Supplementary Material for "Why Is Video Still So Expensive? A Survey of Inference-Efficiency Mechanisms in Video and Audiovisual LLMs"

Killian Steunou, Yannis Tevissen, Mounîm El Yacoubi

S1. Full Inventory of VideoLLM Evaluation Benchmarks

Table S1 extends the compact benchmark table of the main paper to the full inventory of VideoLLM evaluation benchmarks we encountered while screening the surveyed methods, covering their modalities, evaluation format, duration regime and scale. Table S2 complements it with the standard task-level metrics used by the traditional datasets discussed in the main paper. Dense video captioning (DVC) is separated from clip-level captioning because it evaluates both temporal event localization and the language attached to each event: the protocol reports proposal precision/recall and METEOR [1] for captions matched at temporal IoU thresholds, while SODA [2] adds temporally ordered matching and a METEOR-based F-measure that penalizes redundant event captions.

References

[1] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 65–72.

[2] S. Fujita, T. Hirao, H. Kamigaito, M. Okumura, and M. Nagata, "SODA: Story Oriented Dense Video Captioning Evaluation Framework," in *Proc. ECCV*. Springer International Publishing, 2020, pp. 517–531.

[3] M. Ning, B. Zhu, Y. Xie, B. Lin, J. Cui, L. Yuan, D. Chen, and L. Yuan, "Video-Bench: A Comprehensive Benchmark and Toolkit for Evaluating Video-based Large Language Models," 2023.

[4] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo, L. Wang, and Y. Qiao, "MVBench: A Comprehensive Multi-modal Video Understanding Benchmark," in *Proc. CVPR*, 2024.

[5] Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, and L. Hou, "TempCompass: Do Video LLMs Really Understand Videos?" 2024.

[6] Y. Li, X. Chen, B. Hu, L. Wang, H. Shi, and M. Zhang, "VideoVista: A Versatile Benchmark for Video Understanding and Reasoning," 2024.

[7] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, P. Chen, Y. Li, S. Lin, S. Zhao, K. Li, T. Xu, X. Zheng, E. Chen, C. Shan, R. He, and X. Sun, "Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis," in *Proc. CVPR*, 2025.

[8] X. Chen, Y. Lin, Y. Zhang, and W. Huang, "AutoEval-Video: An Automatic Benchmark for Assessing Large Vision Language Models in Open-Ended Video Question Answering," 2024.

[9] Y. Hu, Z. Song, N. Feng, Y. Luo, J. Yu, Y.-P. P. Chen, and W. Yang, "SF2T: Self-supervised Fragment Finetuning of Video-LLMs for Fine-Grained Understanding," 2025.

[10] X. Fang, K. Mao, H. Duan, X. Zhao, Y. Li, D. Lin, and K. Chen, "MMBench-Video: A Long-Form Multi-Shot Benchmark for Holistic Video Understanding," 2024.

[11] H. Wu, D. Li, B. Chen, and J. Li, "LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding," in *Proc. NeurIPS*, 2024.

[12] W. Wang, Z. He, W. Hong, Y. Cheng, X. Zhang, J. Qi, X. Gu, S. Huang, B. Xu, Y. Dong, M. Ding, and J. Tang, "LVBench: An Extreme Long Video Understanding Benchmark," 2025.

[13] J. Zhou, Y. Shu, B. Zhao, B. Wu, Z. Liang, S. Xiao, M. Qin, X. Yang, Y. Xiong, B. Zhang, T. Huang, and Z. Liu, "MLVU: Benchmarking Multi-task Long Video Understanding," in *Proc. CVPR*, 2025.

[14] X. Tan, Y. Luo, Y. Ye, F. Liu, and Z. Cai, "ALLVB: All-in-One Long Video Understanding Benchmark," 2025.

[15] K. Chandrasegaran, A. Gupta, L. M. Hadzic, T. Kota, J. He, C. Eyzaguirre, Z. Durante, M. Li, J. Wu, and L. Fei-Fei, "HourVideo: 1-Hour Video-Language Understanding," 2024.

[16] H. Zou, T. Luo, G. Xie, V. X. J. Zhang, F. Lv, G. Wang, J. Chen, Z. Wang, H. Zhang, and H. Zhang, "HLV-1K: A Large-scale Hour-Long Video Benchmark for Time-Specific Long Video Understanding," 2025.

[17] K. Ataallah, E. Abdelrahman, M. Ahmed, C. Gou, K. Pahwa, J. Ding, and M. Elhoseiny, "InfiniBench: A Benchmark for Large Multi-Modal Models in Long-Form Movies and TV Shows," 2025.

