FPS-Bench: A Benchmark for High Frame-Rate Video Understanding

Supplementary Material

This supplement contains additional annotation details and experiments that were alluded to in the main text. We begin by providing significantly more detailed descriptions of each question category in FPS-Bench, how they differ and the main reasoning capabilities they test. We then describe some additional analysis we conducted on FPS-Bench to understand where its difficulty comes from, along with ablations on the prompt we used in our model evaluation harness. Finally, we include several representative visual examples spanning all the question categories to help readers better understand the composition of the benchmark.

1. Question Category Details

We provide detailed descriptions of each category and provide further details on how they differ and what capabilities they test.

Blink & Miss. These questions target events that are extremely brief and visually localized, often observable in only one or two frames, yet are identifiable without needing to follow motion across time. These events do not rely on temporal continuity; a single frame containing the event is generally sufficient for a human to confirm its occurrence. Examples include a camera flash, a blink, or a collision between objects. The key challenge is temporal sparsity: because the event may appear in only a tiny fraction of frames, low-FPS sampling can easily skip it entirely. Thus, the difficulty arises not from interpreting motion, but from ensuring enough temporal resolution to capture the fleeting frame in which the event appears.

Fine-Grained Motion. These questions require understanding subtle, temporally extended motion patterns that cannot be inferred from any single frame. These tasks involve nuanced differences in trajectory, form, or execution, such as distinguishing between similar martial-arts techniques, evaluating the style of a dive, or identifying a specific variation of a gymnastics flip. Here, temporal continuity is essential: the defining features of the action only emerge when observing how the subject moves over time. Errors occur not because frames are missing entirely, but because coarse temporal sampling collapses motion cues, making distinct movements appear visually identical when undersampled.

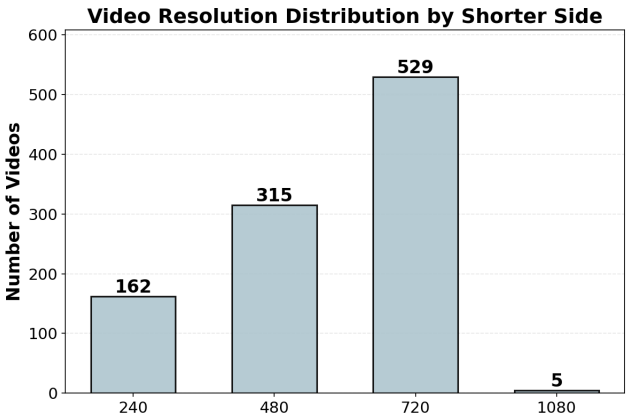
Instance Count. These questions involve counting short, discrete events that are each visually identifiable in isolation but require sufficient temporal resolution to distinguish successive occurrences. While a single frame is enough to recognize the event itself, reliably counting multiple instances also depends on observing the transition back to the baseline or "non-event" state between them. For example, when counting camera flashes, it is necessary to see not only each flash appearing but also the brief return to darkness that separates one flash from the next. If the sampling rate is too low, these intermediate states may be skipped, causing distinct events to merge and leading to systematic undercounting. As a result, the challenge stems not from interpreting motion, but from preserving enough temporal granularity to resolve individual event boundaries.

Repetitive Motion. These questions ask annotators to count cycles of a sustained, periodic action whose identity emerges only through motion, such as tallying the number of flips in a gymnastics sequence, brush strokes, or fan blade turns. Unlike Instance Count, these cycles are not visually atomic; a single frame cannot indicate whether a cycle has occurred. Instead, the defining unit is a motion trajectory or full action sequence, making the task inherently motion-dependent. Undersampling blurs or aliases the motion cycle, often causing different phases of the motion to appear indistinguishable or even inverted, which leads directly to miscounts. High FPS is therefore required both to preserve the periodic motion pattern and to accurately discriminate cycle boundaries.

