MME-VideoOCR: Evaluating OCR-Based Capabilities of Multimodal LLMs in Video Scenarios

Yang Shi[1,2◇*] **Huanqian Wang**[3◇] **Wulin Xie**[4◇] **Huanyao Zhang**[2◇] **Lijie Zhao**[5◇]
Yi-Fan Zhang[4◇‡] **Xinfeng Li**[6] **Chaoyou Fu**[7] **Zhuoer Wen**[2] **Wenting Liu**[2]
Zhuoran Zhang[2] **Xinlong Chen**[4] **Bohan Zeng**[2] **Sihan Yang**[8] **Yushuo Guan**[1]
Zhang Zhang[4] **Liang Wang**[4] **Haoxuan Li**[2] **Zhouchen Lin**[2]
Yuanxing Zhang[1‡] **Pengfei Wan**[1] **Haotian Wang**[3‡] **Wenjing Yang**[♠]
[1]Kling Team [2]PKU [3]THU [4]CASIA [5]CUHKSZ [6]NTU [7]NJU [8]XJTU
[◇] Core Contributor [♠] Project Lead [‡] Corresponding Author
https://mme-videoocr.github.io/

Abstract

Multimodal Large Language Models (MLLMs) have achieved considerable accuracy in Optical Character Recognition (OCR) from static images. However, their efficacy in video OCR is significantly diminished due to factors such as motion blur, temporal variations, and visual effects inherent in video content. To provide clearer guidance for training practical MLLMs, we introduce **MME-VideoOCR** benchmark, which encompasses a comprehensive range of video OCR application scenarios. MME-VideoOCR features 10 task categories comprising 25 individual tasks and spans 44 diverse scenarios. These tasks extend beyond text recognition to incorporate deeper comprehension and reasoning of textual content within videos. The benchmark consists of $1,464$ videos with varying resolutions, aspect ratios, and durations, along with $2,000$ meticulously curated, manually annotated question-answer pairs. We evaluate 18 state-of-the-art MLLMs on MME-VideoOCR, revealing that even the best-performing model (Gemini-2.5 Pro) achieves only an accuracy of 73.7%. Fine-grained analysis indicates that while existing MLLMs demonstrate strong performance on tasks where relevant texts are contained within a single or few frames, they exhibit limited capability in effectively handling tasks that demand holistic video comprehension. These limitations are especially evident in scenarios that require spatio-temporal reasoning, cross-frame information integration, or resistance to language prior bias. Our findings also highlight the importance of high-resolution visual input and sufficient temporal coverage for reliable OCR in dynamic video scenarios.

1 Introduction

In recent years, the rapid advancement of Multimodal Large Language Models (MLLMs)[1–6] has garnered significant attention. These models, capable of processing and integrating information across various modalities (e.g., text, images, and video), have demonstrated considerable potential and significant value across a wide range of real-world applications[7–14].

Optical Character Recognition (OCR) [15], a fundamental technology in visual understanding, serves as a crucial link for enabling structured comprehension of image and video content. It transforms visual information into computationally analyzable semantic data. Within cross-modal learning, OCR provides critical feature support for text-visual alignment, directly impacting the performance of

[*] Work done during an internship at Kling Team.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

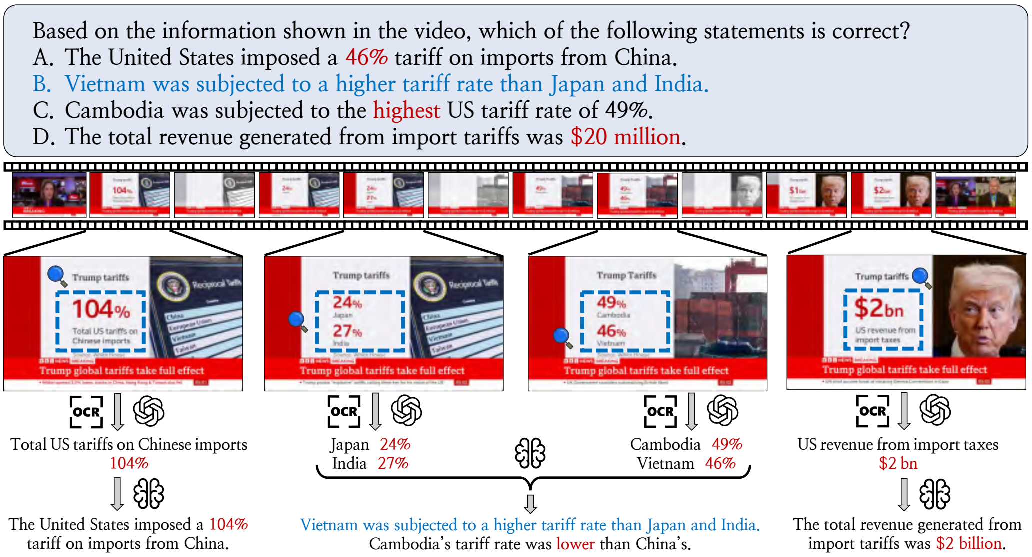

Based on the information shown in the video, which of the following statements is correct?
A. The United States imposed a 46% tariff on imports from China.
B. Vietnam was subjected to a higher tariff rate than Japan and India.
C. Cambodia was subjected to the highest US tariff rate of 49%.
D. The total revenue generated from import tariffs was $20 million.
Trump tariffs
104%
Total US tariffs on Chinese imports
24%
Japan
27%
India
49%
Cambodia
46%
Vietnam
$2bn
US revenue from import taxes
Trump global tariffs take full effect
OCR
Total US tariffs on Chinese imports
104%
Japan 24%
India 27%
Cambodia 49%
Vietnam 46%
US revenue from import taxes
$2 bn
The United States imposed a 104% tariff on imports from China.
Vietnam was subjected to a higher tariff rate than Japan and India.
Cambodia's tariff rate was lower than China's.
The total revenue generated from import tariffs was $2 billion.

Figure 1: **An example in MME-VideoOCR**. The task requires the MLLM to first recognize the textual information distributed across multiple video frames, and then to perform semantic understanding and reasoning over the extracted text to accurately determine the correct answer. The correct information is marked in blue, while misleading information is marked in red.

Table 1: **Key differences between MME-VideoOCR and prior video-based OCR benchmarks**. MME-VideoOCR features a larger number of task types and scenarios, employs fully manual annotations to ensure reliability, supports bilingual content for broader coverage, and enables comprehensive evaluation across perception, understanding, and reasoning.

Benchmarks	#Videos	#QA	#Tasks	#Scenarios	Annotation	Bilingual	Perception	Understanding	Reasoning
OCR Benchmark [23]	25	1,477	1	20+	M	✗	✓	✗	✗
FG Bench [24]	1,028	2,961	6	20+	A&M	✗	✓	✓	✗
MME-VideoOCR	1464	2,000	25	44	M	✓	✓	✓	✓

downstream tasks [16, 17]. Previous OCR-based benchmarks [18–22] primarily focus on evaluating the OCR-based capabilities of MLLMs in static image scenarios. Several studies [23, 24] have initiated preliminary investigations into video scenarios. However, they typically concentrate on perceiving textual content, often neglecting text-based understanding and reasoning.

Considering the unique challenges of video understanding tasks, a comprehensive video OCR evaluation must address three key issues, as illustrated in Figure 1. Firstly, textual information in videos can appear in various forms—such as foreground text, background scenery, on-screen annotations, watermarks, and floating overlays. This requires models to establish robust spatio-temporal visual-text associations and to effectively identify and extract relevant textual information from these diverse and often noisy sources across different shots. Secondly, critical textual information in videos is often distributed across multiple frames, rather than appearing in a single static image. Therefore, models must be capable of effectively recognizing, integrating, and understanding text content over time, leveraging temporal context to reconstruct and interpret fragmented or sequentially presented information. Thirdly, as task complexity increases, models must be able to reason over the recognized text. This reasoning ability is essential for deeper video understanding and remains a significant challenge for current MLLMs.

In this paper, we propose the MME-VideoOCR benchmark, which provides a comprehensive evaluation framework for OCR tasks in video scenarios. Recognizing the limitations of current OCR tasks in existing evaluation datasets, MME-VideoOCR encompasses 10 task categories and 25 specific tasks, incorporating a substantial number of actively collected or custom-created videos. As shown in Table 1, MME-VideoOCR consists of $1,464$ videos, paired with $2,000$ diverse and accurately human-annotated question-answer (QA) pairs. The tasks require answers based on both localized key information and a holistic understanding of the entire video.