[18] R. Rawal, K. Saifullah, M. Farré, R. Basri, D. Jacobs, G. Somepalli, and T. Goldstein, "CinePile: A Long Video Question Answering Dataset and Benchmark," 2024.

[19] K. Mangalam, R. Akshulakov, and J. Malik, "EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding," in *Proc. NeurIPS*, 2023.

[20] G. Chen, Y. Liu, Y. Huang, Y. He, B. Pei, J. Xu, Y. Wang, T. Lu, and L. Wang, "CG-Bench: Clue-grounded Question Answering Benchmark for Long Video Understanding," 2024.

[21] Y. Li, J. Niu, Z. Miao, C. Ge, Y. Zhou, Q. He, X. Dong, H. Duan, S. Ding, R. Qian, P. Zhang, Y. Zang, Y. Cao, C. He, and J. Wang, "OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding?" 2025.

[22] D. Cores, M. Dorkenwald, M. Mucientes, C. G. M. Snoek, and Y. M. Asano, "Lost in Time: A New Temporal Benchmark for VideoLLMs," 2025.

[23] M. Cai, R. Tan, J. Zhang, B. Zou, K. Zhang, F. Yao, F. Zhu, J. Gu, Y. Zhong, Y. Shang, Y. Dou, J. Park, J. Gao, Y. J. Lee, and J. Yang, "TemporalBench: Benchmarking Fine-grained Temporal Understanding for Multimodal Video Models," 2024.

[24] W. Hong, Y. Cheng, Z. Yang, W. Wang, L. Wang, X. Gu, S. Huang, Y. Dong, and J. Tang, "MotionBench: Benchmarking and Improving Fine-grained Video Motion Understanding for Vision Language Models," 2025.

[25] Y. Yuan, H. Zhang, W. Li, Z. Cheng, B. Zhang, L. Li, X. Li, D. Zhao, W. Zhang, Y. Zhuang, J. Zhu, and L. Bing, "VideoRefer Suite: Advancing Spatial-Temporal Object Understanding with Video LLM," 2025.

[26] H. Wang, Q. Chen, C. Yan, J. Cai, X. Jiang, Y. Hu, W. Xie, and S. Gavves, "Object-centric Video Question Answering with Visual Grounding and Referring," 2025.

[27] Z. Cheng, J. Hu, Z. Liu, C. Si, W. Li, and S. Gong, "V-STaR: Benchmarking Video-LLMs on Video Spatio-Temporal Reasoning," 2025.

[28] K. Hu, P. Wu, F. Pu, W. Xiao, Y. Zhang, X. Yue, B. Li, and Z. Liu, "Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos," 2025.

[29] Y. Zhao, L. Xie, H. Zhang, G. Gan, Y. Long, Z. Hu, T. Hu, W. Chen, C. Li, J. Song, Z. Xu, C. Wang, W. Pan, Z. Shangguan, X. Tang, Z. Liang, Y. Liu, C. Zhao, and A. Cohan, "MMVU: Measuring Expert-Level Multi-Discipline Video Understanding," 2025.

[30] E. Song, W. Chai, W. Xu, J. Xie, Y. Liu, and G. Wang, "Video-MMLU: A Massive Multi-Discipline Lecture Understanding Benchmark," 2025.

[31] Z. Zhou, R. Wang, and Z. Wu, "Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities," 2025.

TABLE S1

SUMMARY OF VIDEOLLM BENCHMARKS. MOD. DENOTES THE MODALITIES AVAILABLE IN THE BENCHMARK DATA BEYOND THE QUESTION TEXT: V=VIDEO (FRAMES), A=AUDIO TRACK, T=TRANSCRIPT/SUBTITLES (TEXT), I=IMAGE. (AUDIO/TEXT ARE MARKED ONLY WHEN THEY ARE EXPLICITLY PROVIDED AS PART OF THE BENCHMARK RELEASE.) S/M/L INDICATE WHETHER THE BENCHMARK INCLUDES SHORT (<1 MIN), MEDIUM (1–10 MIN), OR LONG (>10 MIN) VIDEOS.

