VideoAds for Fast-Paced Video Understanding

Zheyuan Zhang[1*] Wanying Dou[1*] Linkai Peng[1*] Hongyi Pan[1] Ulas Bagci[1] Boqing Gong[2]
[1]Northwestern University [2]Boston University
{zheyuan.zhang, monica.dou, linkai.peng, hongyi.pan, ulas.bagci}@northwestern.edu, bgong@bu.edu

Abstract

*Advertisement videos serve as a rich and valuable source of purpose-driven information, encompassing high-quality visual, textual, and contextual cues designed to engage viewers. They are often more complex than general videos of similar duration due to their structured narratives and rapid scene transitions, posing significant challenges to multi-modal large language models (MLLMs). In this work, we introduce VideoAds, the first dataset tailored for benchmarking the performance of MLLMs on advertisement videos. VideoAds comprises well-curated advertisement videos with complex temporal structures, accompanied by **manually** annotated diverse questions across three core tasks: visual finding, video summary, and visual reasoning. We propose a quantitative measure to compare VideoAds against existing benchmarks in terms of video complexity. Through extensive experiments, we find that Qwen2.5-VL-72B, an open-source MLLM, achieves 73.35% accuracy on VideoAds, outperforms GPT-4o (66.82%) and Gemini-1.5 Pro (69.66%); the two proprietary models especially fall behind the open-source model in video summarization and reasoning, but perform the best in visual finding. Gemini-2.5 Pro leads with an accuracy of 80.04%. Notably, human experts easily achieve a remarkable accuracy of 94.27%. These results underscore the necessity of advancing MLLMs' temporal modeling capabilities and highlight VideoAds as a potentially pivotal benchmark for future research in understanding video that requires high FPS sampling. The dataset and evaluation code will be publicly available at* https://videoadsbenchmark.netlify.app.

1. Introduction

Advertisement videos represent a unique and high-value segment of visual media, carefully crafted to capture the audience's attention through rich multimodal content in a short time [42]. Unlike general videos, advertisements integrate carefully designed storytelling, persuasive visual elements, and concise yet information-dense narratives to convey targeted messages efficiently. These videos often include high-quality cinematography, strategic scene transitions, and contextual cues that enhance engagement and retention, making them invaluable in marketing, entertainment, and e-commerce.

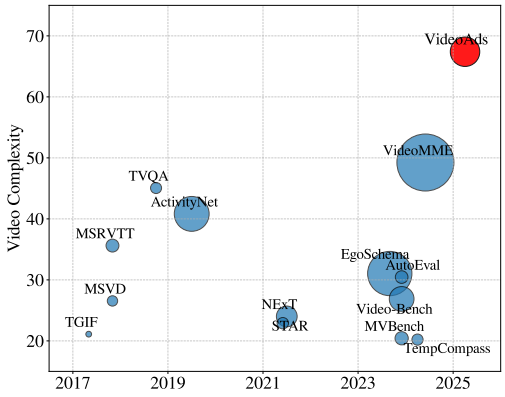

Figure 1. Recent years have witnessed a significant increase in the complexity of video benchmarks, paralleling the rapid progress in the capabilities of multi-modal large language models (MLLMs). In this work, we introduce VideoAds, a complex dataset based on advertisement videos, specifically designed to benchmark the performance of MLLMs on challenging visual comprehension and complex temporal reasoning. The size of each scatter point represents the average duration of videos within each dataset.

The rapid advancements in Large Language Models (LLMs) [2, 27] have revolutionized multi-modal video understanding, enabling models to process and reason over video content with unprecedented efficacy [11, 30, 45]. Generally, these models take videos as input and project them into the embedding or token space of LLMs [22, 24, 27, 48] and leverage LLMs to generate high-quality outputs. These multi-modality LLMs (MLLMs) exhibit remarkable open-ended generation capabilities, including but not limited to temporal, spatial, and causal inference, making them promising for real-world video applications [7, 16].