The main contributions are summarized as follows:

1. MME-VideoOCR introduces a diverse set of video OCR tasks, utilizing manually quality-controlled videos and question-answer pairs. These tasks span multiple dimensions, such as perceptual accuracy, contextual comprehension, and cross-frame reasoning, which together enable a comprehensive evaluation of MLLMs' OCR capabilities in video scenarios.
2. We evaluate 18 state-of-the-art MLLMs, including publicly available models ranging from 7B to 78B in size, as well as closed-source models like GPT-4o and Gemini-2.5 Pro. The results demonstrate strong discriminative power and the challenges posed by MME-VideoOCR. Regarding discriminative power, the worst-performing model, LLaVA-OneVision 7B, has an accuracy of 46.0%, while the best-performing model achieves an accuracy of 73.7%, showing a significant gap in performance. Regarding task difficulty, on several tasks we designed, such as Cross-Frame Text Understanding and Text-Based Video Understanding, most models score below 60%.
3. The evaluation results further reveal significant deficiencies in current models on OCR tasks that require spatio-temporal reasoning and cross-frame information integration, thereby indicating a critical direction for MLLM optimization. Moreover, both high-resolution visual inputs and sufficient temporal coverage are essential for achieving reliable OCR performance in dynamic video settings. Notably, MLLMs exhibit a strong language prior bias during text recognition, frequently favoring semantically plausible outputs over visually accurate transcriptions.

2 Related Work

2.1 Multimodal Large Language Models

Through the integration of a vision encoder into Large Language Models (LLMs) and pretraining on large-scale multimodal data, MLLMs exhibit strong OCR capabilities, making them well-suited for downstream tasks such as document understanding [25, 26], key information extraction [17], and scene text recognition [27, 28]. Building on this foundation, some recent MLLMs [6, 29–33] have further extended their capabilities to handle video inputs, enabling them to process dynamic visual information. This advancement enables MLLMs not only to recognize text in static images, but also to extract text-related information from videos and leverage it for more effective video understanding. However, the increased visual complexity, dynamic content, and temporal dependencies inherent in video characteristics impose greater demands and challenges on MLLMs [3, 34]. Therefore, a comprehensive and effective evaluation of their OCR-based capabilities in video scenarios is crucial.

2.2 OCR Benchmarks for Multimodal Large Language Models

Most existing benchmarks [35] are designed to evaluate the OCR capabilities of MLLMs in static image scenarios, including TextVQA [18], OCR-VQA [19], SEED-Bench2-Plus [20], OCRBench [21] and OCRBench v2 [22]. A few works [23, 24] have extended to video, but they cover only a narrow range of task types, lack diversity in video content, and provide limited insight into the unique characteristics of OCR-based tasks in video scenarios. Moreover, these benchmarks emphasize text recognition while overlooking text-based understanding and reasoning. Some scene-text video QA benchmarks [27, 36, 37] incorporate textual cues into visual QA. However, they often overlook fine-grained text perception, including temporal grounding and attribute recognition, and do not fully evaluate the potential of text as a central driver for video understanding. Moreover, as they focus solely on scene-text understanding, which represents only a narrow application scenario, this is far from sufficient for a comprehensive evaluation of MLLMs' OCR-based capabilities. These benchmarks are also limited in video diversity, task variety, and their exploration of the unique characteristics of OCR-based tasks in video scenarios.

3 MME-VideoOCR

3.1 Task Definition

Challenges inherent in video data, such as motion blur, inter-frame interference, the complexity of cross-modal alignment, difficulties in tracking content across shots, and limited generalization in noisy scenes. These issues pose significant obstacles for both video coding [38, 39] and semantic perception [40, 41], critically impacting the accuracy of MLLMs. To rigorously assess and foster

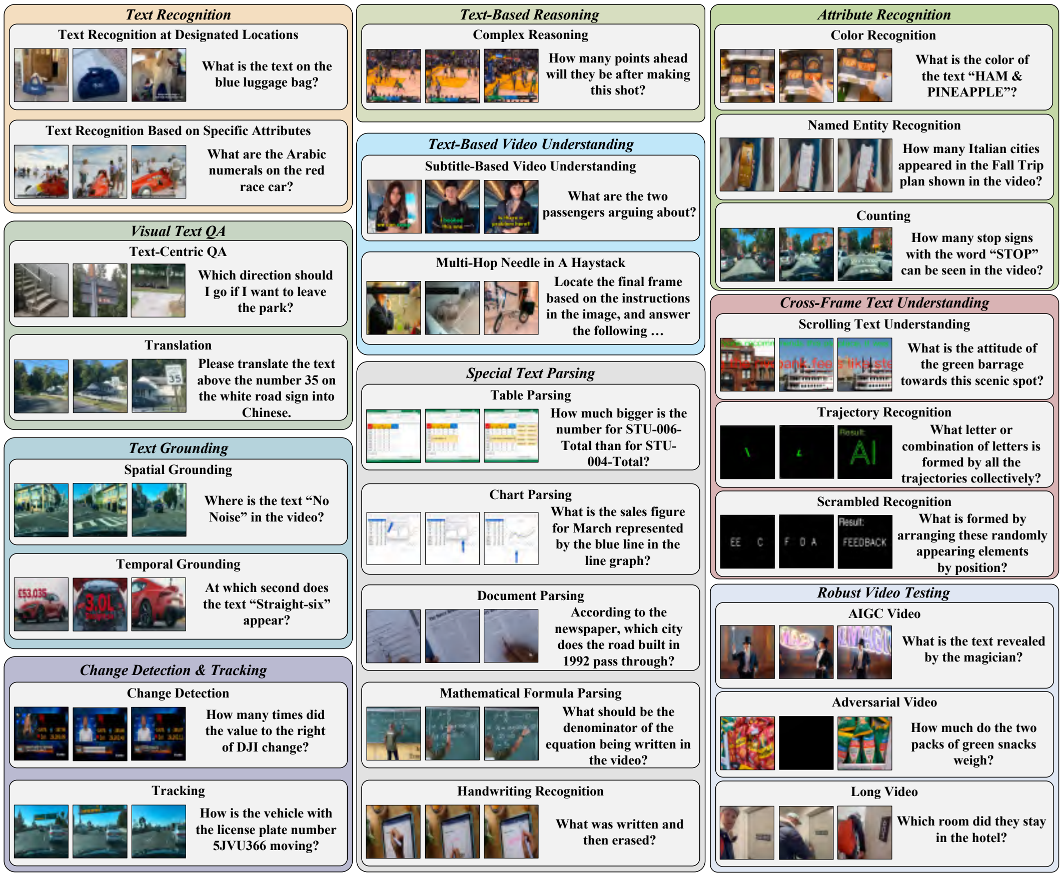

Text Recognition
Text Recognition at Designated Locations
What is the text on the blue luggage bag?
Text Recognition Based on Specific Attributes
What are the Arabic numerals on the red race car?
Visual Text QA
Text-Centric QA
Which direction should I go if I want to leave the park?
Translation
Please translate the text above the number 35 on the white road sign into Chinese.
Text Grounding
Spatial Grounding
Where is the text "No Noise" in the video?
Temporal Grounding
At which second does the text "Straight-six" appear?
Change Detection & Tracking
Change Detection
How many times did the value to the right of DJI change?
Tracking
How is the vehicle with the license plate number 5JVU366 moving?
Text-Based Reasoning
Complex Reasoning
How many points ahead will they be after making this shot?
Text-Based Video Understanding
Subtitle-Based Video Understanding
What are the two passengers arguing about?
Multi-Hop Needle in A Haystack
Locate the final frame based on the instructions in the image, and answer the following ...
Special Text Parsing
Table Parsing
How much bigger is the number for STU-006-Total than for STU-004-Total?
Chart Parsing
What is the sales figure for March represented by the blue line in the line graph?
Document Parsing
According to the newspaper, which city does the road built in 1992 pass through?
Mathematical Formula Parsing
What should be the denominator of the equation being written in the video?
Handwriting Recognition
What was written and then erased?
Attribute Recognition
Color Recognition
What is the color of the text "HAM & PINEAPPLE"?
Named Entity Recognition
How many Italian cities appeared in the Fall Trip plan shown in the video?
Counting
How many stop signs with the word "STOP" can be seen in the video?
Cross-Frame Text Understanding
Scrolling Text Understanding
What is the attitude of the green barrage towards this scenic spot?
Trajectory Recognition
What letter or combination of letters is formed by all the trajectories collectively?
Scrambled Recognition
What is formed by arranging these randomly appearing elements by position?
Robust Video Testing
AIGC Video
What is the text revealed by the magician?
Adversarial Video
How much do the two packs of green snacks weigh?
Long Video
Which room did they stay in the hotel?

