VideoAds for Fast-Paced Video Understanding

Supplementary Material

I. Video filtering criteria

The filtering criteria for our VideoAds during the video selection included: i) Non-advertisement videos (e.g., user-generated content, review videos, or behind-the-scenes footage). ii) Short videos (less than 30 seconds), which typically lack the necessary narrative progression for meaningful temporal reasoning. iii) Low-quality videos (e.g., advertisements consisting solely of several static images, lacking dynamic scene transitions). In particular, steps i and iii filtering are performed by human experts, while step ii filtering is conducted automatically.

II. Further discussion on the video complexity

Here we further provide the influence of time duration d on the V_{cpx}, and we can observe that increasing time duration V_{cpx} tends to increase due to more variance in the long time unless it already reaches the maximum the video durations. And regardless of the time duration, the VideoAds can significantly outperforms other datasets in terms of video complexity.

	5	10	15	20	25	30
TGIF	20.79	21.10	21.11	21.11	21.11	21.11
ActivityNet	37.91	40.82	41.74	42.19	42.44	42.60
TVQA	41.85	45.06	46.11	46.64	46.94	47.12
MSVD	25.36	26.55	26.68	26.70	26.70	26.70
MSRVTT	33.25	35.62	36.00	36.06	36.08	36.09
EgoSchema	25.88	31.04	33.61	35.28	36.51	37.47
AutoEval-Video	26.43	30.46	31.54	31.95	32.14	32.26
TempCompass	16.50	20.21	21.23	21.55	21.63	21.64
NExTVideo	19.54	24.01	26.09	27.26	27.97	28.43
Video-Bench	23.10	26.88	28.31	29.08	29.59	29.97
MVBench	17.39	20.43	21.71	22.36	22.67	22.83
VideoMME-S	52.06	56.10	57.44	58.09	58.47	58.72
VideoMME-M	32.98	42.38	47.08	50.05	52.14	53.71
VideoAds	58.71	67.40	70.65	72.32	73.31	73.89

Table 1. The influence of time duration in calculating the video complexity, with the increasing time duration V_{cpx}, tends to increase due to more variance in the long term.

It is also worth noticing that the video complexity score only measures the complexity of the video, and can not fully represent the complexity in terms of VQA benchmarks. For example, TempCompass dataset [10] show the challenging of time order in the video understanding for MLLMs. While the video itself can be easy, there exists the possibility that the VQA question can be still challenging.

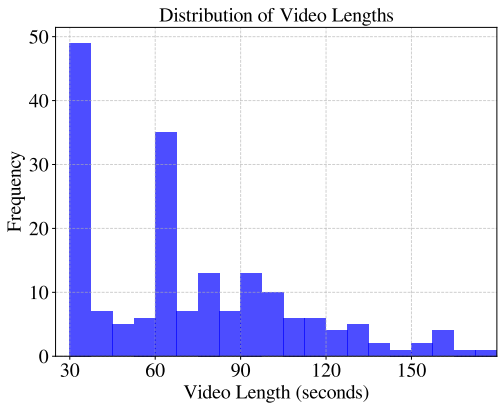

Figure 1. The video duration distribution of proposed VideoAds. We can find most advertisement videos are particularly focusing on 30 seconds, 60 seconds, and 90 seconds.

III. Duration Distribution of VideoAds

Here we provide the distribution of video durations in Figure 1, Interestingly, we can find that most advertisement videos are particularly focusing on 30 seconds, 60 seconds, and 90 seconds. This aligns with the real-world advertisement video distribution given the property of advertisement markers.

IV. Failure case study

Here we provide several failure cases by common MLLMs study in VideoAds benchmark as shown in Figure 2,3,4,5,8. Interestingly, for some reasoning questions when it is against common sense like Figure 2, MLLMs tend to generate answers based on common sens rather than visual information.

V. Why not CLIP?

The reason for choosing DINO rather than CLIP [12] is that DINO is more focused on the changes of the low-level visual components, while CLIP is noisier due to semantic encoding [13]. Here we provide several examples in Figure 9 where the CLIP provides one high video complexity but the video actually contains low complexity.

