Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis

Supplementary Material

7. Detailed Experimental Settings

Models. We conduct a comprehensive evaluation on four commercial models and nine representative open-source video-based multimodal large language models. To further demonstrate the adaptability of our benchmark to multi-image scenarios, we also include three widely utilized image-based MLLMs as part of the evaluation. The complete list of models evaluated is provided below.

- **Commercial MLLMs**: GPT-4V, GPT-4o, Gemini 1.5 Flash, and Gemini 1.5 Pro.
- **Open-source Video MLLMs**: Video-LLaVA, ST-LLM, ShareGPT4Video, VideoChat2-Mistral, VILA-1.5, Chat-UniVi-V1.5, VITA-1.0, VITA-1.5, LLaVA-NeXT-Video.
- **Advanced Image MLLMs**: Qwen-VL-Chat/Max and InternVL-Chat-V1.5.

Frame Extraction. A standard approach involves extracting a sequence of frames from the video and interpreting the resulting multi-image inputs. For Gemini 1.5 Pro which supports extremely long multimodal contexts, we sample frames at 1 frame per second for short and medium videos, and at 1 frame every 2 seconds for long videos to ensure API stability. For all other models, frame extraction adheres to their respective official guidelines, uniformly sampling a specified number of frames from the video. The specific numbers of sampled frames are as follows: 10 frames for GPT-4V, 384 for GPT-4o, 8 for Video-LLaVA, 16 for VideoChat2-Mistral, 16 for ShareGPT4Video, 64 for ST-LLM, 64 for Chat-UniVi-V1.5, 32 for VITA-1.0, 16 for VITA-1.5, 32 for LLaVA-NeXT-Video, 8 for VILA-1.5, 4 for both Qwen-VL-Chat and Qwen-VL-Max, 10 for InternVL-Chat-V1.5.

Subtitle Utilization. In the subtitle-enabled setting, all models utilize subtitles corresponding to the timestamps of the sampled video frames. For instance, if 10 frames are sampled from a video, the 10 subtitles that correspond to the respective timestamps of those frames are selected. This approach ensures accurate coherent synchronization between the visual and textual multimodal input for evaluation.

Evaluation. The evaluation follows the format: "entire video frames + complete subtitles/audios (optional) + question with prompt." Whenever possible, the model's default prompt is utilized for multiple-choice questions. If unavailable, a standardized prompt is employed as follows:

> *This video's subtitles are listed below: [Subtitles] Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option. [Question] The best answer is:*

The accuracy is computed by extracting the model's output using regular expressions and comparing it directly with the ground-truth answer, without relying on external judges like commonly-used ChatGPT.

8. Additional Analysis

How do MLLMs perform on the two highlighted cases in Figure 1? We conduct qualitative evaluation (using frames and subtitles) on the two cases in Figure 1. As analyzed in Section 3.2, these two cases comprehensively examine the model's capabilities in OCR, attribute perception, object recognition, and long-range temporal reasoning, making them highly challenging. **For the date-related question in Case 1**, Video-LLaVA identifies the date (May 31st) from the frame at 01:10 and subtitles, but fails to perform reasoning based on context and incorrectly determines the year of the event, leading to the erroneous selection of option A. The remaining open-sourced models miscalculate the date 10 days after May 31st during the reasoning process, resulting in the incorrect choice of option C. **For the event-related question in Case 2**, Video-LLaVA, VideoChat2, and ST-LLM incorrectly associate the target person with nearby events, resulting in the selection of incorrect options A or C. In contrast, LLaVA-NeXT-Video and Gemini 1.5 Pro accurately track the events involving the target individual across the entire video, showcasing robust long-range temporal modeling capabilities. They correctly link the target person's injury at 03:35 with his reappearance at 27:30, identifying the true cause of the injury (option D). In summary, the questions in our benchmark pose significant challenges to the models, which motivates MLLMs to advance both their perception and reasoning capabilities.

Could additional modalities benefit the performance across categories? Figure 4 presents the results of Gemini 1.5 Pro across the 30 subcategories of Video-MME, under the testing modes of frames, frames + subtitles, and frames + audio. The results indicate that subtitles and audio positively contribute to video understanding in multimodal large models. However, the extent of improvement provided by these modalities varies across different domains.