Figure 2: **Example videos and their annotated questions from the MME-VideoOCR benchmark**, encompassing 25 tasks across 10 categories. Each task is designed to evaluate models' capabilities in various aspects such as text recognition, localization, reasoning, and comprehensive video understanding. The figure displays representative video samples and their corresponding questions.

advancements in MLLMs against these challenges, we introduce MME-VideoOCR, a comprehensive benchmark comprising 25 distinct tasks across 10 categories (details can be found in Appendix B.1). Figure 2 showcases representative examples, illustrating the specific nature and scope of each task.

Text Recognition involves *Text Recognition at Designated Locations* and *Text Recognition Based on Specific Attributes* to evaluate the fine-grained text recognition capability.

Visual Text QA employs *Text-Centric QA* and *Translation*. Both tasks challenge the model's ability to not only perceive but also comprehend multimodal semantics.

Text Grounding introduces *Spatial Grounding* and *Temporal Grounding* to assess the model's ability on localizing and interpreting text across both spatial-temporal dimensions within dynamic scenes.

Attribute Recognition is composed of three tasks: *Color Recognition*, where models are expected to identify the color of the text; *Named Entity Recognition*, which focuses on extracting and classifying named entities; and *Counting*, where models must accurately identify the number of textual elements that meet specified criteria.

Change Detection & Tracking contains *Change Detection* and *Tracking* to identify textual changes over time and monitor text elements as they change position across frames, respectively.

Special Text Parsing includes five tasks: *Table Parsing*, *Chart Parsing*, *Document Parsing*, *Mathematical Formula Parsing*, and *Handwriting Recognition*. These tasks require models to accurately identify and understand text with either special structures or highly variable visual forms.

Cross-Frame Text Understanding includes three subtasks: *Scrolling Text Understanding*, which focuses on recognizing dynamic text streams that move across frames and may only be fully readable when aggregated over time; *Trajectory Recognition*, where the motion path of an object in the

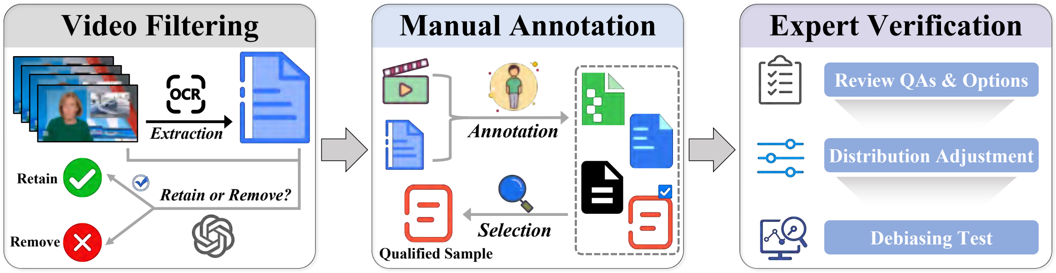


Figure 3: **Overview of the MME-VideoOCR construction process**. Video filtering ensures sufficient visual dynamics and meaningful textual content. Manual annotation provides high-quality QA pairs, and expert verification further enhances sample reliability and mitigates potential biases.

video forms a recognizable text, and the model must interpret this trajectory as the intended message; *Scrambled Recognition*, which involves identifying and reconstructing a complete text from characters that appear out of order across different positions in the video frames.

Text-Based Reasoning requires models to go beyond surface-level understanding by synthesizing dispersed cues, identifying implicit relation, and resolving ambiguity or misleading content.

Text-Based Video Understanding introduce a) *Subtitle-Based Video Understanding* which resembles real-world scenarios like conversations, tutorials, or news, where subtitles provide key information that visuals alone cannot capture; b) *Multi-Hop Needle in A Haystack* which requires reasoning over multiple pieces of subtitle content to find the correct answer.

Robust Video Testing contains three specialized video types: *AIGC Videos*, *Long Videos*, and *Adversarial Videos*. *AIGC Videos*, generated by AI systems [42], assess model adaptability to increasingly common synthetic content. *Long Videos* test the ability to extract relevant information from lengthy sequences with substantial redundancy. *Adversarial Videos* strategically insert all-black frames into normal videos, designed to mislead the MLLMs.

3.2 Benchmark Construction

Video Collection & Filtering. MME-VideoOCR covers as many diverse scenarios as possible in order to provide a comprehensive evaluation. To achieve this, we employ three distinct data collection methods, balancing diversity and efficiency in the construction of the benchmark.

Reconstructing from Existing Video QA Data. To maximize data collection efficiency, we leverage existing text-based video QA datasets, including BOVText [43], M4-ViteVQA [36], NewsVideoQA [44], LSVTD [45], RoadText-1K [46], RoadTextVQA [37], EgoTextVQA [27], NIAH-Video [47], and DSText [48]. For each video in these datasets, we uniformly sample 5-10 frames and extract the text using PaddleOCR [49]. The sampled frames, along with the extracted text, are then processed using GPT-4o to evaluate whether the video exhibits sufficient visual dynamics and contains semantically meaningful text. Only videos that meet these criteria are retained for further use.

Manual Collection of Publicly Available Videos. Existing benchmarks often lack the diversity needed to fully satisfy the requirements of our 25 OCR tasks. Therefore, we manually collect additional data from publicly available online sources (e.g., YouTube, Bilibili, Kuaishou) to further enhance diversity and ensure coverage of specific scenarios that are underrepresented in current datasets, such as webpages, charts, and mathematical formula derivations. Additionally, since most existing MLLMs are primarily trained on horizontally oriented videos, we intentionally include vertically formatted video content to improve distributional balance and better reflect real-world usage scenarios.

AI-Generated Videos. The task of recognizing and understanding the text in AI-generated videos is becoming more critical. To cover this emerging scenario, we manually create a set of videos designed to diversify the dataset and introduce controlled challenges. We initially generated $2,000$ everyday phrases. These phrases were then expanded into scene descriptions using Llama3.1-8B [50], with the requirement that each scene must incorporate the corresponding text and include a narrative element detailing its appearance or disappearance. Subsequently, these descriptions were provided to Wan [42] for text-to-video generation. From the resultant videos, we selected those exhibiting accurate text rendering, high visual-scene integration, and plausible narratives for our evaluation set. These videos are not only useful for evaluating the model's ability to understand AI-generated content but also address specific cases that are difficult to obtain from existing datasets or online sources, such as occluded text revealed over time.

Manual Annotation. In order to circumvent errors and biases that may arise from model-based annotations [51, 52], we opt for manual annotation to ensure the dependability of our samples. Human annotators are tasked with carefully examining each video and developing 3-4 QA pairs per video, adhering to the specified task requirements. Next, a second expert implements a selection process, retaining 1-2 high-quality QA pairs per video. This sequential two-stage screening process is expected to substantially ensure the generation of high-quality QA pairs exhibiting significant relevance, clarity, and challenging attributes, effectively preventing biases from individual annotators.

Expert Verification. To uphold the highest data quality, expert annotators meticulously verify the constructed dataset against stringent standards. This verification process specifically addresses potential issues such as *ambiguous questions*, *inaccurate answers*, and *insufficiently challenging problems*. Initially, annotators review and rectify any errors or ambiguities within the QA pairs. Subsequently, for multiple-choice questions, they thoroughly assess all options, confirming that each is meaningful, poses an appropriate level of challenge, and functions as a plausible distractor. To mitigate potential biases stemming from imbalanced answer option frequencies [53, 54], we ensure a uniform distribution of correct answers across all options. Furthermore, to identify and eliminate residual biases that could compromise evaluation reliability, we conduct a dedicated debiasing test, as detailed in Section 3.4. The complete data construction process is illustrated in Figure 3.

3.3 Evaluation Criteria

Considering the characteristics of different tasks, we employ three distinct evaluation metrics to balance accuracy and efficiency in evaluation.

Containment Match. For *Text Recognition* and *Handwriting Recognition*, where the model must accurately identify the recognized text, we simply check whether the ground truth appears in model's response. This straightforward yet effective method is widely adopted in previous work [21, 24, 55].