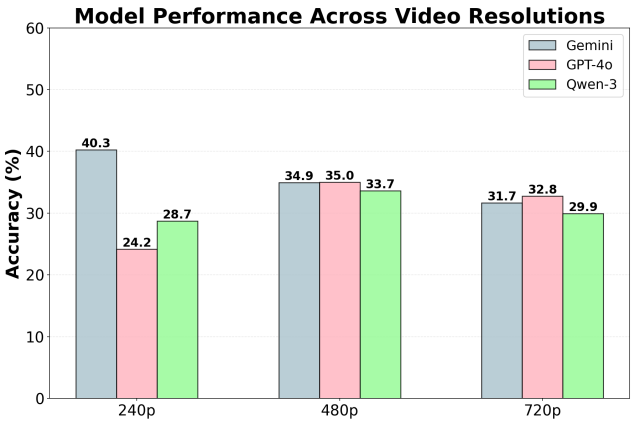
Speed Recognition. These questions require inferring the change in speed of on object or comparing the relative velocity of objects where the change or difference is only perceptible when observing motion at a sufficiently high frame rate. These tasks rely on temporal continuity: subtle variations in acceleration or traversal speed often appear visually similar when undersampled, leading faster and slower motions to become indistinguishable. High FPS sampling preserves the fine-grained temporal cues needed to judge pacing, periodicity, or rate of movement, enabling accurate speed discrimination.

Action Order. These questions require determining the temporal sequence of two or more events that occur nearly simultaneously or in rapid succession. Because the relevant events may be separated by only a few frames, low FPS sampling can either merge them into a single perceived event or invert their apparent order. High temporal resolution is therefore essential for preserving event boundaries and allowing annotators to pinpoint which action occurred first, especially when interactions happen at high speed or with minimal temporal gap.

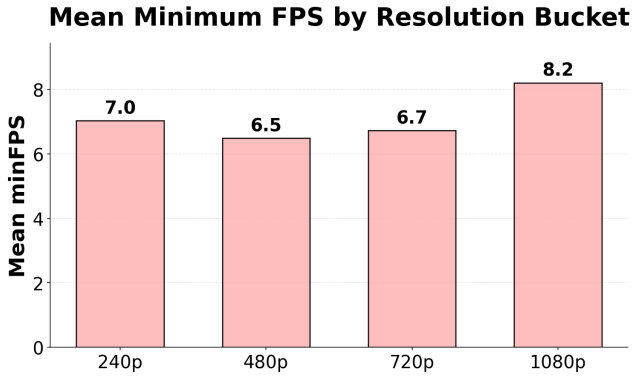
State at Event These questions focus on identifying the precise configuration, pose, or attribute of an object at the moment of a brief, transient interaction. Unlike Blink & Miss, where the challenge is detecting that a fleeting event



(a) **Resolution Distribution.** Most videos in FPS-Bench are 720p, but a significant fraction are 480p and 240p. There are a very small number of 1080p videos in the benchmark.



(b) **Accuracy vs Resolution.** There is no clear relationship between model performance and input video resolution, showing that FPS-Bench's difficulty is from its demanding motion understanding requirements.



(c) **MinFPS vs Resolution.** When measuring the average minFPS across different resolution buckets, there is no clear relationship, indicating that tasks are equally difficult despite their resolution.

Figure 1. **Analyzing the effect of video resolution**. (a) Distribution of video resolutions in the benchmark dataset. (b) Model accuracy across different input resolutions shows no clear correlation with resolution quality. (c) Comparing minFPS to video resolution shows no relationship between task difficulty and resolution.

occurred at all, the event here is typically easy to perceive; what is fleeting is the specific information needed to answer the question. The decisive frame expressing the required state may last only an instant, and coarse temporal sampling can easily skip or blur this moment, even though the broader event remains visible. High FPS sampling is therefore essential to capture the exact state at the critical instant and to reliably verify the intended attribute.

Causality Detection questions require determining the precise temporal relationship between two events to identify which one directly triggered or led to the other. These tasks depend on capturing subtle leading–lagging dynamics: the cause must be observed occurring just before the resulting effect, often within only a few frames. Undersampling can blur or collapse this ordering, making the initiating event appear simultaneous with, or even later than—the effect. High FPS sampling preserves the fine temporal offsets needed to infer causality and to distinguish genuine cause–effect links from coincidental co-occurrences.