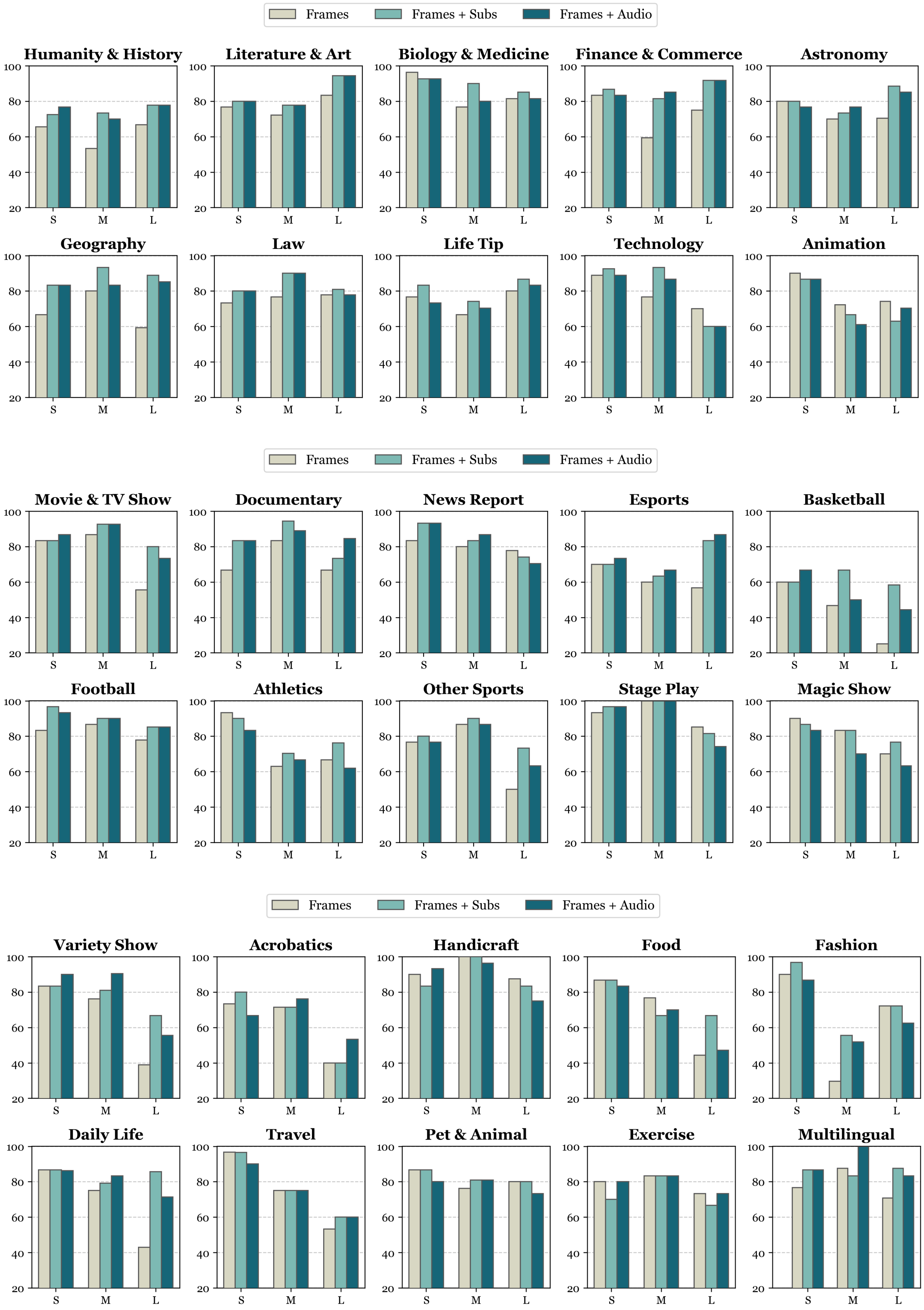

Figure 4. Evaluation results of Gemini 1.5 Pro across different video subcategories.