Video Panels for Long Video Understanding

Supplementary Material

6. Dataset Details

- **VideoMME** [10] (VMME). A multi-modal evaluation benchmark that spans across 6 different domains (Knowledge, Film & Television, Sports Competition, Artistic Performance, Life Record, and Multilingual). Video durations range between 11 seconds to 1 hour, and are categorized as *short* videos (80 seconds on average), *medium* (500 seconds) and *long* (2500 seconds).
- **TimeScope** [49] inserts multiple short video clips ("needles") into long videos. These needles contain the key information required to answer the questions. As opposed to the usual Needle In A Haystack (NIAH) evaluation, this forces models to understand the whole video. Videos are divided into 13 different lengths, ranging from 60 to $36,000$ seconds. We divide the evaluation into `short` (up to 3 hours, average 2590 seconds) and `long` (between 5 and 10 hours, average $27,600$).
- **MLVU** [47] covers a wide range of video lengths, from 3 minutes to 2 hours, with an average length of 15 minutes. It includes both real-world videos (*e.g.*, egocentric, movies) and simulated videos (*e.g.*, video games, cartoons). The benchmark provides multiple-choice and open-ended QA tasks across 9 different categories. We evaluate on the *dev* set, which consists of 2593 QA pairs over 1730 videos.
- **MF2** [41] evaluates comprehension and recall of key narrative information from full-length movies (50-170 minutes). It includes 53 complete movies with an average duration of 88.3 minutes. Unlike other benchmarks, the task is to discriminate between true and false claims across 850 claim-pairs. The claims span five categories: character motivations and emotions, memorable moments, causal chains, and event order.
- **VNBench** [46] (VNB) targets three aspects of video understanding: temporal perception, chronological ordering, and spatio-temporal coherence. It includes tasks such as retrieval, ordering, and counting, and consists of 1350 samples ranging from 10 to 180 seconds, collected from 150 videos.

7. Full Results for Failure Analysis

We present the full results that have been discussed in Section 4.7, using the Qwen2.5 7B model with 32-frame input on the MLVU dataset in Tab. 5. As discussed in the main paper, panels outperform the standard representation in all tasks, except anomaly recognition, where they perform equally, and ordering, where performance drops by 1.2%. This highlights a limitation in the zero-shot panels representation, as temporal information is not explicitly encoded. In the tasks where panels improve the results, the gain is typically around 1-4%. For the counting task, the performance increases from 23.3% to 39.8%. This demonstrates that panels are also effective in tasks requiring spatial detail.

8. Adding Explicit Temporal Encoding

We conduct an experiment to explore whether incorporating some form of temporal encoding to video panels increases overall performance. We implement a temporal order encoding in two ways based on NumPro [36], by overlaying the frame index (panels-F) and the frame time (panels-T) on the panels. We evaluate these methods against the base video panels without explicit temporal encoding in Tab. 6. The results show that adding explicit temporal information improves performance on temporal tasks such as answering questions about order. Adding the frame index (panels-F), however, reduces performance for some other types, such as counting. While adding the frame index to the panels overall decreases the performance, adding the frame time slightly increases the performance on MLVU. Thus, explicitly including temporal order is beneficial for specific question types and can be used when required by a specific task.

9. Prompt

Our main results show that existing VLMs are already capable of interpreting the paneled images without any additional information in the prompt. Nonetheless, adding additional directions to the textual prompt can, in some cases, further improve results. We show results for LLaVA-OV 7B and Qwen2.5-VL 7B with three different prompts:

- **Prompt 1** (P1): "You are given a sequence of images. Each image is a composite grid of video frames arranged in temporal order: panels are ordered from left to right, then top to bottom — like reading a book. Within each composite, the panels represent consecutive frames from the video. Across the sequence, the composites are shown in chronological order. When answering, interpret the full temporal sequence, not individual panels in isolation." The prompt is added before the question.
- **Prompt 2** (P2): "When answering, treat the panels as frames from one video, in order from left to right, then top to bottom." The prompt is added before the question.
- **Prompt 3** (P3): "Each image is divided into {r} rows and {c} columns of panels. Read them in left-to-right top-to-bottom order as consecutive video frames. Answer with the option's letter from the given choices directly."

	Needle	Count	Ego	Topic Reas.	PlotQA	Anomaly Recogn.	Order
Qwen2.5-7B	71.0	23.3	55.7	85.6	63.1	**72.0**	**49.8**
+ ours	**72.1**	**39.8**	**56.8**	**89.0**	**65.3**	**72.0**	48.6

Table 5. **Analyzing failure cases on MLVU.** We report results for Qwen2.5 7B with 32 frames as input.

	Count	PlotQA	Order	Overall
Qwen2.5-7B	23.3	63.1	49.8	60.1
+ panels	39.8	65.3	48.6	63.4
+ panels-F	35.4	64.9	50.6	62.4
+ panels-T	41.3	65.1	50.6	63.9

Table 6. **Effect of adding explicit temporal information.** We report results for Qwen2.5 7B with 32 frames on MLVU.

	No prompt	P1	P2	P3
LLaVA-OV 7B	58.9	60.1	59.4	58.8
Qwen2.5-VL	62.4	61.9	61.8	62.9

Table 7. **Effect of changing the prompt.** We report results of using additional prompts on VMME using LLaVA-OV 7B and Qwen2.5-VL.

The prompt is added after the question. {r} and {c} are replaced by the number of rows and columns.

