: (1) refusal detection and (2) correctness judgement. The full prompt provided to GPT-4o is shown below.

Judgement prompt for GPT-4o

```
You are a strict evaluator assessing answer correctness. You must output 1 for correct answers and 0 for any other case.
### Input
Question: {question}
Ground Truth Answer: {answer}
Model Prediction: {response}

### Evaluation Rules
Step 1: Refusal Detection - Determine whether the model refused to answer the question.
- Set refusal = 1 if the prediction indicates that the video lacks sufficient information to answer the question directly.
- If the model attempts to answer (even incorrectly), set refusal = 0.

Step 2: Judgement of Correctness
Evaluate judgement based on the ground-truth answer type.
Case A: The ground truth has a definite answer
- If refusal = 1, then judgement = 0.
- If refusal = 0, assess semantic consistency:
* Set judgement = 1 if the predicted answer is semantically consistent with the ground truth, even when expressed indirectly or with mild reasoning.
* Ignore hedging phrases (e.g., ``maybe``, ``possibly``, ``not sure``).
* Minor formatting differences are acceptable.
* Numerical values must match within a reasonable tolerance.
* For unit-based questions, both value and unit must be correct.
* For order-based questions, the event order and the selected events must match exactly.
* Otherwise, set judgement = 0.
Case B: The ground truth is ``The video does not provide enough information``
- Set judgement = 1 only if refusal = 1.
- If the model provides any concrete answer (refusal = 0), then judgement = 0.

### Strict Output Format
Respond strictly in JSON (each field must be either 0 or 1).
Example:
{"refusal": 0, "judgement": 1}
```

5. Additional Cases of VirtueBench

In this section, we provide additional qualitative examples from VirtueBench to further illustrate the critical impact of frame sampling density on video understanding and the validity of ground-truth answers. These cases demonstrate how questions that are answerable under high-frame settings may become unanswerable or misleading when the visual evidence is lost due to down-sampling.

As shown in Figure 2, we present a fine-grained temporal reasoning scenario involving a sports score update. Under the 64-frame setting, the key moment of the score change is skipped, rendering the question unanswerable. Consequently, models unable to trace the complete temporal progression are prone to relying on incorrect assumptions or priors.

Furthermore, Figure 3 illustrates a complex social interaction requiring the recognition of emotional cues and character identities. While the 1024-frame input captures the full narrative arc, including the specific emotional breakdown of the character, the sparse 64-frame input misses these subtleties. In the absence of these critical visual cues, models frequently resort to hallucination or misinterpret the character relationships.

In the first game, when the challenger scores for the first time, how many points has Bugs & Daffy scored?

64f

0:05 0:10 0:25

Gemini-2.5-Flash: The provided frames do not show the exact moment when the challenger scores for the first time in the first game. The first score display is at 0:05, showing "Bugs & Daffy 3, Challengers 0". The next score display at 0:25 shows the final score of the first game as "Bugs & Daffy 10, Challengers 5". Therefore, the frames do not provide enough information to answer when Bugs & Daffy scored when the challenger scored for the first time.

Qwen3-VL-8B: Based on the visual content of the frames: 1. At 0:04, the scoreboard shows "Bugs & Daffy 3" and "Challengers 0". This indicates that Bugs & Daffy have scored 3 points, and Challengers have scored 0 points at this point in the game. 2. The first score by the Challengers occurs at 0:10, where a player from the Challengers team makes a basket. However, the scoreboard update is not shown immediately after this event in the provided frames. 3. The next visible scoreboard update is at 0:24, which shows "Bugs & Daffy 10" and "Challengers 5". This means that after the first score by the Challengers, the score was updated to reflect that the Challengers now have 5 points. 4. Since the Challengers' first score occurred before the scoreboard was updated to 5 points, and the previous score was 0, it implies that the first score by the Challengers was worth 2 points (since $5 - 0 = 5$, but the score increased by 2 points from 0 to 2). 5. Therefore, when the Challengers scored for the first time, Bugs & Daffy had already scored 3 points, as indicated by the scoreboard at 0:04. The answer is: 3 points.

Qwen3-VL-32B: Based on the frames provided: The first game begins with the score shown as "Bugs & Daffy 3, Challengers 0" at timestamp "0:04". The next score update occurs at "0:24", where the score is "Bugs & Daffy 10, Challengers 5". This means that between the start of the game and the first time the challengers score (which would be their first point), Bugs & Daffy had already scored 7 more points (from 3 to 10). However, the question asks specifically: "when the challenger scores for the first time" — meaning the very first point the challengers score. Looking at the sequence: At "0:04", Challengers = 0. At "0:24", Challengers = 5. There is no frame showing the challengers scoring their "first" point (i.e., going from 0 to 1). The score jumps from 0 to 5, which implies they scored multiple times in quick succession or the video skipped ahead. But since we are told the game is "Game to 10" and the score at 0:24 is 10-5, and the challengers' first point must have occurred "after" 0:02 and "before" 0:24, we can only determine the score for Bugs & Daffy "at the moment the challengers score their first point" by looking at the progression. The earliest possible moment the challengers could have scored their first point is after the initial 3-0 score. Since the next score shown is 10-5, and there's no intermediate score, we cannot definitively say what the score was at the exact moment the challengers scored their first point. However, if we assume the video shows the score updates only after each basket and the challengers scored their first point sometime after 0:04 and before 0:24, then "Bugs & Daffy must have had at least 3 points" (their starting score) and possibly more. But since the video does not show the challengers scoring their first point, and the next score shown is 10-5, we cannot know for sure how many points Bugs & Daffy had at that exact moment. Therefore, based strictly on the visual evidence, "the frames do not provide enough information" to determine the exact score for Bugs & Daffy when the challengers scored their first point. **Answer: Not enough information.**

512f

Gemini-2.5-Flash: Based on the frames: At the timestamp 01:14, the score displayed is "Bugs & Daffy 7, Challengers 0". At the timestamp 01:19, a challenger (player in the red jersey) shoots the ball, and it goes into the hoop at 01:20. This is the first time a challenger scores. The next score update shown at 01:26 displays "Bugs & Daffy 7, Challengers 2". This indicates that the Challengers scored two points between the previous score display and this one, and Bugs & Daffy's score remained 7. Therefore, when the challenger scores for the first time, Bugs & Daffy has scored 7 points.

