BOLT: Boost Large Vision-Language Model Without Training for Long-form Video Understanding
[Supplementary Material]

Shuming Liu Chen Zhao Tianqi Xu Bernard Ghanem
King Abdullah University of Science and Technology (KAUST)

In this appendix, we provide additional experimental results and further analyses. Specifically, we present results on Video-MME with subtitles in Section A. Then, we discuss inference costs in Section B. In the end, we provide additional visualization examples in Section C.

A. Additional Results on Video-MME

In the main paper, we presented benchmark results of various VLMs on the Video-MME dataset without subtitles. To further validate the effectiveness of the proposed method, we incorporate subtitles into the VLMs' text input. As shown in Table 1, our method consistently improves overall performance across different frame budgets. Particularly, with only 8 input frames, BOLT increases the accuracy from 58.7% to 61.0%. The performance across short, medium, and long videos also improves, demonstrating the effectiveness and robustness of our approach in leveraging both visual and textual information.

Table 1. **Benchmark results on Video-MME dataset with subtitles.** Our proposed inverse transform sampling can consistently enhance the overall performance under different frame budgets.

Model	BOLT	Overall	Short	Medium	Long
LLaVA-OneVision$^{8\text{ frame}}$	✗	58.7	70.8	56.1	49.4
	✓	**61.0**	**71.7**	**58.4**	**52.9**
LLaVA-OneVision$^{16\text{ frame}}$	✗	60.3	72.7	57.1	51.1
	✓	**61.7**	**74.0**	**59.3**	**51.8**
LLaVA-OneVision$^{32\text{ frame}}$	✗	61.9	**75.7**	58.4	**51.6**
	✓	**62.7**	75.1	**61.4**	51.4

B. Analysis of Inference Cost

We also evaluate the inference cost of our training-free approach. Our method consists of three main steps: CLIP-based frame feature encoding, inverse transform sampling, and VLM inference. In terms of memory usage, our approach requires nearly the same GPU memory as the baseline method that uses uniform sampling. In terms of inference time, as shown in Table 2, the total inference time per sample increases by approximately 90%, primarily due to the CLIP feature encoding step. In contrast, the inverse transform sampling itself is highly efficient. Although our method introduces some additional inference time, it remains acceptable considering that no training or fine-tuning is required.

Table 2. **Inference time analysis.** The inference time is evaluated by one A100 GPU. We utilize the LLaVA-OneVision-7B with an input of 16 frames. CLIP-L/14 is used to extract visual features.

Step	Time	Increase
VLM inference	1.32 s	
CLIP visual feature	1.21 s	
Inverse Transform Sampling	0.003 s	
Total	2.53 s	+90.9%

Additionally, our inference pipeline relies solely on basic CLIP for visual-text matching. While incorporating auxiliary alignment models or external tools, such as object detectors or OCR models, could further improve VQA performance, it would inevitably increase computational overhead. Our frame selection method is orthogonal to such approaches, as it focuses on selecting query-relevant frames to improve the effectiveness of downstream VLM inference.

C. Additional Qualitative Results

We provide additional visualizations of inverse transform sampling in Figure 1. The blue curve represents the similarity scores across the entire video sequence, while the red lines indicate the selected frames.

As shown in the figure, the proposed method effectively selects frames with high visual-query similarity. In addition, it preserves certain background elements, helping to maintain important contextual information required for accurate video understanding.

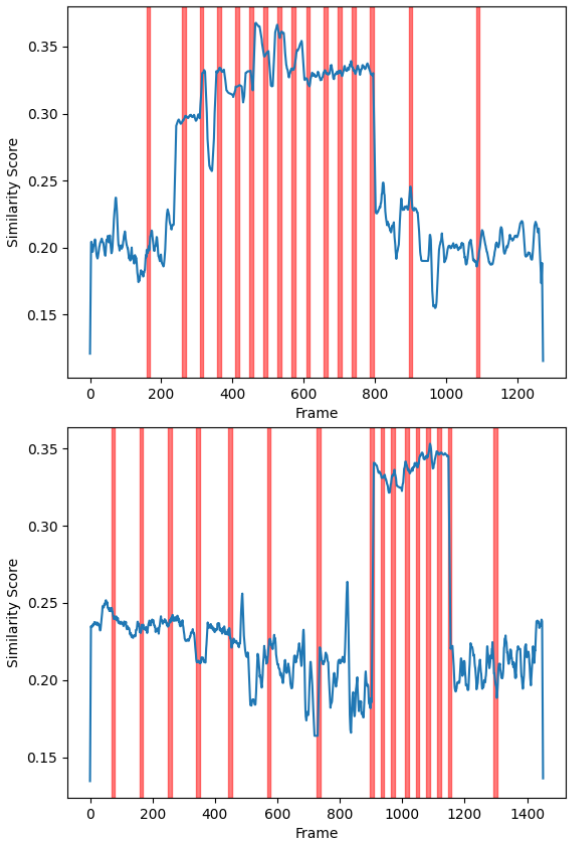

Figure 1. **Additional visualization results.**

In Figure 2, we present an example where the visual-query similarity scores remain relatively similar throughout the video. In such cases, the cumulative distribution function becomes approximately linear, causing the final selected frames to approximate uniform sampling. This indicates that most clips in the video may contribute equally to answering the question.

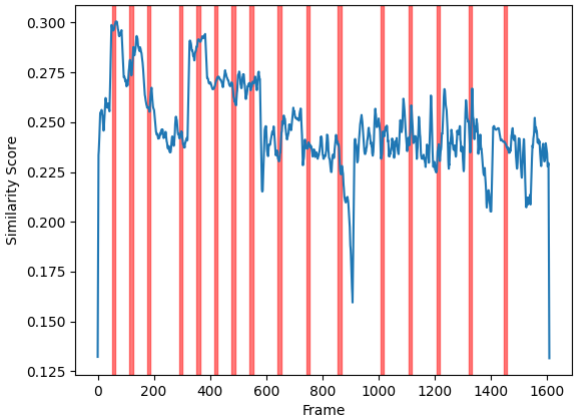

Figure 2. **When visual-query similarity scores remain similar across the video, inverse transform sampling simplifies to uniform sampling.**