As can be seen from Tab. 7, even among these models, there is no consistently best prompt. However, with model-specific prompts, performance can get another boost, such as Prompt 1 for VMME with LLava-OneVision, and Prompt 3 for VMME with Qwen2.5-VL. While model-specific prompts improve the results further, we do not include them for the results in the paper since our focus is on a model-agnostic approach.

10. Qualitative Results

Figure 6 shows an example from the VideoMME benchmark using LLaVA-Video 7B as the VLM. The question asks what the protagonist does after *feeding the ducks* and *riding the bike*. Without our paneling strategy, the model processes only a sparse set of frames and observes only the 'feeding the ducks' portion, entirely missing the segment where the person rides the bike. In contrast, our Video Panels capture both events by providing finer temporal coverage, allowing the model to correctly answer the question. We also present an example of the challenging Needle In A Haystack (NIAH) task in Figure 7, where the goal is to answer the question based on a very short, discriminative clip embedded within a much longer video. In this example, the question concerns a man walking along the shore. Without our paneling strategy, the crucial frames containing the target event are never passed to the VLM, preventing it from retrieving the correct evidence. In contrast, Video Panels ensures that these sparse but essential frames are included, enabling an accurate answer.

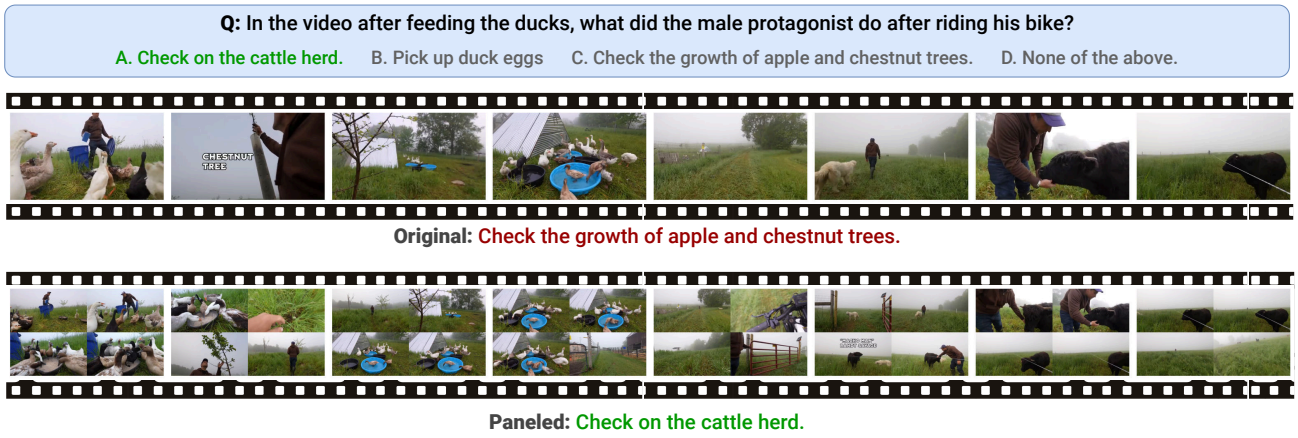

Figure 6. **Qualitative example on VideoMME.** We use LLaVA-Video 7B as the VLM. Without panels, the relevant information to answer the question (frames with the bike) is absent.

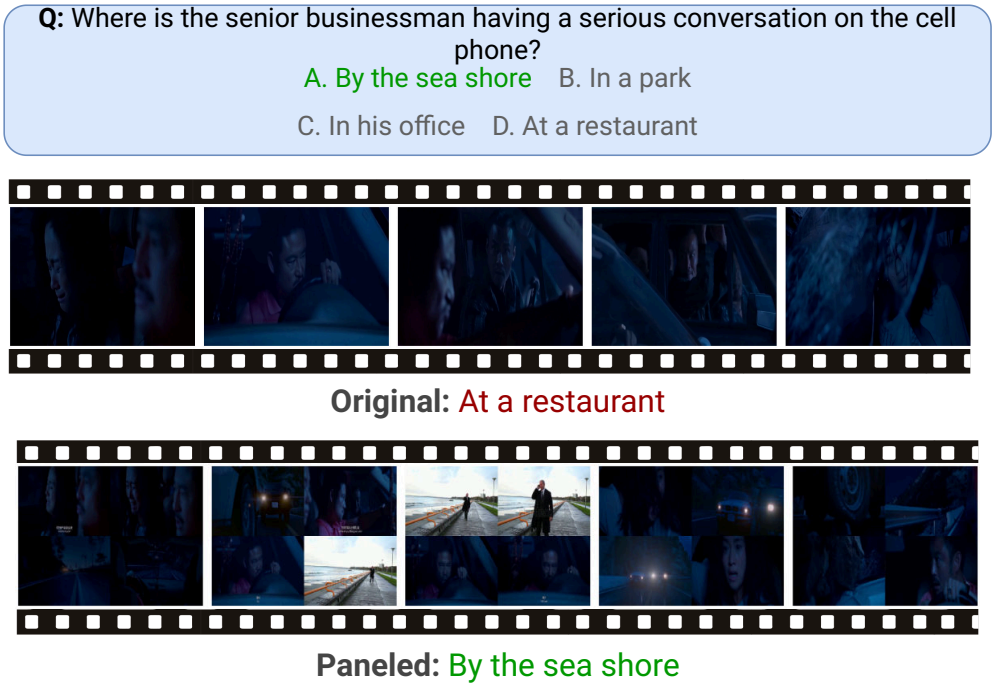

Figure 7. **Qualitative example on MLVU.** We use LLaVA-OV 7B as the VLM for the Needle In A Haystack task. Without panels, the relevant 'needle' to answer the question is absent.