Synchronization Assessment questions focus on evaluating whether one or more actions occur in unison or with slight temporal offsets. The challenge lies in resolving very small differences in timing across parallel motion streams. For example, determining whether two people dribble a basketball simultaneously or whether two machine parts cycle in perfect sync. At low sampling rates, these actions can appear artificially aligned or misaligned due to aliasing, obscuring their true temporal relationship. High FPS input ensures that the relative phase and timing of each action are faithfully captured, enabling accurate synchronization judgments.

Models	RM	IC	SR	FM	AO	BM	SE	CD	SY	Overall
Random	20.0%	20.0%	20.0%	20.0%	20.0%	20.0%	20.0%	20.0%	20.0%	20.0%
Qwen-3-8B-instruct	18.9%	8.0%	34.9%	29.6%	37.6%	34.0%	26.5%	37.4%	25.9%	28.0%
Qwen-3-32B-instruct	29.7%	18.8%	33.9%	32.4%	39.4%	33.0%	25.5%	30.6%	33.0%	30.7%
InternVL-3.5-8B	19.8%	15.2%	31.2%	29.6%	34.9%	32.0%	30.3%	34.3%	32.1%	28.7%
InternVL-3.5-14B	20.7%	14.3%	32.1%	29.6%	42.2%	28.0%	23.2%	30.6%	30.3%	27.9%
OmniVinci	21.6%	17.0%	33.0%	27.8%	33.9%	31.0%	29.6%	34.6%	28.7%	28.5%
LLaVA-OneVision	21.6%	14.3%	30.3%	28.3%	33.9%	33.3%	25.8%	29.6%	26.6%	27.0%
LLaVA-NeXT-Video	16.2%	8.9%	29.4%	28.3%	37.6%	23.0%	35.1%	26.9%	34.9%	26.5%
LLaVA-1.5	19.6%	17.2%	16.4%	33.3%	35.8%	33.0%	23.6%	30.0%	35.1%	27.1%
Oryx	15.3%	11.6%	42.2%	25.0%	48.6%	34.0%	34.3%	37.4%	34.3%	31.3%
Deepseek-tiny	13.5%	16.1%	24.8%	22.2%	31.2%	27.3%	31.3%	24.3%	32.4%	24.6%
Deepseek-small	19.8%	10.7%	34.9%	24.1%	33.9%	27.3%	25.3%	33.6%	27.8%	26.3%
Deepseek-base	12.6%	7.1%	30.3%	24.1%	32.1%	20.2%	31.3%	28.0%	31.5%	24.0%
Gemini-2.5-Pro	29.1%	22.3%	33.9%	27.8%	32.7%	29.7%	28.3%	28.0%	28.3%	28.9%
GPT-4o	34.2%	32.1%	25.7%	25.0%	35.8%	33.0%	37.4%	39.8%	23.9%	31.8%

Table 1. **Evaluation Results with "None of the Above" Option.** Performance of VLMs when "None of the Above" is included as an additional option, reducing random chance to 20%. Cells colored in red indicate decreased performance, where the model is less certain, and cells colored in green indicate increased performance.

2. Additional Results

Effect of Resolution. One possibility for FPS-Bench's difficulty could be that the videos are low-resolution. Labelers for FPS-Bench often reported that labeling videos with low-resolutions was harder, and anecdotally they might require higher minFPS. We evaluate this with a set of measurements comparing the relationship between accuracy, minFPS and resolution. The results in Figure 1 demonstrate that this is clearly not the case. Firstly, we show the distribution of video resolutions in Figure 1a. Most of the videos are 720p, but a significant fraction is 480p and 240p as well. However, Figure 1c shows that there is no clear relationship with minFPS and resolution: lower resolutions do not have notably lower minFPS than higher resolutions. Furthermore, Figure 1b shows that the accuracy does not degrade with lower resolution. These results provide strong evidence that FPS-Bench's difficulty is due to the inherent difficulty of high-frame-rate understanding rather than low-resolution samples.

Model Uncertainty. Multiple-choice benchmarks often include a "None-of-the-above" option for models to express uncertainty. We evaluated all models in the main text on a version of FPS-Bench with this additional option, the results of which are shown in **??**. None-of-the-above is *never* the answer to a question, so the change in performance allows us to understand which categories models are least or most certain about. We observe that all the evaluate models are significantly less certain about categories involving counting, like *instance count* and *repetitive motion*. *Causality detection* is also challenging: including an uncertain option reduces performance. However, most other categories actually see an improvement. Surprisingly, models tend to get an overall higher accuracy with a none-of-the-above option present, despite the fact that none-of-the-above is never the correct answer.