GPT-Assisted Scoring. In the *Translation* task, multiple valid answers may exist. These answers may vary in form but remain consistent in meaning. To ensure flexibility and prevent unnecessary constraints on the model, we incorporate GPT-Assisted Scoring. Given the reference answer and the model's response, GPT-4o-0806 [56] serves as the evaluator, assessing their consistency and assigning a binary score of either 0 or 1. The prompt is shown in Appendix B.3.

Multiple-Choice. Tasks like *Visual Text QA* and *Spatial Grounding* allow for highly flexible responses. Since both Containment Match and GPT-Assisted Scoring may introduce evaluation errors, we use a multiple-choice format for assessment. In this setting, the model only needs to select the most appropriate option, which simplifies evaluation and reduces ambiguity in scoring. We use a common prompt as shown in Appendix B.3.

3.4 Debiasing Test

Since the underlying LLM of an MLLM is pretrained on large-scale textual corpora, it may introduce biases into the evaluation [53]. One source of bias arises from textual priors. For example, when asked "`What does the text on the red warning sign say?`", the model is more likely to answer "‘`STOP`" than "`EXIT`" due to common co-occurrence patterns in pretraining data. Another issue is potential knowledge leakage. For instance, for a question like "`According to the video, when was the United States founded?`", the model may provide the correct answer without relying on visual input, thereby compromising the reliability of the evaluation.

To mitigate these biases, we introduce a debiasing test designed to quantify and minimize the influence of textual priors and knowledge leakage. Specifically, we evaluate Qwen2.5-VL-7B [30] by presenting questions and options without providing any meaningful visual input. Under this setup, the model relying solely on textual priors should ideally achieve accuracy close to 0% for tasks evaluated via Containment Match and GPT-Assisted Scoring, indicating minimal textual bias. For multiple-choice questions, random guessing should yield an accuracy close to 25%. After each debiasing test iteration, expert annotators review and revise samples flagged as potentially problematic. As summarized in Table 2, the final results demonstrate the effectiveness of our approach, confirming that textual biases have been significantly suppressed, thus ensuring greater reliability and fairness of our evaluation.

Table 2: **Accuracy of the debiasing test**. Through multiple rounds of testing and revision, potential biases were effectively suppressed, ensuring the validity and reliability of MME-VideoOCR.

Model	Visual Input	Containment Match	GPT-Assisted Scoring	Multiple-Choice
Qwen2.5-VL	None	0%	0%	25.1%
Qwen2.5-VL	Black Image	0%	0%	27.4%

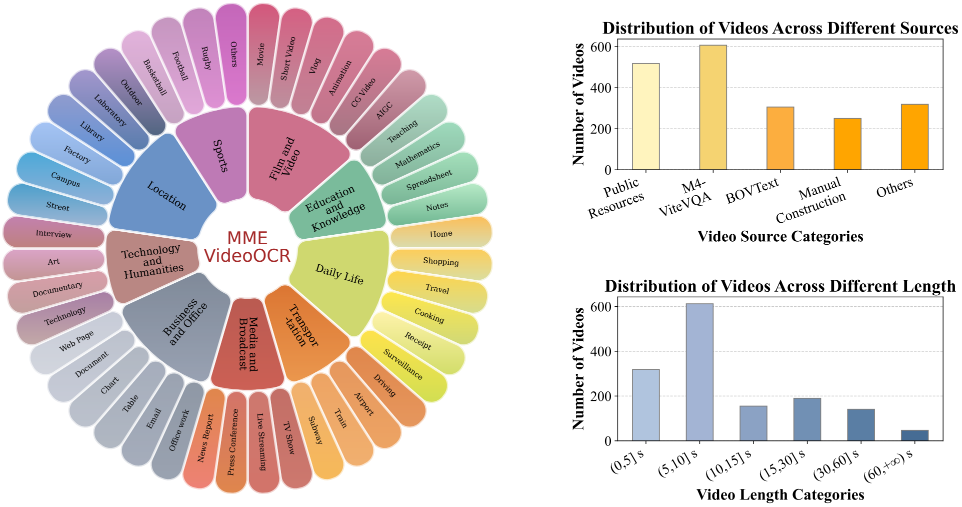


Figure 4: **Overview of MME-VideoOCR Statistics**. The videos in MME-VideoOCR covers 9 major scenario categories comprising 44 specific scene types, offering fine-grained coverage of diverse video contexts. The benchmark features a balanced distribution of video durations and sources, with a significant portion of the videos newly collected from public resources or manually curated.

3.5 Statistics

Through rigorous video selection, manual annotation, and expert-level validation, we collect a total of $1,464$ videos along with $2,000$ high-quality QA annotations. As illustrated in Figure 4, these videos span 9 major scenario categories, such as daily life, education and knowledge, and sports, encompassing 44 specific scenarios. The videos vary considerably in duration, resolution, and aspect ratio. They originate from diverse sources, with a substantial proportion newly collected from public resources or manually constructed.

4 Experiments

We evaluate a total of 18 mainstream MLLMs, including 3 cutting-edge closed-source models and 15 open-source models. The closed-source models involve GPT-4o [56], Gemini-2.5 Pro [57] and Gemini-1.5 Pro [5]. In selecting open-source models, we consider two factors. First, models are categorized by parameter size into small (7B/8B), medium (16B/32B/38B), and large (72B/78B) groups. Second, models are differentiated by their video processing strategies, including **(a)** sparse frame sampling (InternVL3 [31], LLaVA-OneVision [58], VITA-1.5 [2], LLaVA-Video [29], Kimi-VL [59], Qwen2.5-VL [30], Oryx-1.5 [60]), **(b)** dense sampling with token compression (VideoLLaMA 3 [61], VideoChat-Flash [47]), and **(c)** the slow-fast frame sampling approach (Slow-fast MLLM [62]). Please refer to Appendix C for details of the experimental setup.

4.1 Main Results

We evaluate the performance of all baseline models on MME-VideoOCR and display the accuracy for each task category and the overall accuracy, as shown in Table 3. Our observations indicate that among the 18 evaluated models, Gemini-2.5 Pro is the top performer, yet achieves an accuracy of only 73.7%. Concurrently, five models that demonstrate strong performance on other video understanding tasks achieved an accuracy below 50% on MME-VideoOCR. This performance landscape underscores the challenging nature and discriminative capability of the MME-VideoOCR benchmark.

Table 3: **Evaluation results on MME-VideoOCR**. “TR” denotes Text Recognition, “VTQA” Visual Text QA, “TG” Text Grounding, “AR” Attribute Recognition, “CDT” Change Detection & Tracking, “STP” Special Text Parsing, “CFTU” Cross-Frame Text Understanding, “TBR” Text-Based Reasoning, “TBVU” Text-Based Video Understanding, and “RVT” Robust Video Testing. The highest accuracy of each task is in red, and the second highest is underlined.

Model	**Size**	**TR**	**VTQA**	**TG**	**AR**	**CDT**	**STP**	**CFTU**	**TBR**	**TBVU**	**RVT**	**Total**
Closed-source MLLMs												
Gemini-1.5 Pro	-	76.7%	77.6%	61.5%	64.7%	55.0%	74.0%	31.3%	68.7%	53.5%	68.0%	64.9%
GPT-4o	-	83.3%	81.6%	60.5%	74.7%	51.5%	68.0%	30.7%	60.7%	59.0%	75.3%	66.4%
Gemini-2.5 Pro	-	83.0%	91.6%	64.5%	74.0%	70.0%	84.4%	48.7%	74.0%	56.5%	72.0%	73.7%
Small-scale MLLMs												
LLaVA-OneVision	7B	42.0%	50.0%	49.0%	54.0%	41.0%	46.4%	20.0%	45.3%	52.0%	60.0%	46.0%
VideoChat-Flash	7B	36.7%	48.0%	60.0%	60.0%	49.0%	46.0%	19.3%	50.0%	54.0%	60.7%	47.8%
Slow-fast MLLM	7B	46.0%	54.8%	52.0%	60.0%	47.0%	48.0%	20.0%	43.3%	48.5%	54.0%	47.8%
VITA-1.5	7B	49.0%	58.4%	43.0%	61.3%	49.0%	53.2%	20.0%	51.3%	47.0%	58.7%	49.5%
Oryx-1.5	7B	51.7%	54.0%	50.5%	54.7%	44.5%	52.8%	23.3%	48.7%	47.0%	64.0%	49.6%
LLaVA-Video	7B	47.0%	59.2%	61.0%	68.7%	48.5%	50.0%	21.3%	47.3%	56.5%	68.7%	52.8%
VideoLLaMA 3	7B	47.3%	57.6%	68.0%	64.7%	50.0%	54.0%	21.3%	48.7%	55.0%	67.3%	53.5%
Qwen2.5-VL	7B	70.3%	70.0%	58.0%	68.7%	48.5%	66.4%	17.3%	49.3%	53.0%	71.3%	59.1%
InternVL3	8B	61.3%	72.0%	60.0%	69.3%	56.5%	62.4%	23.3%	57.3%	55.0%	71.3%	59.8%
Middle-scale MLLMs												
Oryx-1.5	32B	50.3%	60.0%	63.5%	62.7%	46.0%	60.4%	21.3%	54.7%	61.0%	68.0%	55.2%
Kimi-VL	16B	54.7%	66.4%	59.0%	62.7%	48.0%	57.6%	23.3%	56.7%	57.5%	71.3%	56.2%
Qwen2.5-VL	32B	58.3%	77.2%	62.5%	68.7%	52.0%	70.4%	22.7%	68.7%	54.5%	65.3%	61.0%
InternVL3	38B	67.0%	76.8%	65.0%	76.0%	61.0%	69.6%	24.7%	76.0%	61.5%	76.7%	66.1%
Large-scale MLLMs												
InternVL3	78B	70.0%	77.6%	67.5%	76.0%	65.5%	71.6%	24.7%	77.3%	57.0%	75.3%	67.2%
Qwen2.5-VL	72B	80.7%	80.0%	65.0%	74.0%	56.5%	79.6%	26.7%	74.7%	57.0%	78.7%	69.0%