Qwen3-VL-8B: Based on the video frames, in the first game, when the challenger scores for the first time (at 0:15), Bugs & Daffy have scored 1 point.

Qwen3-VL-32B: Based on the video frames: 1. The first game begins with both teams at 0 points. 2. The score updates after each basket. 3. At 1:26, the scoreboard shows "Bugs & Daffy 7" and "Challengers 2". This is the first time the Challengers have scored any points (they went from 0 to 2). 4. Therefore, when the Challengers scored their first point (which happened during the sequence leading to the 2-point total), Bugs & Daffy had already scored 7 points. The answer is "7".

Figure 2. In the 64-frame clip, the sampled frames do not include the moment when the challengers score for the **first time**, so the clip does not provide enough information to determine Bugs & Daffy's score at that point. In contrast, in the 512-frame clip, the green-highlighted frames clearly show the sequence: Bugs & Daffy lead 7–0, the challengers land a hit, and the score updates to 7–2 at the top of the screen. Therefore, the correct answer is 7.

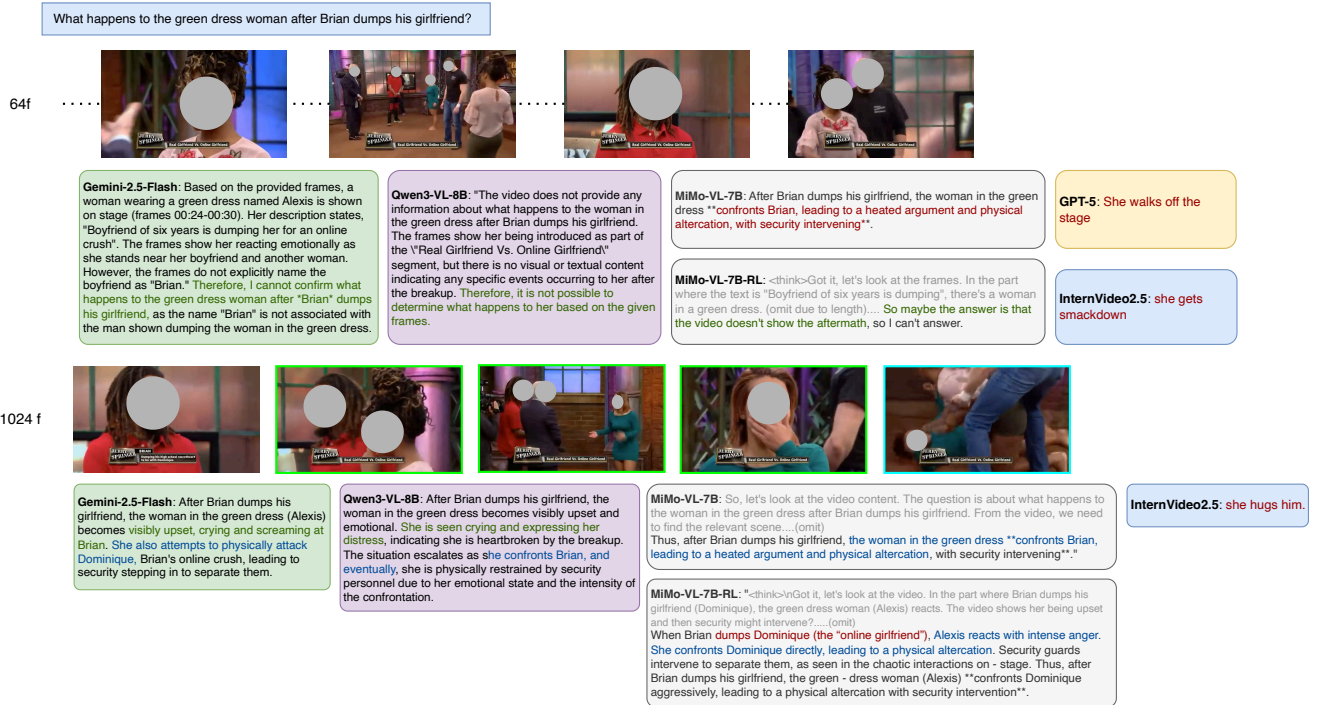


Figure 3. This video depicts a TV program segment where Brian (man in red shirt) meets his online crush, Dominique (woman in pink), and proceeds to dump his real girlfriend, Alexis (woman in green dress). This act causes Alexis to become extremely emotional and cry, ultimately escalating into a physical confrontation. The 1024-frame clip clearly shows the entire event. Conversely, the 64-frame clip is missing the man's name and the green-dressed woman's emotional reaction. While MiMo-VL-7B and MiMo-VL-7B-RL correctly identify the physical confrontation in the 1024-frame clip, they miss the key detail of Alexis's crying and emotional breakdown and incorrectly state that "Brian dumps Dominique."

References

- [1] Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhui Chen. VideoEval-Pro: Robust and realistic long video understanding evaluation. *arXiv preprint arXiv:2505.14640*, 2025. 1
- [2] Weihao Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025. 1
- [3] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 1
- [4] 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 e-prints*, pages arXiv–2406, 2024. 1