Model	System Prompt	Accuracy
Qwen-3 [1]	ALLVB	32.80%
	LongVideoBench [3]	31.60%
	Video-MME [2]	32.90%
	Ours	**32.40%**
InternVL	ALLVB	31.50%
	LVBench	30.10%
	VideoMME	31.90%
	Ours	**32.00%**

Table 2. System prompt ablation results comparing different prompt styles across models. Our prompt is comparable with that of Video-MME.

3. Prompts and Prompt Ablation

The prompt is an important part of VLM evaluation , and the prompt used for the evaluation harness has a strong effect on the downstream result. We measure the performance of three different widely used open-source evaluation harnesses: Video-MME [2], LongVideoBench [3] and ALLVB. We found that the Video-MME prompt generally worked well, but models would often choose an option that was not provided, such as *UNKNOWN*. We added extra language to encourage the model to only pick among the provided options to fix this issue. The prompts from each setup, along with ours are provided below.

Evaluation Prompt - Ours

```
Please analyze the video frames and
answer the following multiple choice
question.
Question: [Question]
Options: [Options]
IMPORTANT: Your response MUST start
with ONLY the letter of your answer
([available letters]), followed
by a period or space, then your
explanation.
Format your response exactly like
this:
A. (your explanation here)
```

Evaluation Prompt - Video-MME

```
Select the best answer to the
following multiple-choice question
based on the video. Respond with only
the letter ([available letters]) of
the correct option.
[Question]
[Options]
```

Evaluation Prompt - LVBENCH

```
[Question]
[Options]
Please select the best answer from
the options ([available letters])
above and directly provide the letter
representing your choice without
giving any explanation.
```

Evaluation Prompt - ALLVB

```
You are an expert at analyzing videos
and their accompanying subtitles.
Carefully observe the details in the
video frames.
Based on your observations, select the
best option for the question provided.
Question: [Question]
[Options]
Please strictly provide the letter
representing your choice.
```

4. Visual Examples

We include several samples from FPS-Bench, beginning on the next page, across a range of categories to better visualize what different tasks constitute.

Blink & Miss

What facial gesture does the blonde man make?

A. He blinks
B. He widens his smile
C. He winks
D. He raises both eyebrows

He winks

Action Order

What order of attacks does the character on the left perform?

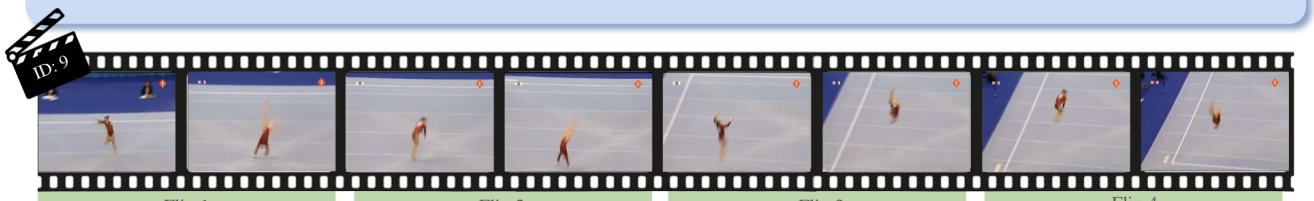
A. lowknick, knee, punch, punch, roundhouse kick
B. punch, punch, knee, lowkick, roundhouse kick
C. low kick, punch, punch, knee, roundhouse kick
D. low kick, punch, knee, punch, roundhouse kick

Low kick | Punch | Punch | Knee | Roundhouse Kick

Repetitive Motion

How many rotations does the gymnast do?