Next, it is clear that models with larger parameter scales tend to achieve higher accuracy, with a clear scaling effect evident in the Qwen2.5-VL [30], InternVL3 [31], and Oryx-1.5 [60] series. Meanwhile, model architecture significantly impacts performance. Despite achieving high scores on general video understanding multiple-choice benchmarks (e.g., Video-MME [51], MLVU [63]), token compression methods show a clear disadvantage on MME-VideoOCR. Representative approaches such as VideoChat-Flash [47] and Slow-fast MLLM [62] illustrate this limitation, suggesting that critical information may be lost during the token merging process.

In addition, our benchmark presents strong discriminative power across task categories, which could be taken as the potential direction for MLLM optimization. For tasks such as Text Recognition, Visual Text QA, and Text-Based Reasoning, the performance gap between the best and worst-performing models exceeds 30%, clearly distinguishing model capabilities across different levels of perception, understanding, and reasoning.

Furthermore, the benchmark reveals several common defects of the mainstream MLLMs. In tasks such as Change Detection & Tracking and Text-Based Video Understanding, most models achieve an accuracy below 60%, indicating significant challenges in dynamic scene comprehension and temporal alignment. For Cross-Frame Text Understanding, which requires multi-frame integration and memory, the baseline models generally achieve an accuracy below 25%, underscoring their limited capacity for semantic integration across frames.

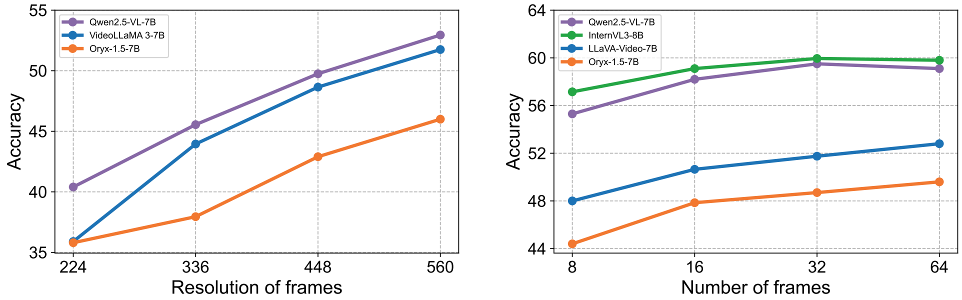
4.2 Analysis and Findings

Table 4 presents the accuracy of the top-5 performing models among the 18 evaluated MLLMs on each task. The full results for all models are provided in Appendix C.3.

Resolution and Number of Frames. To investigate the impact of resolution and frame count on models' performance in OCR tasks, we conduct two sets of comparative experiments. For the resolution study, we use Qwen2.5-VL [30], VideoLLaMA 3 [61], and Oryx-1.5 [60], all of which support dynamic resolution settings. In this experiment, the maximum number of input frames per sample is fixed at 32. Subsequently, the original video resolution is adjusted by scaling the longer edge of each frame to 224, 336, 448, or 560 pixels. As shown in Figure 5a, increasing the input resolution consistently leads to performance improvements across all models. To analyze the effect of input frame count, we select Qwen2.5-VL [30], InternVL3 [31], LLaVA-Video [29], and Oryx-1.5 [60], as these models are equipped with relatively long context windows. As illustrated in Figure 5b, increasing the number of input frames generally leads to a notable improvement in

Table 4: **Accuracy of top-5 performing evaluated MLLMs on each task**. Fine-grained task types offer an accurate reflection of the models' capabilities and limitations across multiple dimensions.

Task Category	**Task**	**Gemini 2.5-Pro**	**Qwen2.5-VL 72B**	**InternVL3 78B**	**GPT-4o**	**InternVL3 38B**
Text Recognition	Text Recognition at Designated Locations	86.0%	80.5%	72.5%	82.0%	70.0%
	Text Recognition Based on Specific Attributes	77.0%	81.0%	65.0%	86.0%	61.0%
Visual Text QA	Text-Centric QA	93.5%	83.5%	80.0%	84.5%	78.0%
	Translation	84.0%	66.0%	68.0%	70.0%	72.0%
Text Grounding	Spatial Grounding	88.0%	81.0%	77.0%	67.0%	83.0%
	Temporal Grounding	41.0%	49.0%	58.0%	54.0%	47.0%
Attribute Recognition	Color Recognition	76.0%	90.0%	90.0%	88.0%	88.0%
	Named Entity Recognition	84.0%	78.0%	74.0%	74.0%	76.0%
	Counting	62.0%	54.0%	64.0%	62.0%	64.0%
Change Detection & Tracking	Change Detection	57.0%	44.0%	55.0%	46.0%	48.0%
	Tracking	83.0%	69.0%	76.0%	57.0%	74.0%
Special Text Parsing	Table Parsing	92.0%	74.0%	56.0%	52.0%	60.0%
	Chart Parsing	84.0%	72.0%	68.0%	68.0%	66.0%
	Document Parsing	92.0%	94.0%	76.0%	90.0%	76.0%
	Mathematical Formula Parsing	68.0%	88.0%	90.0%	62.0%	80.0%
	Handwriting Recognition	86.0%	70.0%	68.0%	68.0%	66.0%
Cross-Frame Text Understanding	Scrolling Text Understanding	70.0%	64.0%	70.0%	62.0%	70.0%
	Trajectory Recognition	0.0%	0.0%	0.0%	0.0%	0.0%
	Scrambled Recognition	76.0%	16.0%	4.0%	30.0%	4.0%
Text-Based Reasoning	Complex Reasoning	74.0%	74.7%	77.3%	60.7%	76.0%
Text-Based Video Understanding	Subtitle-Based Video Understanding	86.0%	96.0%	96.0%	93.0%	95.0%
	Multi-Hop Needle in A Haystack	27.0%	18.0%	18.0%	25.0%	28.0%
Robust Video Testing	AIGC Videos	88.0%	88.0%	84.0%	82.0%	88.0%
	Long Videos	44.0%	58.0%	68.0%	60.0%	62.0%
	Adversarial Videos	84.0%	90.0%	74.0%	84.0%	80.0%
Total	-	73.7%	69.0%	67.2%	66.4%	66.1%



(a) Different resolution settings.

(b) Different frame sampling settings.

Figure 5: **Model performance on MME-VideoOCR under different resolution and frame sampling settings**. Both lower resolution and reduced frame count significantly degrade performance, underscoring the importance of visual coverage and clarity in OCR tasks.

model performance. However, we observe a slight performance drop for Qwen2.5-VL and InternVL3 when the number of input frames increases from 32 to 64. This suggests that when the context becomes excessively long, the models may struggle to focus on task-relevant content, potentially due to limitations in attention allocation or memory compression within long sequences. These findings highlight the importance of both high resolution and sufficient temporal coverage for OCR tasks.