Benchmark	Mod.	Type	Fmt.	S	M	L	#V	#Q	Notes
Video-Bench [3]	V + A	General eval	Mix	✓			5,917	15,033	Benchmark+toolkit for video-based LLM evaluation.
MVBench [4]	V	Multi-task temporal	MCQ	✓			3,641	4,000	20 tasks spanning perception→cognition.
TempCompass [5]	V	Temporal understanding	Mix	✓			410	7,540	Many instruction variants over a compact video set.
VideoVista [6]	V + T + I	Broad understanding+reasoning	OE/Mix	✓	✓	✓	3,402	24,906	Diverse categories, many task types, varied durations.
Video-MME [7]	V+A+T	Comprehensive eval	MCQ	✓	✓	✓	900	2,700	Multi-scale (short/med/long) + modality settings.
AutoEval-Video [8]	V	Auto-eval diagnostic	Mix	✓			327	327	Small set proposed for automatic evaluation/diagnosis.
FineVidBench [9]	V + I	Fine-grained video QA	Mix	✓	✓		910	22,718	Emphasizes fine-grained temporal diversity.
MMBench-Video [10]	V + T	Multi-shot understanding	Mix	✓	✓		600	2,000	Long-form multi-shot video understanding benchmark.
LongVideoBench [11]	V+T	Long-context QA	MCQ		✓	✓	3,763	6,678	Video–subtitle interleaved inputs up to 1 hour.
LVBench [12]	V	Extreme long video	MCQ/Mix			✓	103	1,549	Long-video comprehension and information extraction.
MLVU [13]	V	Multi-task long video	MCQ		✓	✓	1,730	3,102	Long videos (minutes→hours) with multiple tasks.
ALLVB [14]	V + T	Long video suite	MCQ/Mix			✓	1,376	252k	Very long videos on average; very large QA volume.
HourVideo [15]	V	Hour-long QA	MCQ/Mix			✓	500	12,976	Hour-level long-context video understanding.
HLV-1K [16]	V	Hour-long benchmark	MCQ/Mix			✓	1,009	14,857	Hour-long videos ("1K" scale) to stress long context.
InfiniBench [17]	V + T	Very-long benchmark	MCQ/Mix			✓	1,217	87.7k	Stress-test very long-range video understanding.
CinePile [18]	V + T	Long narrative QA	Mix			✓	9,396	305k	Movie-scale long video QA with large QA volume.
EgoSchema [19]	V	Long-form diagnostic	MCQ		✓		5,063	5,063	Long-form egocentric diagnostic benchmark.
CG-Bench [20]	V+A+T	Clue-grounded long QA	MCQ			✓	1,219	12,129	Requires retrieving temporally localized clues.
OVO-Bench [21]	V	Online/streaming	Mix		✓	✓	644	2,814	Timestamped queries under multiple online scenarios.
TVBench [22]	V	Temporal MCQ	MCQ	✓			2,525	2,654	Designed to reduce bias, isolate temporal understanding.
TemporalBench [23]	V	Fine-grained temporal	Mix	✓			2,179	9,867	Focuses on subtle temporal state transitions.
MotionBench [24]	V	Motion understanding	MCQ/Mix	✓			5,385	8,052	Motion-oriented categories for fine-grained motion.
VideoRefer-Bench [25]	V	Referring/object reasoning	MCQ	✓			198	1,000	Spatial-temporal object understanding via curated MCQs.
VideoInfer [26]	V	Object-centric inference QA	Mix	✓	✓		1,620	28,811	Emphasizes inference beyond direct observation.
V-STaR [27]	V	Spatiotemporal reasoning	Mix	✓	✓		743	6,282	Diagnostic spatiotemporal reasoning benchmark.
Video-MMMU [28]	V + A	Knowledge acquisition	MCQ		✓		300	900	Perception/comprehension/adaptation-aligned questions.
MMVU [29]	V	Expert multi-discipline	MCQ		✓		1,529	3,000	Expert-authored QAs + rationales and domain knowledge.
Video-MMLU [30]	V	Lecture understanding	OE/Mix		✓		1,065	15,746	Notes (captioning) + quiz (reasoning) evaluation.
Daily-Omni [31]	V+A	Audio-visual reasoning	MCQ	✓	✓		684	1,197	Explicit temporal alignment across audio and vision.
OmnixR [32]	V + A + T + I	Cross-modal reasoning	MCQ	✓	✓	✓	1,500	1,500	Synthetic + real-world cross-modal scenarios.
LongVALE [33]	V + A + T	Audiovisual long video	OE	✓	✓		8,411	105,730	Fine-grained events and audiovisual correlation captions.
MT-Video-Bench [34]	V	Multi-turn dialogue	OE/Mix	✓	✓	✓	135	5,887	Multi-turn video dialogue with open-ended interaction.
MVU-Eval [35]	V	Multi-video reasoning	Mix	✓	✓		4,959	1,824	Reasoning over sets of multiple videos.
SCVBench [36]	V	Story-centric reasoning	Mix		✓	✓	925	7,280	Story-centric evaluation (event ordering/decomposition).
VRBench [37]	V	Multi-step long narrative	OE/Mix			✓	960	8,243	Long videos (avg. ∼1.6 h) with multi-step reasoning.
ExAct [38]	V	Expert action analysis	MCQ	✓			3,521	3,521	Expert skill understanding across multiple domains.
X-LeBench [39]	V	Ultra-long egocentric	Mix			✓	432	11,313	Life-log durations from minutes to many hours.
MMR-V [40]	V	Deep video reasoning	OE/Mix	✓	✓	✓	317	1,257	Long-range multimodal deep reasoning benchmark.