However, we contend that existing video-language benchmarks do not thoroughly assess MLLMs' ability to comprehend complex temporal dependencies. Advertisement videos, a valuable but underexplored video type in MLLM evaluation, while typically short, exhibit intricate multi-stage

Figure 2. VideoAds comprises three challenging tasks: Visual Finding, Visual Summary, and Visual Reasoning, specifically designed to evaluate MLLMs' temporal reasoning capabilities on videos with complex temporal structures that have never been investigated before. Unlike many previous datasets that focus on recognizing isolated actions or events, VideoAds demands that models derive the correct answers only through multistep reasoning over multi-modal visual clues.

narratives, rapid scene transitions, and implicit causal relationships, making them particularly difficult for MLLMs to comprehend. In contrast, existing benchmarks for video understanding primarily focus on generic action recognition, instructional videos, movie comprehension, and general videos collected from YouTube [7, 16, 18, 20]. These videos are rich and diverse in actions and scenes, but they are not comparable to advertisement videos in terms of how to purposely use complex visual cues, background context, and temporally disjointed sequences to convey persuasive messaging. As a result, advertisement videos are a more challenging testbed than generic videos for assessing MLLMs' inference ability across time and space, multi-step reasoning about narratives, and long-range summarization, among

other essential abilities needed for video understanding.

To this end, we introduce VideoAds, the first dedicated dataset for evaluating MLLMs in the context of advertisement video understanding. It is important to position VideoAds in the landscape of existing benchmarks. VideoAds is not intended to replace larger, more comprehensive benchmarks like Video-MME [7], but rather to complement them. VideoAds presents unique challenges for video understanding due to its complex temporal structures and high-density semantic content, making it a testbed with high potential for identifying limitations of current MLLMs in temporal reasoning and, accordingly, driving the development of next-generation AI systems with advanced temporal modeling, cross-modal interactions, and multi-stage reasoning capabilities. Beyond academic impact, VideoAds addresses real-world challenges in automated content analysis, brand sentiment prediction, and strategic ad placement, facilitating AI-empowered applications in the advertisement sector.

We summarize our main contributions as follows:

- **First Advertisement Video Benchmark for MLLMs**: We present VideoAds, a carefully curated benchmark dataset consisting of high-quality and high-complexity advertisement videos, featuring **manually** annotated questions covering three groups of tasks: visual finding, video summarization, and visual reasoning.
- **Quantitative Video Complexity Measure**: We introduce a novel quantitative measure for video complexity, enabling a holistic analysis of how scene transitions, narrative coherence, and temporal event dependencies influence MLLM performance. This measure also accommodates a comparison between VideoAds and existing video understanding datasets in terms of complexity.
- **Comprehensive Benchmarking of State-of-the-Art Models**: We evaluate leading MLLMs, including GPT-4o [12], Gemini 1.5 Pro [31], and open source models like LLaVa-Video [48], Qwen2.5-VL [33] over multiple-choice VQA tasks, revealing significant gap from human performance when the models attempt to reason across complex temporal structures. Moreover, MLLMs perform well in finding visual patterns from videos but struggle with video summarization and reasoning.
- **Insights about Audio and Chain of Thought in MLLMs**: We conduct further analyses about the influence of speech audio transcripts and Chain of Thoughts (CoT). We find that large MLLMs benefit from CoT more than small models, aligning with LMMs' properties, and the additional information from speech transcripts almost always improves MLLMs' performance. These results signify the need for future MLLMs to reinforce cross-modal interactions and reasoning.

2. What makes an advertisement video?

Advertisement videos are a unique form of visual storytelling, meticulously crafted to capture audience attention, convey persuasive messages, and influence consumer behavior within a short time [13, 29]. For example, the second video in Figure 2, "CHANEL N°5 the Film" is a 120-second short advertisement directed by Baz Luhrmann and with a budget of $33 million [37]. This advertisement tells a story: "A famous celebrity runs away in a pink dress in the middle of Times Square in New York City, only to get into a cab with a man who does not recognize her. After four days in his Lower East Side apartment, her secretary insists that she return as a celebrity." This advertisement video shares a plotline similar to Roman Holiday but condenses it into just 120 seconds and provides us with high-density information that has rarely been seen in other video types. The high commercial value translates into complex narrative structures and information-dense sequences, making advertisement videos more challenging to analyze for the current MLLMs which often rely on low FPS sampling.