A. 2
B. 3
C. 4
D. 5

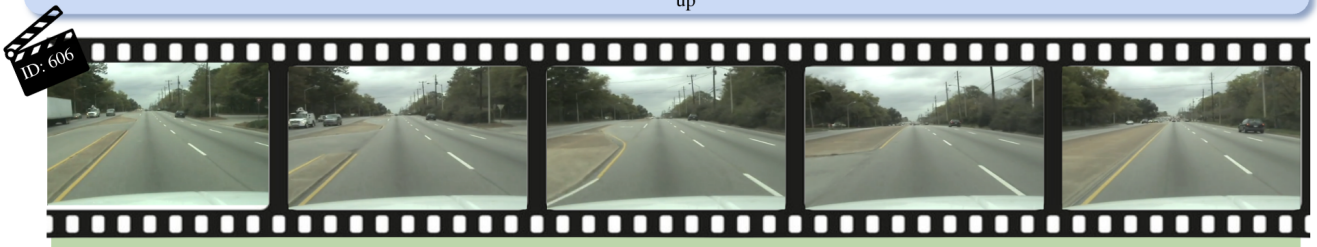


Flip 1 | Flip 2 | Flip 3 | Flip 4

Speed Recognition

How does the speed of the car change throughout the video?

A. The car stays at a constant speed
B. The car slows down
C. The car slows down, then speeds up
D. The car speeds up

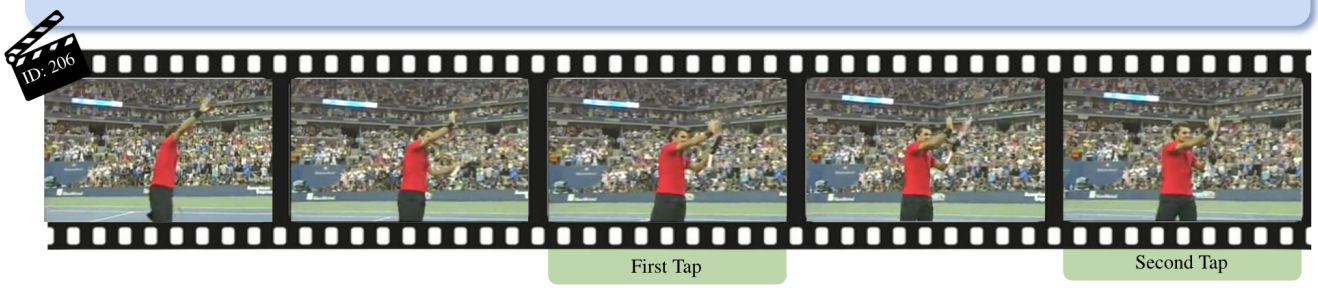


Car accelerates throughout video clip

Instance Count

How many times did he hit the racquet with his palm?

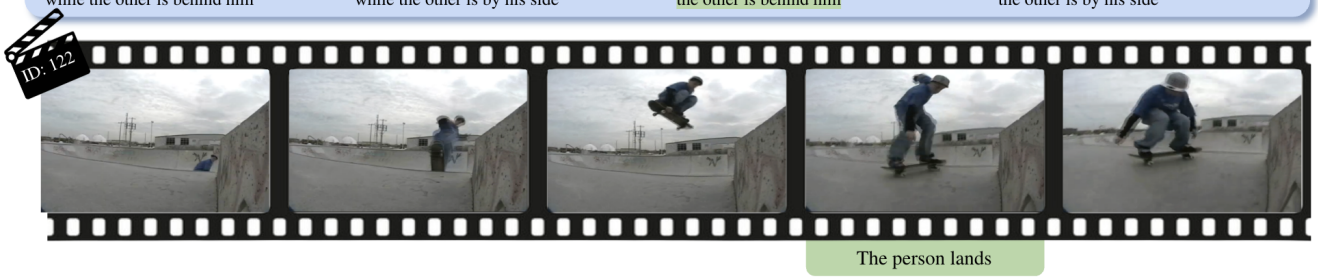
A. He did not hit the racquet | B. 1 | C. 2 | D. 3



State at Event

How are the person's hands positioned when he lands?