[32] L. Chen, H. Hu, M. Zhang, Y. Chen, Z. Wang, Y. Li, P. Shyam, T. Zhou, H. Huang, M.-H. Yang, and B. Gong, "OmnixR: Evaluating Omni-modality Language Models on Reasoning across Modalities," 2024.

[33] T. Geng, J. Zhang, Q. Wang, T. Wang, J. Duan, and F. Zheng, "LongVALE: Vision-Audio-Language-Event Benchmark Towards Time-Aware Omni-Modal Perception of Long Videos," 2024.

[34] Y. Pan, Q. Xie, G. Zhang, Z. Wang, Y. Wen, Y. Zhang, H. Hu, Z. Pan, Y. Huang, Z. Gan, Y. Lin, A. Ping, S. Li, Y. Wang, T. Peng, and J. Liu, "MT-Video-Bench: A Holistic Video Understanding Benchmark for Evaluating Multimodal LLMs in Multi-Turn Dialogues," 2026.

[35] T. Peng, H. Wang, Y. Zhang, Z. Wang, Z. Wang, G. Chang, J. Yang, S. Li, Y. Wang, X. Wang, H. Li, W. Ji, P. Wan, S. Huang, Z. Zhang, and J. Liu, "MVU-Eval: Towards Multi-Video Understanding Evaluation for Multimodal LLMs," 2025.

[36] S. You, B. Yuan, and B.-K. Bao, "SCVBench: A Benchmark with Multi-turn Dialogues for Story-Centric Video Understanding," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 2287–2295.

[37] J. Yu, Y. Wu, M. Chu, Z. Ren, Z. Huang, P. Chu, R. Zhang, Y. He, Q. Li, S. Li, Z. Li, Z. Tu, C. He, Y. Qiao, Y. Wang, Y. Wang, and L. Wang, "VRBench: A Benchmark for Multi-Step Reasoning in Long Narrative Videos," 2025.

[38] H. Yi, Y. Pan, F. He, X. Liu, B. Zhang, O. Oguntola, and G. Bertasius, "ExAct: A Video-Language Benchmark for Expert Action Analysis," 2025.

[39] W. Zhou, K. Cao, H. Zheng, Y. Liu, X. Zheng, M. Liu, P. O. Kristensson, W. Mayol-Cuevas, F. Zhang, W. Lin, and J. Shen, "X-LeBench: A Benchmark for Extremely Long Egocentric Video Understanding," 2025.

[40] K. Zhu, Z. Jin, H. Yuan, J. Li, S. Tu, P. Cao, Y. Chen, K. Liu, and J. Zhao, "MMR-V: What's Left Unsaid? A Benchmark for Multimodal Deep Reasoning in Videos," 2025.

[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[42] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[43] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," 2014.

[44] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," 2016.

TABLE S2
COMMON EVALUATION METRICS BY TASK IN VIDEO UNDERSTANDING. ↑: HIGHER IS BETTER; ↓: LOWER IS BETTER.

Task	**Metrics**
Action classification	Top-1 / Top-5 accuracy ↑
Action detection / localization	mAP at temporal IoU thresholds ↑
Clip-level video captioning	BLEU; METEOR; ROUGE-L; CIDEr; SPICE ↑ [1], [41]–[44]
Dense video captioning	Proposal precision / recall at temporal IoU thresholds; METEOR; SODA ↑ [2]
Video–text retrieval	R@1/5/10 ↑; median rank ↓
Temporal grounding	R@K at temporal IoU thresholds; mean IoU ↑
Summarization / chaptering	F1 overlap; mean IoU; caption metrics ↑
Video QA / dialogue	Accuracy; exact match; soft matching ↑