3. VideoAds

To build the VideoAds benchmark, we take three steps: video collection followed by manual filtering; curation of VQA tasks; and quality review.

3.1. Video collection and filtering

We begin by collecting YouTube advertisement videos, using targeted keyword searches to encompass various commercial domains. Specifically, we retrieve videos using commercial search terms about Cars, Sports, Movie Trailers, Food, Tech, Health, Travel, and Financial Services. This broad list ensures our dataset covers multiple industries, allowing for diverse visual storytelling techniques, product promotions, and narrative-driven ad structures. After this step, we collected 2,700 videos from YouTube.

To refine the dataset, we implement a systematic filtering process to remove irrelevant, excessively short, low-quality, and static-content videos that do not align with our benchmarking objectives. The specific filtering criteria are attached in the Appendix. After applying these filtering steps, we keep 200 high-quality, visually complex advertisement videos from the original 2,700 candidates. The data size of 200 is a sweet trade-off between the benchmark's coverage (in terms of diversity and visual quality) and utility (in computational and financial costs). These videos have densely packed content, including but not limited to rich visual storytelling, diverse editing styles, and complex temporal structures, making them a versatile testbed for evaluating the temporal reasoning capabilities of MLLMs.

Figure 3. Visualization of Video Complexity Scores: Examples of videos with high, medium, and low complexity scores based on the proposed video complexity metrics. Videos with high complexity scores exhibit frequent scene transitions, dynamic interactions, and complex visual transformations, posing significant challenges for temporal reasoning. In contrast, videos with low complexity display static or minimally changing frames, resembling image slideshows with minimal narrative progression.

Table 1. The comparison of various minutes long video benchmarks: total number of videos (**#Videos**), number of clips (**#Clips**), average duration of the videos (**Len.**), number of QA pairs (**#QA Pairs**), method of annotation (**Anno.**, M/A means the manually/automatic manner), average number of QA pair tokens (**QA Tokens**), complexity duration(**Complexity Duration**), video complexity (**Video Complexity**)

Benchmarks	#Videos	#Clips	Len.(s)	#QA Pairs	Anno.	Complexity Duration (↑)	Video Complexity (↑)
TGIF-QA [14]	9,575	9,575	3.0	8,506	A&M	0.67	21.10
STAR [38]	914	7,098	11.9	7,098	A	6.25	22.90
NExT-QA [39]	1,000	1,000	39.5	8,564	A	9.23	24.01
MSVD-QA [40]	504	504	9.8	13,157	A	1.75	26.55
MSRVTT-QA [40]	2,990	2,990	15.2	72,821	A	3.68	35.62
ActivityNet-QA [44]	800	800	111.4	8,000	M	47.59	40.82
TempCompass [18]	410	500	11.4	7,540	A&M	1.62	20.21
MVBench [16]	3,641	3,641	16.0	4,000	A	3.34	20.43
Video-Bench [21]	5,917	5,917	56.0	17,036	A&M	16.58	26.88
EgoSchema [20]	5,063	5,063	180.0	5,063	A&M	54.30	31.04
AutoEval-Video [3]	327	327	14.6	327	M	3.46	30.46
Video-MME-S [7]	300	300	80.7	900	M	44.81	56.10
Video-MME-M [7]	300	300	515.9	900	M	220.88	42.38
VideoAds	200	200	79.6	1100	M	60.32	67.40

3.2. VQA annotations

We ask human experts to manually construct video question answering (VQA) tasks about the videos. The annotation process is designed to emphasize complex temporal reasoning, visual interpretation, and high-level narrative understanding, addressing key aspects of video understanding. We categorize the VQA tasks into three core types.

- **Visual finding**: This category assesses the model's ability to extract concrete visual patterns from video frames. It includes but is not limited to object recognition, attribute identification (*e.g.*, color, shape, size), detecting spatial relationships among objects, and scene recognition. Some exemplar questions are as follows. "What color is the second car in the advertisement?" "Where is the ball on the table?" "In which city does the first scene take place?"
- **Visual summary**: This category evaluates the model's ability to summarize key events and thematic elements in the video. Questions focus on event description, sequential understanding, changes across time, and topic discovery. For example, "How many types of insurance are mentioned in the ads?", "What is the main product being advertised?", "What happened before the targeted product is introduced?"
- **Visual reasoning**: This is the most challenging category, requiring the model to perform high-order inference based on the video's context. The questions involve cause-and-effect analyses, interpersonal dynamics, emotional interpretation, and logical problem-solving. For example, "Why does the man finally decide to purchase the prod-

uct?", "What emotion does the actor express after using the product?", "How does the advertisement convey the product's benefits using different scenes?"