VI. More discussion in related work

Traditional video evaluation benchmarks focus on one specific task by collecting data from the corresponding domain. For example, ActivityNet-QA [15] focuses on human activity recognition, and MSVD-QA [14] are generated by video description datasets and focus on action and object

Question: Why does the dog give the tennis ball to the man in the video?

Type: Reasoning *Complexity: 77.03* *Answer: A*

A. The dog gives the tennis ball to the man because he wants to attract his attention from other dogs towards the man.

B. The dog gives the tennis ball to the man because it is a form of bonding and exercise for both the dog and the man.

C. The dog gives the tennis ball to the man because he wants to play with him.

D. The dog gives the tennis ball to the man because it is trained to do so and enjoys playing fetch.

Gemini Answer: C GPT-4o Answer: D Qwen2.5-VL Answer: C LLava-Video Answer: D

Figure 2. Selected case visualization with MLLMs prediction

Question: Which film shares the same storyline as this video?

Type: Reasoning *Complexity: 70.69* *Answer: A*

A. Roman Holiday

B. The City of Love

C. The Devil Wears Prada

D. Before Sunrise

Gemini Answer: A GPT-4o Answer: A Qwen2.5-VL Answer: C LLava-Video Answer: C

Figure 3. Selected case visualization with MLLMs prediction

recognition. Along with the rapid progress of the current LLMs [2, 6], the recent VLM enables the process of complex

Question: What emotion is the woman expressing at the end of the video?

Type: Reasoning *Complexity: 69.63* *Answer: D*

A. Joyful excitement

B. Tearful gratitude

C. Contentment

D. Touched

Gemini Answer: B GPT-4o Answer: D Qwen2.5-VL Answer: B LLava-Video Answer: B

Figure 4. Selected case visualization with MLLMs prediction

Question: What occurs upon the women's return to land?

Type: Summary *Complexity: 74.76* *Answer: A*

A. The man left by car at the same time.

B. The women are greeted by a group of people on the beach.

C. They are seen walking along the shoreline, carrying their surfboards.

D. The women are captured in a moment of joy and celebration, hugging each other.

Gemini Answer: D GPT-4o Answer: C Qwen2.5-VL Answer: C LLava-Video Answer: C

Figure 5. Selected case visualization with MLLMs prediction

video in more than one single motion or short video clip [5]. Correspondingly, the recent video evaluation benchmarks

Question: In which city do the women go to a restaurant at the initial part of the video?

Type: Finding *Complexity: 67.45* *Answer: B*

A. The women go to a restaurant in Paris.

B. The women go to a restaurant in New York.

C. The women go to a restaurant in London.

D. The women go to a restaurant in Tokyo.

Gemini Answer: D GPT-4o Answer: D Qwen2.5-VL Answer: B LLava-Video Answer: D

Figure 6. Selected case visualization with MLLMs prediction

Question: How many people are in the video?

