Video-MMLU: A Massive Multi-Discipline Lecture Understanding Benchmark

Enxin Song [1] Wenhao Chai [2] Weili Xu[1,3] Jianwen Xie [4] Yuxuan Liu [3] Gaoang wang [1]

[1]Zhejiang University [2]Princeton University [3]University of Illinois Urbana-Champaign [4]Lambda, Inc.

Contributions

Video-MMLU pushes LMMs to the limits—can the model really understand real-world lectures?

- We build **Video-MMLU**, which requires strong reasoning capabilities and world knowledge compared to the previous benchmarks for video LMMs.
- We evaluate more than 90 proprietary models and open-source models of varying sizes on VideoMMLU. Our findings indicate that existing models generally perform poorly, with accuracy ranging from only 10% to 50%.
- We explore how the number of visual tokens and the base LLMs influence performance, offering insights into the interplay between multimodal perception and reasoning in lecture comprehension.

Benchmark Statistics

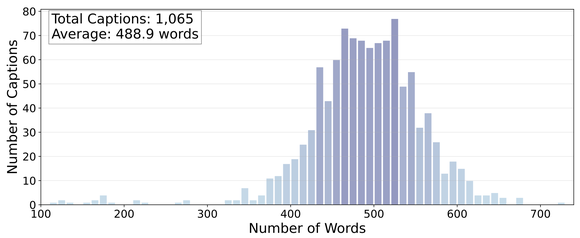

Figure 1. Video detailed captions length distribution.

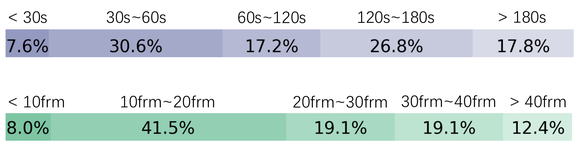

Figure 2. Video length and keyframes number distribution.

Overview

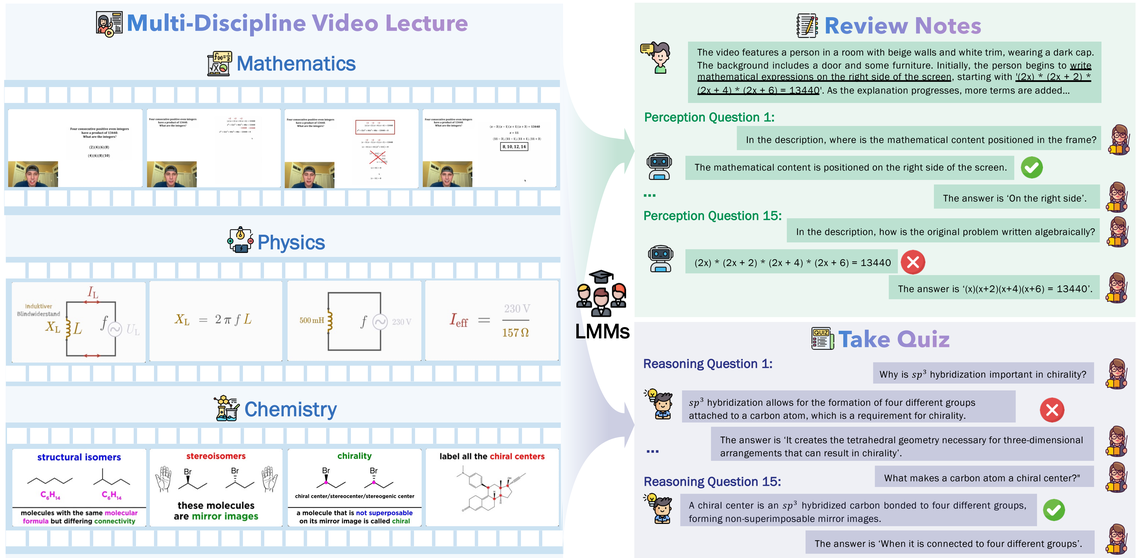

Figure 3. The benchmark includes multi-discipline lecture videos in mathematics, physics, and chemistry, featuring theorem demonstrations and problem-solving. Evaluation consists of: (1) **Review Notes**, where models generate detailed video captions to assess visual perception, and (2) **Take Quiz**, where models answer reasoning questions to test comprehension.

Influence of model size

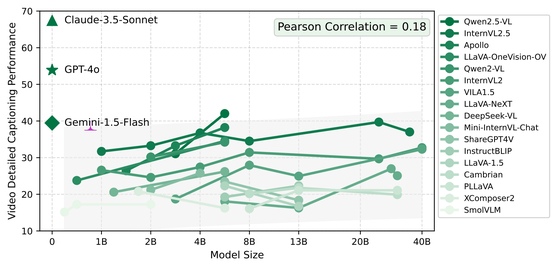

Figure 4. **Relationship between model size and video captioning performance.** The shaded region shows the confidence interval.

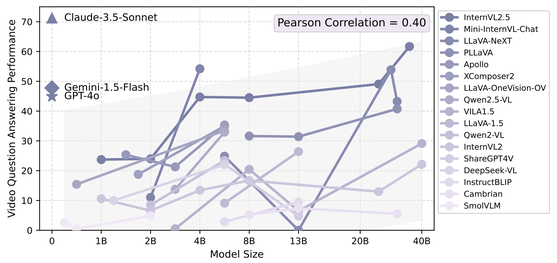

Figure 5. **Relationship between model size and video QA performance.** The shaded region shows the confidence interval.

Influence of LLM Architecture

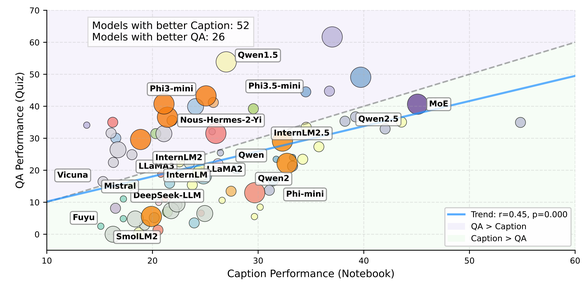

Figure 6. Relationship between captioning and QA across LLM.

Ability of Token Compression Models

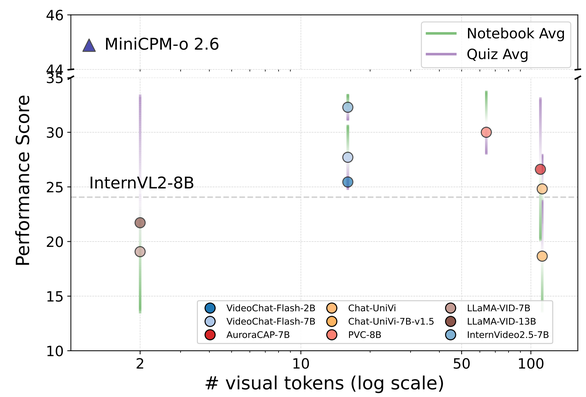

Figure 7. Impact of vision tokens number.

Our findings indicate that significant token reduction is feasible while maintaining or even surpassing the performance of the base model. Despite efficiency gains, a substantial gap remains between token-compressed models and the state-of-the-art, indicating challenges in preserving fine-grained details essential for complex lecture reasoning with existing token-compressed models.