While semi-automated VQA annotation methods using powerful MLLMs [19, 41] can easily generate a large number of question-answer pairs using human-provided captions, they are inherently limited by the granularity of the captions and MLLMs' potential bias. Hence, in this work, the VQA tasks are mainly generated by human annotations. Human annotators are required to create a question, a correct answer, and an incorrect (yet confusing) answer for each VQA task. For every video, an annotator is required to generate 10 VQA pairs, leading to 2,000 VQA tasks in total. We then employ the LLava-Next-72B model [48] to generate four candidate incorrect answers. Then, OpenAI GPT4o [11] is used to merge the question, correct answer, and human-provided and model-generated incorrect answers into standardized four-option multiple-choice questions. We show examples of all stages of this annotation process in the Appendix.

3.3. Quality review

To ensure the accuracy and depth of the generated VQA tasks, we apply a quality control step to preserve the most informative and challenging questions, ensuring a balanced distribution across the three primary VQA categories and removing trivial queries. Specifically, we remove VQA pairs that are trivial or overly simplistic, as well as those that contain ambiguities, multiple correct answers, or annotation errors. After expert filtering, we retain 1,100 high-quality VQA tasks from the original 2,000. The numbers of visual finding, visual summary, and visual reasoning tasks are 425, 312 and 363, respectively, and evenly distributed across all question types. Overall, this refinement significantly improved the VQA quality, making it a challenging benchmark for assessing MLLMs.

3.4. Dataset statistics

We analyze VideoAds and contrast it with existing ones and, through this comparison, we notice that VideoAds differs from others primarily by its high visual complexity across time. The following first defines video complexity, a quantitative measure, and then presents dataset statistics.

3.4.1. Definition of video complexity

Although many video datasets claim to collect complex videos from diverse application scenarios, a quantitative definition of video complexity remains elusive in video analysis. Inspired by [18, 32] and moving beyond qualitative observations, we propose a quantitative measure of video complexity based on the variance of visual features over time, addressing the need for a structured approach to measuring complexity in MLLM-based video understanding.

Intuitively, a video that exhibits greater changes over time within a given duration is more complex and challenging to understand than a static or minimally changing video. In the context of MLLMs, this translates to greater difficulty in understanding temporal coherence/changes, causal reasoning, and narrative understanding. In this work, we define video complexity based on visual feature variance with DINO-v2 (referred to as f) as a feature extractor [23]. The reason for choosing DINO rather than CLIP [25] is that DINO preserves more visual details than CLIP [32].

Given a specific video V, we sample one frame I_i per second from it and denote by n the total number of video frames. Denote by $[-d, d]$ the neighbor of any frame. We define *complexity duration* D_{cpx} as:

$$D_{cpx} = n - \sum_{i=1}^{n} \frac{\sum_{j=\max(i-d,1)}^{\min(i+d,n)} e^{-\frac{|j-i|}{2d}} \cos(f(I_i), f(I_j))}{\sum_{j=\max(i-d,1)}^{\min(i+d,n)} e^{-\frac{|j-i|}{2d}}}, \tag{1}$$

where $j \neq i$, $\cos(.,.)$ indicates cosine similarity, and the exponential weighting $e^{-\frac{|j-i|}{2d}}$ ensures that nearby frames contribute more to the complexity calculation than distant frames. The complexity duration is named in analogy to time duration. Indeed, Egoschema [20] videos are long in duration (180 seconds per video, on average), but their average complexity duration is 54.3. We further define *video complexity* V_{cpx} (also called complexity density) as the average of complexity duration across time:

$$V_{cpx} = 100 D_{cpx}/n, \tag{2}$$

A high video complexity V_{cpx} value indicates great visual changes over unit time, necessitating high FPS sampling strategies for MLLMs to comprehend the video thoroughly. Figure 3 illustrates videos with high, medium, and low video complexities, validating our measure. We analyze the impact of neighborhood size d in the Appendix.