Type: Summary *Complexity: 28.86* *Answer: A*

A. Two people are in the video.

B. Only one person is visible in the video.

C. There are three people in the video.

D. There are four people in the video.

Gemini Answer: A GPT-4o Answer: A Qwen2.5-VL Answer: A LLava-Video Answer: A

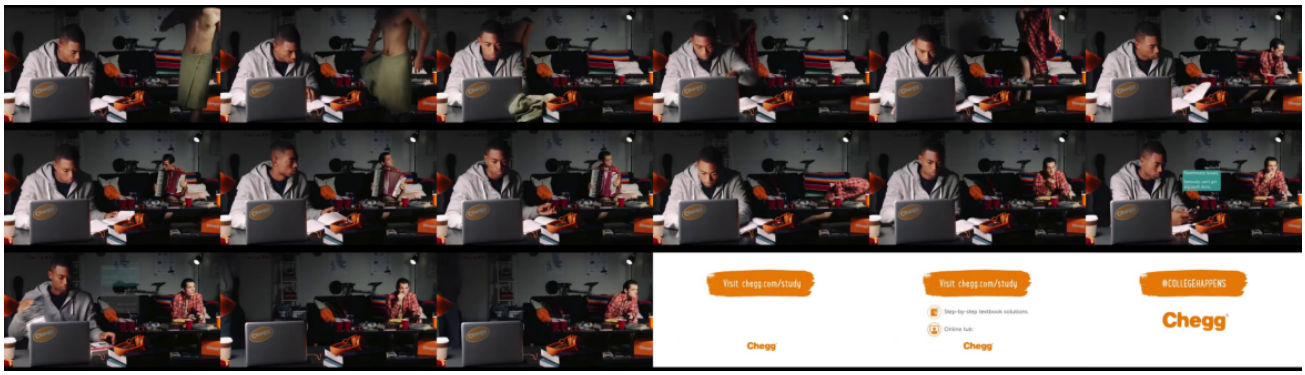

Figure 7. Selected visualization for case with low complexity with MLLMs prediction

are more focused on more challenging and comprehensive video understanding. Video-MME [4] is one widely used

Question: What vegetables are included in the McRib?

Type: Finding | *Complexity: 3.78* | *Answer: A*

A. The McRib includes onions and pickles as vegetables.

B. The McRib is topped with onions and a slice of tomato.

C. The McRib is garnished with onions and a pickle spear.

D. The McRib is served with onions and a side of coleslaw.

Gemini Answer: A | GPT-4o Answer: A | Qwen2.5-VL Answer: A | LLava-Video Answer: A

Figure 8. Selected visualization for case with low complexity with MLLMs prediction

CLIP Complexity Score: 55.59 vs DINO Complexity Score: 30.83

CLIP Complexity Score: 65.90 vs DINO Complexity Score: 33.68

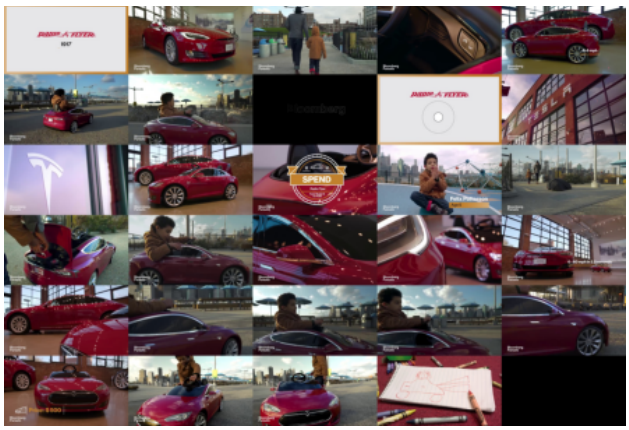

Figure 9. Some examples that CLIPs generate over-estimation for the video complexity.

comprehensive video evaluation benchmark containing different domains and various video lengths (from seconds to hours). MVBench [8] covers 20 video tasks for spatial understanding and temporal understanding. This benchmark also builds a semiautomatic reannotation pipeline using ChatGPT for existing video datasets with original annotations. MMBench-Video [9] focuses on free-form questions from lengthy YouTube videos and introduces GPT-4 [1] for automated assessment. Video-Bench [11] includes three types of questions, including video-exclusive Understanding, prior Knowledge-based question-answering, and comprehension and decision-making based on the real-world video. TempCompass [10] focuses on on the temporal perception ability of MLLM by collecting videos that share the same static content but differ in a specific temporal aspect. AutoEval-Video [3] develop a novel adversarial annotation mechanism and constructs open-ended video-questions across 9 skill dimensions. Another interesting recent work named Video-MMMU [7] focuses on long professional videos that systematically evaluate knowledge acquisition capabilities.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and etc. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[3] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pages 179–195. Springer, 2024. 5

[4] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 4

[5] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024. 3

[6] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *ArXiv preprint*, 2023. 2

[7] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. 5

[8] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *ArXiv preprint*, 2023. 5

[9] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *ArXiv preprint*, 2023. 5

[10] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *ArXiv preprint*, 2024. 1, 5

[11] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *ArXiv preprint*, 2023. 5

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[13] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 1

[14] D. Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 1

[15] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019. 1