Effective Utilization of Textual Information. In *Subtitle-Based Video Understanding*, most models achieve relatively strong performance. We investigate the task samples and reveal that the correct answers typically appear in a single frame or a small number of frames within the video. This suggests that leading MLLMs are capable of effectively utilizing textual information embedded in videos, and can combine it with visual context to perform accurate video understanding.

Limitations in Temporal Integration Capability. As shown in Table 3, all models exhibit clear shortcomings in Cross-Frame Text Understanding tasks, with most models achieving accuracies around 20%. Table 4 further breaks down the performance on individual tasks within this category. All of the top-5 performing models yield an accuracy of 0% on *Trajectory Recognition*, and 4 out of 5 achieve less than 35% accuracy on *Scrambled Recognition*. These results underscore a common deficiency in the temporal integration capability of current MLLMs. The large performance gap between the two subtasks under the Text-Based Video Understanding category further supports this observation. Both *Subtitle-Based Video Understanding* and *Multi-Hop Needle in A Haystack* require effective video understanding grounded in textual information. However, the key difference lies in the distribution of relevant content: in the former, useful information appears in just a few frames, whereas in the latter, it is scattered across multiple frames and requires the model to perform effective memory and integration. This contrast reveals a critical limitation in current MLLMs: rather than effectively aggregating information across multiple frames, most models appear to rely primarily on information from a small number of frames for video OCR.

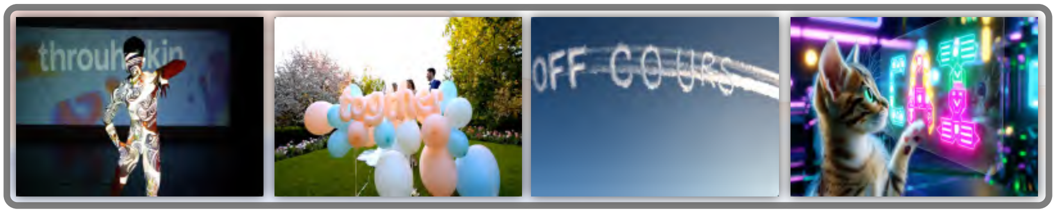


Figure 6: **Examples illustrating language prior bias in MLLMs**. The models tend to incorrectly recognize the text based on plausible language priors—for instance, "`throuh skin`" as "`through skin`", "`togther`" as "`together`", "`OFF COURS`" as "`OFF COURSE`", and "`CAI`" as "`CAT`". These cases highlight the strong influence of language priors on MLLM responses.

Significant Language Prior Bias. One notable failure mode in MLLMs is their tendency to over-rely on language priors when interpreting visually presented text. As illustrated in Figure 6, these models often convert visibly misspelled text into contextually plausible forms, even when the input is visually clear and unambiguous. This indicates that MLLMs frequently prioritize semantic likelihood over visual fidelity, generating interpretations that reflect linguistic expectations rather than the actual visual content. This bias poses a serious challenge for OCR-related tasks, where character-level accuracy is essential. Notably, the misrecognitions are not arbitrary; they follow consistent patterns that favor high-frequency or semantically familiar words over rare, misspelled, or out-of-vocabulary terms. Such behavior underscores the dominant role of language priors, which can override visual evidence—particularly when visual and textual signals are not sufficiently disentangled.

5 Conclusions, Discussions and Limitations

This paper introduces MME-VideoOCR, a benchmark designed for the comprehensive evaluation of video OCR capabilities. The benchmark comprises 25 practical OCR tasks, encompassing bilingual, perceptual, comprehension, and reasoning abilities. Experimental results demonstrate that MME-VideoOCR possesses sufficient difficulty and discrimination to expose the deficiencies of current MLLMs, thereby offering directions for the potential optimization.

While we endeavored to collect and construct videos from 9 diverse scenario categories with manually annotated, precise ground truth, the inherent richness of visual elements in videos means that some concepts may be underrepresented by samples. This may lead to score fluctuations in certain subcategories due to model sensitivity to sparse data. Although augmenting the dataset with more samples could mitigate this, we constrained the total number of items to $2,000$ due to considerations of annotation and evaluation costs. Furthermore, to assess fundamental abilities, we deliberately structured the questions into easy, medium, and hard difficulty tiers. Cutting-edge MLLMs have demonstrated strong performance on the easy and medium-difficulty questions. We intend to supplement the current version with more challenging samples as MLLM capabilities advance, ensuring its continued relevance in guiding future development.

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.

[2] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

[3] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024.

[4] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[5] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[6] Yang Shi, Jiaheng Liu, Yushuo Guan, Zhenhua Wu, Yuanxing Zhang, Zihao Wang, Weihong Lin, Jingyun Hua, Zekun Wang, Xinlong Chen, et al. Mavors: Multi-granularity video representation for multimodal large language model. *arXiv preprint arXiv:2504.10068*, 2025.

[7] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024.

[8] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*, 2024.

[9] Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226, 2023.

[10] Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, et al. Embodiedeval: Evaluate multimodal llms as embodied agents. *arXiv preprint arXiv:2501.11858*, 2025.

[11] Gabriel Sarch, Lawrence Jang, Michael Tarr, William W Cohen, Kenneth Marino, and Katerina Fragkiadaki. Vlm agents generate their own memories: Distilling experience into embodied programs of thought. *Advances in Neural Information Processing Systems*, 37:75942–75985, 2025.

[12] Fangtong Sun, Junjie Zhu, Zunlin Fan, Yiying Li, Zhiyuan Wang, and Ke Yang. Mmtp: Meta-learning-based multi-textual prompt tuning for visual-language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[13] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025.

[14] Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023.

[15] Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. *Optical character recognition*. John Wiley & Sons, Inc., 1999.

[16] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. 2024.

[17] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.

[18] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[19] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[20] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

[21] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[22] Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024.

[23] Sankalp Nagaonkar, Augustya Sharma, Ashish Choithani, and Ashutosh Trivedi. Benchmarking vision-language models on optical character recognition in dynamic video environments. *arXiv preprint arXiv:2502.06445*, 2025.

[24] Yulin Fei, Yuhui Gao, Xingyuan Xian, Xiaojin Zhang, Tao Wu, and Wei Chen. Do current video llms have strong ocr abilities? a preliminary study. *arXiv preprint arXiv:2412.20613*, 2024.

[25] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

[26] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640, 2024.

[27] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. *arXiv preprint arXiv:2502.07411*, 2025.

[28] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International conference on computer vision*, pages 1457–1464. IEEE, 2011.

[29] 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.

[30] 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.

[31] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[32] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

[33] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024.

[34] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2024.

[35] 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.

[36] Minyi Zhao, Bingjia Li, Jie Wang, Wanqing Li, Wenjing Zhou, Lan Zhang, Shijie Xuyang, Zhihang Yu, Xinkun Yu, Guangze Li, et al. Towards video text visual question answering: Benchmark and baseline. *Advances in Neural Information Processing Systems*, 35:35549–35562, 2022.

[37] George Tom, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and CV Jawahar. Reading between the lanes: Text videoqa on the road. In *International Conference on Document Analysis and Recognition*, pages 137–154. Springer, 2023.

[38] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25963–25973, 2024.

[39] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

[40] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019.

[41] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

[42] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[43] Weijia Wu, Yuanqiang Cai, Debing Zhang, Sibo Wang, Zhuang Li, Jiahong Li, Yejun Tang, and Hong Zhou. A bilingual, openworld video text dataset and end-to-end video text spotter with transformer. *arXiv preprint arXiv:2112.04888*, 2021.

[44] Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Watching the news: Towards videoqa models that can read. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4441–4450, 2023.

[45] Zhanzhan Cheng, Jing Lu, Yi Niu, Shiliang Pu, Fei Wu, and Shuigeng Zhou. You only recognize once: Towards fast video text spotting. In *Proceedings of the 27th ACM international conference on multimedia*, pages 855–863, 2019.

[46] Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Roadtext-1k: Text detection & recognition dataset for driving videos. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11074–11080. IEEE, 2020.

[47] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024.

[48] Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Mike Zheng Shou, Umapada Pal, Dimosthenis Karatzas, and Xiang Bai. Icdar 2023 video text reading competition for dense and small text. *arXiv preprint arXiv:2304.04376*, 2023.

[49] PaddlePaddle. Paddleocr. https://github.com/PaddlePaddle/PaddleOCR, 2020. Accessed: 2025-05-09.

[50] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[51] Chaoyou Fu, Yuhan Dai, Yongdong 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.

[52] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

[53] Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Debiasing multimodal large language models. *arXiv preprint arXiv:2403.05262*, 2024.

[54] 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? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[55] Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024.