3.4.2. VideoAds analysis and comparison with others

VideoAds distinguishes itself from existing video benchmarks by presenting a unique combination of high video complexity and challenging question-answer (QA) pairs, despite containing relatively short videos: videos in [8] are up to hours long. Unlike conventional datasets, such as MSVD-QA [40], Youcook2 [50], and ActivityNet-QA [44], which primarily focus on action recognition or object identification with lower video complexity, VideoAds emphasizes complex temporal reasoning on challenging tasks. This is achieved by curating advertisement videos that compress rich visual narratives and multi-stage events into one to two minutes, demanding long-range contextual tracking and multi-hop reasoning. Compared to Video-MME [7], MMBench-Video [17], and recent Video-MMMU [10], which focus on longer video durations or multi-domain understanding, VideoAds presents a higher density of temporal changes.

	Finding	Summary	Reasoning	Total
#Question	9.58	8.85	9.32	9.29
#Options	33.64	40.38	50.09	40.98
Duration	83.28	78.45	78.01	79.60
QA Tokens	53.63	59.95	71.80	61.42
Number	425	312	363	1100

Table 2. Summary of key statistics across the three question types in VideoAds, including average word count for questions (#Question), answer options (#Options), as well as average video duration (Duration) and average QA tokens (QA Tokens).

Table 1 uses the video complexity measure to highlight that VideoAds videos contain significantly more temporal dynamics than other benchmarks. This compact complexity makes it particularly challenging for MLLMs to maintain temporal coherence and causal reasoning, even more so than processing long-form but slowly progressing videos. Furthermore, unlike datasets [8, 17] that use free-form questions, VideoAds employs four-option multiple-choice questions facilitating easy evaluations across different models. Finally, Table 2 presents some statistics for each type of question. In summary, VideoAds raises the bar in video understanding by introducing complex and diverse advertisement videos and VQA tasks.

It is crucial to understand why benchmarks like TempCompass and MVBench score lower on our complexity metric despite being used for fine-grained temporal understanding. Our metric, based on DINO-v2 feature variance, excels at capturing coarse visual shifts and significant content changes typical of advertisements with rapid cuts between distinct scenes. In contrast, videos in TempCompass or MVBench often feature more subtle temporal changes (e.g., an object moving slightly), which results in simpler overall video content and thus a lower complexity score. VideoAds therefore, serves as a complementary benchmark, focused on challenging models with fast-paced, semantics-rich visual transitions rather than fine-grained temporal shifts.

4. Experiments and Results

4.1. Experiment setting

MLLMs: We evaluate our collected dataset using open-source MLMMs, including InternVideo2 [34], LongVA [47], InternVL2 [5], LLaVA-Next-Video [48], LLaVA-Video [49], Qwen2.5-VL [1], which covers models from 7B to 72B parameters; and commercial models GPT-4o [12] , Gemini 1.5 Pro and 2.5 Pro [6, 28]. We also compare the influence of the number of sampled frames on various models for those models that accept different numbers of frames as input.

LLM performance: To ensure VideoAds serves as a reliable benchmark for video understanding, it is crucial to eliminate the possibility of MLLMs exploiting textual patterns rather than genuinely understanding visual content. Therefore, we conduct a baseline experiment using GPT-4o with text-only input, excluding all visual signals. Specifically, the model was prompted with instructions such as, 'Based only on the provided question, answer the following multiple-choice question,' without any indication that visual input was required. This process helps quantify the model's reliance on visual information versus textual bias.

Human performance: To assess the performance of humans in such a challenging task, we recruit two master's students as human experts and instruct them to complete the following tests: Given that most advertisement videos are designed to contain specific information for human, it is relatively easy to answer accurately by the evaluators. Each participant is instructed to watch each video once and then answer the corresponding VQA pairs without revisiting the video. This approach mimics the real-world constraints that MLLMs face, ensuring a fair comparison between human and model performance.