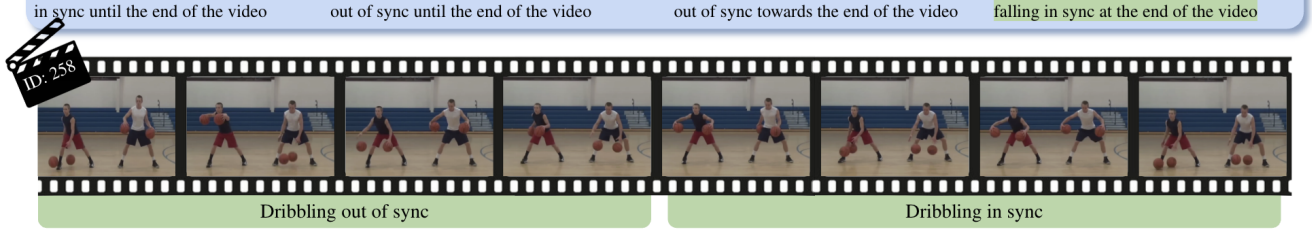
A. One hand is grabbing the skateboard, while the other is behind him | B. One hand is grabbing the skateboard while the other is by his side | C. One hand is reached downwards while the other is behind him | D. One hand is reached downwards while the other is by his side



Synch. Assessment

Do the two people begin dribbling in sync? What about towards the end of the video?

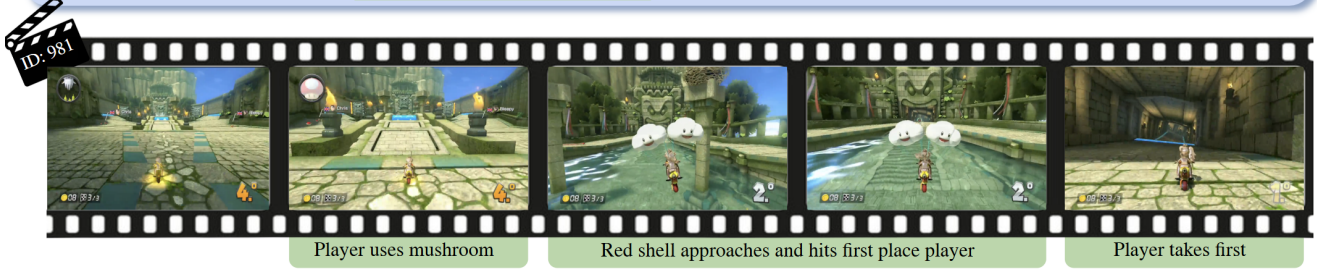
A. They begin dribbling in sync, staying in sync until the end of the video | B. They begin dribbling out of sync, remaining out of sync until the end of the video | C. They begin dribbling in sync, falling out of sync towards the end of the video | D. They begin dribbling out of sync, falling in sync at the end of the video



Causality Detection

How does the player manage to move to first place?

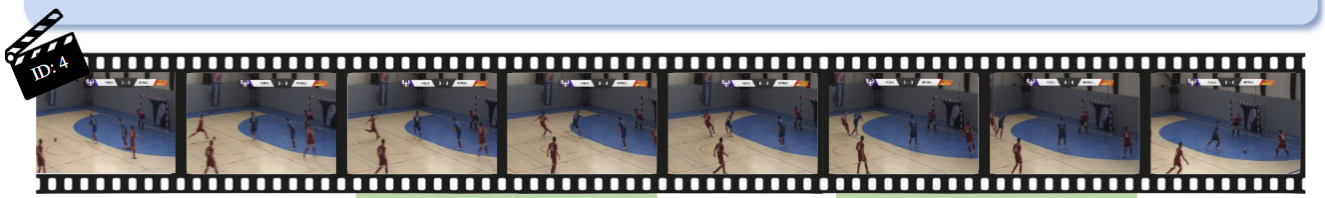
A. He uses 3 mushrooms, which gives him a speed boost that gets him to second place. Then, the person in first place is hit by a red shell, allowing the player to pass | B. He uses a mushroom, which gives him a speed boost that gets him to second place. Then, the person in first place is hit by a red shell, allowing the player to pass | C. He uses a mushroom, which gives him a speed boost that gets him to second place. Then, the person in first place is hit by a green shell, allowing the player to pass | D. He uses 3 mushrooms, which gives him a speed boost that gets him to second place. Then, the person in first place is hit by a green shell, allowing the player to pass



Fine-Grained Motion

What foot does the player shoot with, and does he score?

A. Right foot, no he does not score B. Left foot, no he does not score C. Right foot, yes he does score D. Left foot, yes he does score



shoots the ball with his left foot ball bounces off the goal post

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[2] Chaoyou Fu, Yuhan Dai, Yongdong 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 3

[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. 3