[56] 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.

[57] 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.

[58] 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.

[59] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[60] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024.

[61] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[62] Min Shi, Shihao Wang, Chieh-Yun Chen, Jitesh Jain, Kai Wang, Junjun Xiong, Guilin Liu, Zhiding Yu, and Humphrey Shi. Slow-fast architecture for video multi-modal large language models. *arXiv preprint arXiv:2504.01328*, 2025.

[63] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[64] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section. 4

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No Theorem and Lemma.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: https://huggingface.co/datasets/DogNeverSleep/MME-VideoOCR_Dataset

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer:[NA]

   Justification: All evaluations were performed using greedy decoding with temperature=0, resulting in minimal variance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section D.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: all the original papers that produced the code package or dataset are cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

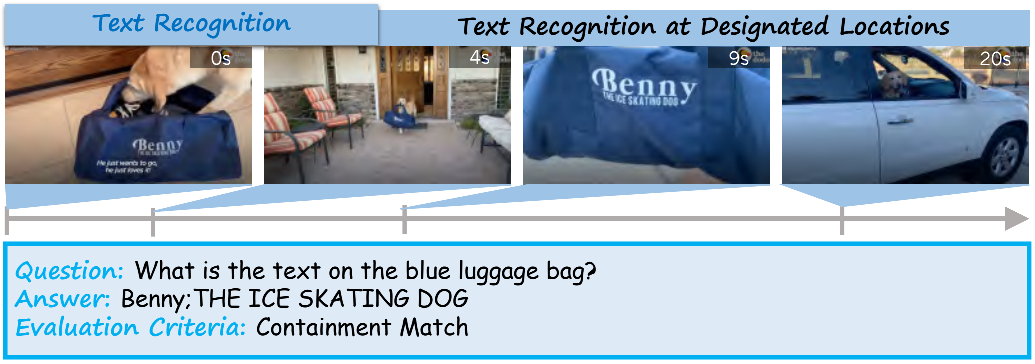
Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

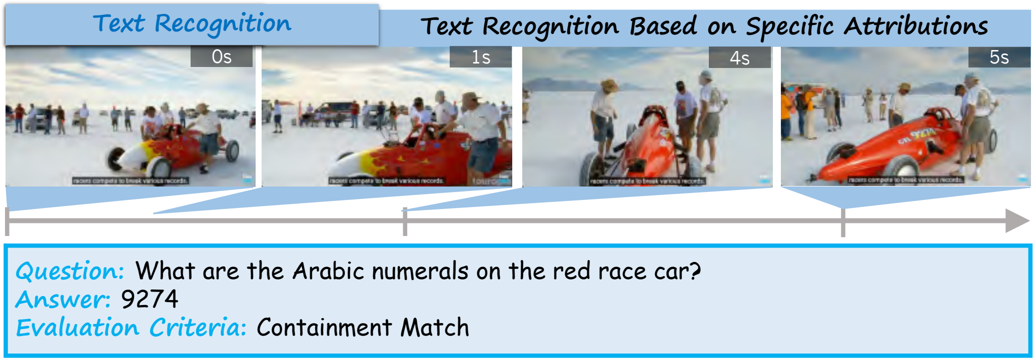
A Representative Examples from MME-VideoOCR

To comprehensively illustrate the characteristics of tasks in MME-VideoOCR, we present one representative example for each task.



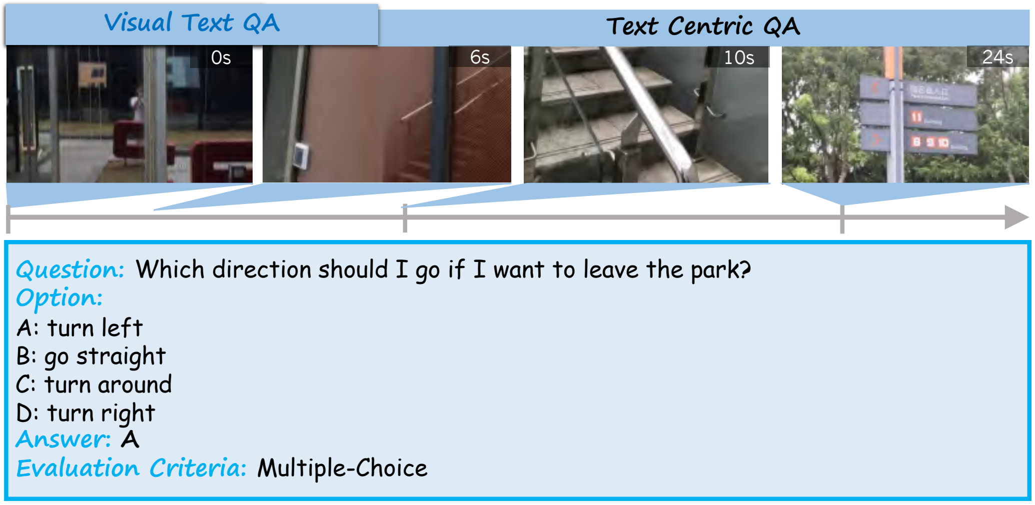
Question: What is the text on the blue luggage bag?
Answer: Benny;THE ICE SKATING DOG
Evaluation Criteria: Containment Match

Figure 7: An example QA of the Text Recognition at Designated Locations task in MME-VideoOCR.



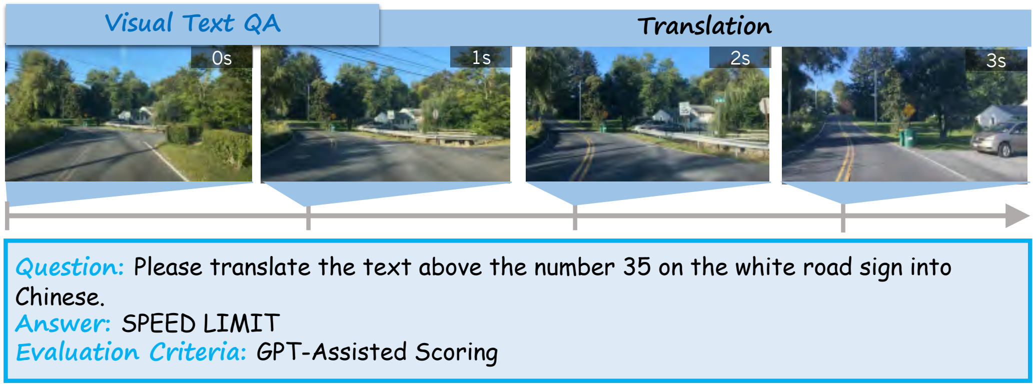
Question: What are the Arabic numerals on the red race car?
Answer: 9274
Evaluation Criteria: Containment Match

Figure 8: An example QA of the Text Recognition Based on Specific Attributes task in MME-VideoOCR.



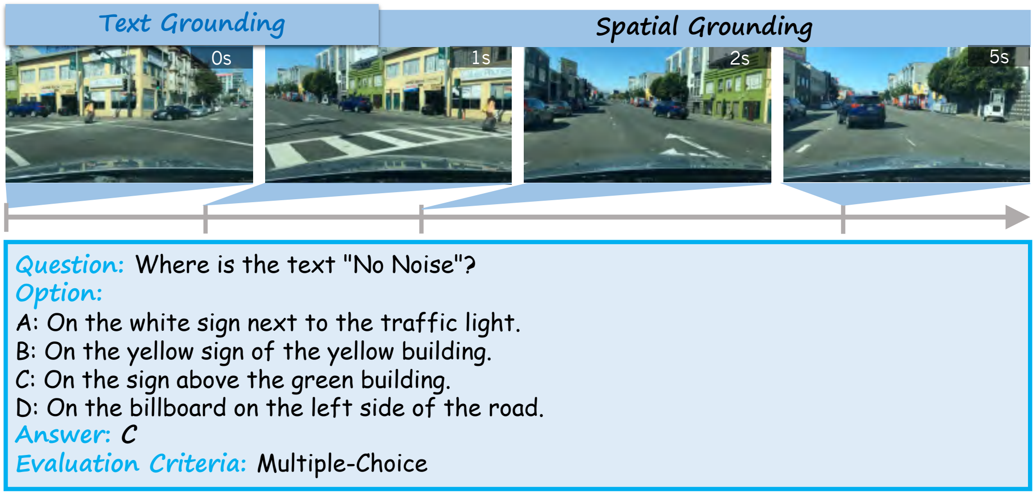
Question: Which direction should I go if I want to leave the park?
Option:
A: turn left
B: go straight
C: turn around
D: turn right
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 9: An example QA of the Text-Centric QA task in MME-VideoOCR.