Evaluation: To ensure a standardized and fair comparison across different MLLMs, we organize the VideoAds dataset into a four-option multiple-choice format. Each question is accompanied by one correct answer and three carefully curated distractors, which are semantically plausible but incorrect. We merge our dataset into the LMM-Eval package [46], which has been widely used in MLLMs evaluation such that other researchers can use this dataset easily.

4.2. Benchmarking results

Table 3 presents a comprehensive comparison of model performance across three groups: LLM performance without visual input, MLLMs with video input, and human experts. This comparison provides several critical insights into the effectiveness of visual reasoning in video and highlights the significant gap between human cognitive abilities and current MLLMs.

LLM without visual information gives rise to low accuracy on VideoAds: To assess the reliance on visual features, we conduct a baseline test using GPT-4o without any video input. The model achieves an accuracy of 21.27%, which is close to the 25% random guess baseline for a four-option multiple-choice format. This result confirms that VideoAds is effectively designed to eliminate trivial text-based shortcuts, ensuring that visual reasoning is indispensable for answering the questions. It also demonstrates that commonsense knowledge and language priors alone are insufficient to solve the complex visual and temporal reasoning tasks posed by this benchmark.

Human performance substantially surpasses state-of-the-art MLLMs: Human experts consistently outperform all tested MLLMs, achieving an impressive accuracy of 94.27%. This highlights the inherent cognitive advantage of humans in narrative comprehension, causal inference, and high-level reasoning. Notably, humans perform best on the Visual Reasoning tasks, achieving 95.04% accuracy, demonstrat-

Model	LLM params.	Frames	Finding	Summary	Reasoning	Overall
Baseline without visual information						
Baseline (GPT4o-text only [12])	-	-	21.88	21.47	20.39	21.27
Open-source MLLMs						
LongVA [47]	7B	32	49.41	40.38	37.19	42.33
Qwen2.5-VL [33]	7B	32	60.47	58.97	44.63	54.69
LLaVA-OneVision [15]	7B	32	67.76	52.56	44.90	55.08
MiniCPM-V2-6 [43]	7B	64	67.53	53.85	50.14	57.17
InternVideo2-chat [34]	8B	16	42.82	35.58	41.87	40.09
InternVL2 [5]	8B	32	53.88	44.55	46.56	48.33
LLaVA-NeXT-Video [48]	32B	32	63.06	60.26	53.44	58.92
LLaVA-Video [48]	72B	32	66.35	68.91	64.46	66.58
		64	71.06	66.03	66.12	67.73
		96	72.94	66.67	67.22	68.94
		128	72.94	66.03	66.39	68.45
Qwen2.5-VL [33]	72B	32	67.76	63.78	61.43	64.33
		64	74.59	69.87	66.67	70.38
		96	73.65	73.08	69.70	72.14
		128	75.53	73.72	70.80	73.35
Close-source MLLMs						
GPT-4o [12]	-	32	70.59	66.35	59.78	65.57
		64	73.65	64.42	59.50	65.86
		128	75.06	63.14	62.26	66.82
Gemini-1.5 Pro [28]	-	1 fps	75.29	67.31	66.39	69.66
Gemini-2.5 Pro [6]	-	-	84.47	78.53	77.13	80.04
Human Performance						
Human Performance	-	-	93.41	94.55	95.04	94.27

Table 3. Benchmark results for different MLLMs. We can observe that human performance substantially surpasses all SOTA MLLMs, while Gemini 2.5 Pro leads the best performance of 80.04%. Some prediction cases are shown in the Appendix.

ing their ability to synthesize information across disjointed scenes and understand implicit messaging strategies typical of advertisement videos. In stark contrast, most MLLMs struggle to reach even 70% accuracy on the reasoning task, with the best-performing model, Qwen2.5-VL-72B, attaining 70.80% accuracy, despite its large model size and state-of-the-art architecture. This substantial gap between human and model performance also highlights the key challenges raised by our VideoAds dataset.

Current MLLMs excel in static visual recognition but struggle with complex reasoning, while humans give contradictory results: From Table 3, a clear performance disparity emerges across the three task types: visual finding, visual summary, and visual reasoning. Specifically, most MLLMs perform the best on the visual finding tasks, followed by visual summary, with the worst performance observed on Visual Reasoning tasks. For example, GPT-4o achieves an accuracy of 75.06% on the visual finding tasks but drops significantly to 62.26% on the visual reasoning tasks with 128 frames. This pattern is consistent across other models, such as LLaVA-Video and Qwen2.5-VL under different sampling frames. This performance gradient suggests that current MLLMs are more adept at static visual recognition and shallow temporal tracking but struggle with tasks requiring complex narrative reasoning. We also observe an interesting result for human performance in Table 3 that the performance on the most challenging reasoning tasks (accuracy 95.04%) is even higher than the visual finding (accuracy 93.41%). This is reasonable given that human experts find it easier to find the visual reasoning structures in advertisement videos at a high level but find it hard to remember every detail.

State-of-the-art open source MLLMs can beat GPT-4o and Gemini 1.5 Pro, while Gemini 2.5 Pro continues to lead performance: For a long time commercial LLMs [12, 22, 31] can easily beat all opensource models in video understanding. VideoAds, however, produces the opposite result, with Qwen2.5-VL outperforming GPT-4o and Gemini 1.5 Pro in Table 3. GPT-4o and Gemini 1.5 Pro can still achieve SOTA results (above 75%) on the visual finding task but struggle with the visual summary and reasoning (around 66%-67%). However, Gemini 2.5 Pro, the most recent model, keep leading the performance with an accuracy of 80.04%. This further emphasizes that rather than taking video as a grid of images, it is necessary to train MLLMs on video data specifically to extract the temporal

relationship within the video like used in Gemini 2.5 Pro [6].

Influence of increasing sampled frames varies across the models: When a video is simple, a limited set of sampled frames can include all necessary information. However, in complex videos like those in VideoAds, increasing the number of frames can significantly impact performance. We observe that Qwen2.5-VL benefits more from additional frames than GPT-4o and LLaVA-Video. The latter two models show notable improvements in visual finding tasks but only marginal gains in reasoning tasks when we increased the number of frames per video from 32 to 128. We hypothesize that, due to high video complexity in VideoAds, it actually requires the MLLMs to process high fps and long context tokens, which are underexplored in the previous work [12, 28, 33, 48].

5. Discussion

5.1. Impact of Speech Audio Transcript

To evaluate the influence of audio transcripts on model performance, we generate subtitles for each video using OpenAI Whisper [26]. We further test the performance of the models with transcripts and the results are shown in Table 4. Comparing transcript-enhanced models to those using only frame-based information, we observe that introducing audio transcripts significantly improves performance across all MLLMs. This suggests that many MLLMs for video are focusing on visual information, while the important cross-modality reasoning capability remains underexplored.

Model	Modality	Finding	Summary	Reasoning	Overall
LongVA	frames	49.41	40.38	37.19	42.33
	+ subtitle	50.35	48.72	45.45	48.18
LLaVA-Video	frames	66.35	68.91	64.46	66.58
	+ subtitle	71.53	71.47	72.18	71.73
Qwen2.5-VL	frames	67.76	63.78	61.43	64.33
	+ subtitle	68.94	67.95	65.56	67.48
GPT-4o	frames	73.65	64.42	59.50	65.86
	+ subtitle	72.24	64.10	63.09	66.47

Table 4. Impact of audio transcript: we can observe that a significant performance gain, particularly in the reason tasks can be achieved by various MLLMs with the help of audio transcript.

5.2. Impact of Chain of Thought

The research on the influence of CoT on the multimodality data remains limited, and most studies focus on the image data [4, 36]. Here we present the first study of CoT's influence on video MLLMs using our challenging dataset. In our CoT prompting, we provided explicit intermediate questions to guide reasoning, rather than using generic 'think step-by-step' instructions. For example, for a visual reasoning task, we might pose intermediate questions like, 'Q1: What product appears in the first 5 seconds?' followed by 'Q2: What emotion does the character express at the end?' The model's answers to these intermediate questions were then used as context to generate the final response to the main question. We adopted this guided setup because many current MLLMs do not reliably support open-ended, multi-round conversational reasoning from a single generic prompt.

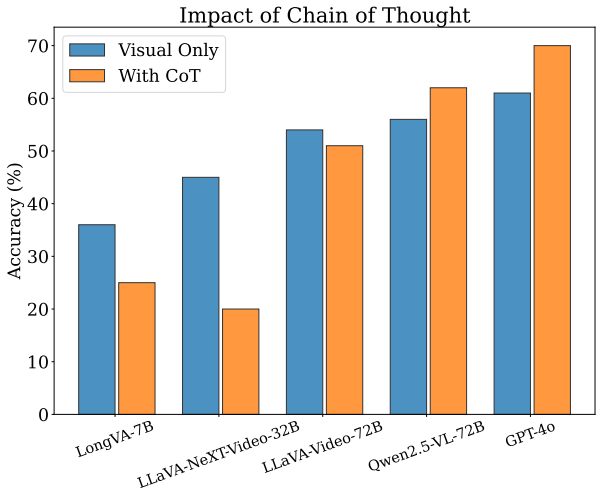

Figure 4. Impact of Chain of thought on the model's performance for challenging reasoning tasks, we can observe a variance in the model's performance along the model size.

Interestingly, the influence of CoT differs across models: Generally, larger models tend to benefit more from the CoT as shown in the NLP field [9, 35]. We observe that LongVA-7B [47], LLava-Next-32B [48], and LLava-Video-72B [48] show a performance drop when dealing long reasoning contexts provided by CoT. However, models like Qwen2.5-VL [33] and GPT-4o [11] show significant performance improvements. Particularly, GPT-4o increases its reasoning accuracy from 61% to 70%, indicating strong reasoning performance given CoT. This also underscores the importance of introducing long context and multi-round VQA in MLLM training [4, 36].

6. Conclusion

In this work, we introduce VideoAds, the first benchmark dataset specifically designed to evaluate Multi-modality Large Language Models (MLLMs) on complex temporal reasoning and narrative understanding in advertisement videos. Our novel quantitative complexity metric provides a structured framework for evaluating temporal dynamics in video benchmarking. Extensive benchmarking experiments show that there is still a significant gap between human cognitive reasoning and the ability of current MLLMs to understand complex temporal structures like advertisement videos, highlighting the urgent need for advanced temporal modeling and narrative comprehension techniques.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[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. 1

[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. *ArXiv preprint*, 2023. 4

[4] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461*, 2023. 8

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *ArXiv preprint*, 2024. 6, 7

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6, 7, 8

[7] 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. 1, 2, 3, 4, 5

[8] 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. 5, 6

[9] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022. 8

[10] 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

[11] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 5, 8

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3, 6, 7, 8

[13] Joan Ikonomi. Tv commercial: Concept and production structure. *Interdisciplinary Journal of Research and Development*, 6(1):99–99, 2019. 3

[14] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 4

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 7

[16] 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. 1, 2, 4

[17] 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, 6

[18] 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. 2, 4, 5

[19] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 5

[20] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2024. 2, 4, 5

[21] 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. 4

[22] OpenAI. GPT-4V(ision) system card, 2023. 1, 7

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *ArXiv preprint*, 2023. 5

[24] Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. Reka core, flash, and edge: A series of powerful multimodal language models. *ArXiv preprint*, 2024. 1

[25] 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. 5

[26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recog-

nition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 8

[27] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, 2024. 1

[28] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, and etc. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, 2024. 6, 7, 8

[29] William Smethurst. *How to Write for Television 6th Edition: A guide to writing and selling successful TV Scripts*. How To Books, 2009. 3

[30] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023. 1

[31] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3, 7

[32] 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. 5

[33] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 7, 8

[34] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 6, 7

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022. 8

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 8

[37] Wikipedia contributors. No. 5 the film — Wikipedia, the free encyclopedia, 2024. [Online; accessed 7-March-2025]. 3

[38] Bo Wu and Shoubin Yu. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2024. 4

[39] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 4

[40] 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. 4, 5

[41] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 5

[42] Keng-Chieh Yang, Chia-Hui Huang, Conna Yang, and Su Yu Yang. Consumer attitudes toward online video advertisement: Youtube as a platform. *Kybernetes*, 46(5):840–853, 2017. 1

[43] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 7

[44] 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. 4, 5

[45] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024. 1

[46] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 6

[47] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 6, 7, 8

[48] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 1, 3, 5, 6, 7, 8

[49] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 6

[50] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5