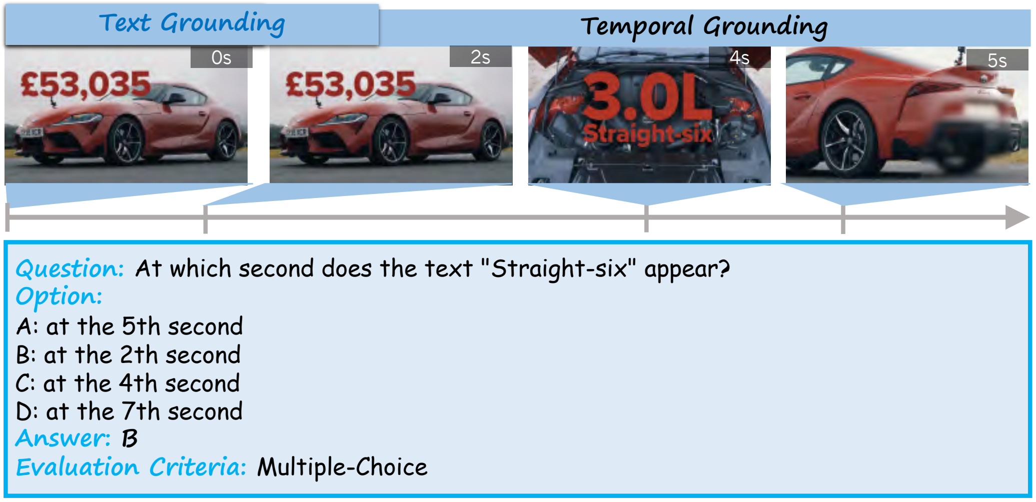
Question: Please translate the text above the number 35 on the white road sign into Chinese.
Answer: SPEED LIMIT
Evaluation Criteria: GPT-Assisted Scoring

Figure 10: An example QA of the Translation task in MME-VideoOCR.



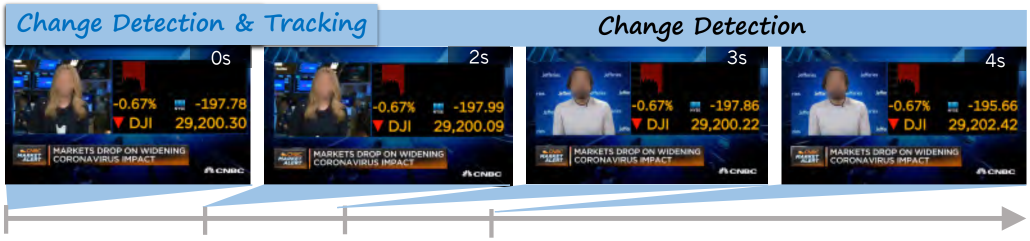
Question: Where is the text "No Noise"?
Option:
A: On the white sign next to the traffic light.
B: On the yellow sign of the yellow building.
C: On the sign above the green building.
D: On the billboard on the left side of the road.
Answer: C
Evaluation Criteria: Multiple-Choice

Figure 11: An example QA of the Spatial Grounding task in MME-VideoOCR.



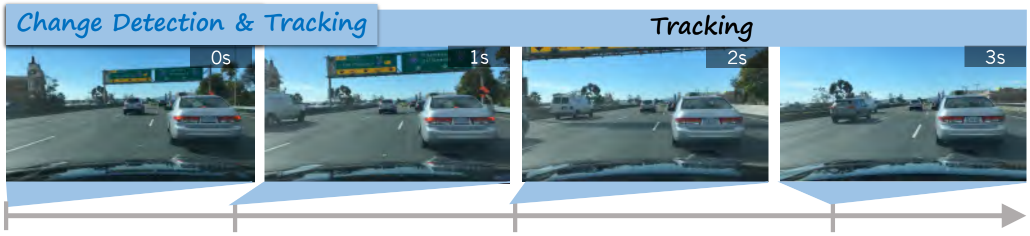
Question: At which second does the text "Straight-six" appear?
Option:
A: at the 5th second
B: at the 2th second
C: at the 4th second
D: at the 7th second
Answer: B
Evaluation Criteria: Multiple-Choice

Figure 12: An example QA of the Temporal Grounding task in MME-VideoOCR.



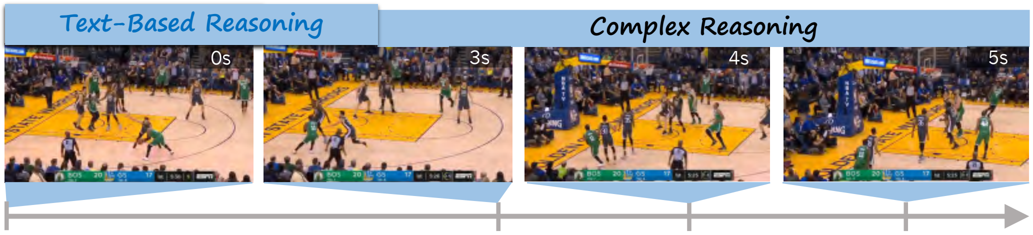
Question: How many times did the value to the right of DJI change?
Option:
A: 5
B: 3
C: 4
D: 2
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 13: An example QA of the Change Detection task in MME-VideoOCR.



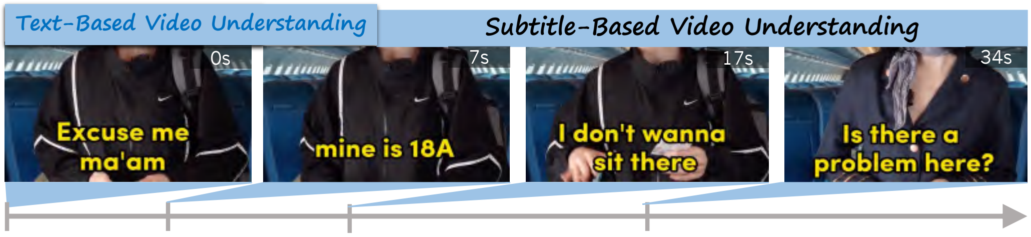
Question: How is the vehicle with the license plate number 5JVU366 moving?
Option:
A: Keeps going straight.
B: Moves into the left lane.
C: Moves into the right lane.
D: It stopped.
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 14: An example QA of the Tracking task in MME-VideoOCR.



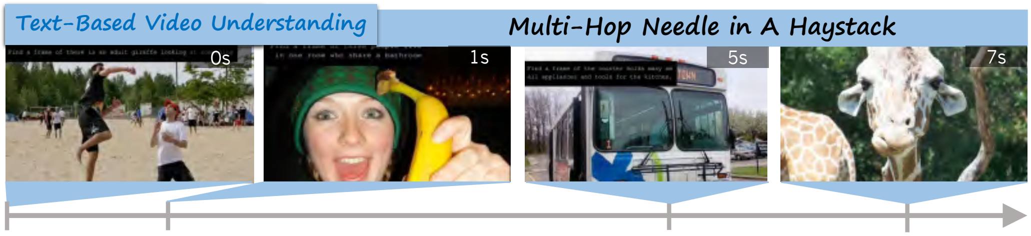
Question: How many points ahead will they be after making this shot?
Option:
A: 5
B: 4
C: 3
D: 6
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 15: An example QA of the Complex Reasoning task in MME-VideoOCR.



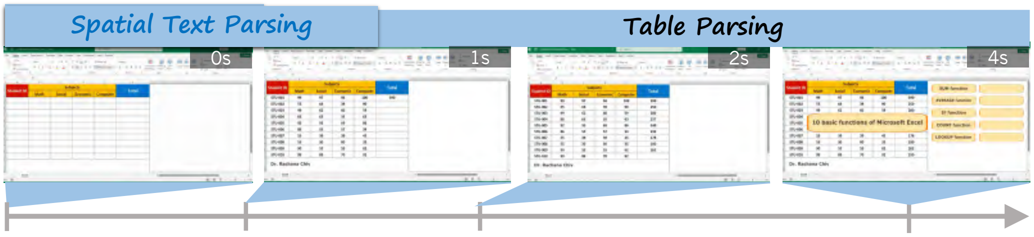
Question: What are the two passengers arguing about?
Option:
A: One passenger is unwilling to exchange seats with another passenger.
B: One passenger is unwilling to give up his seat to another passenger.
C: One passenger is unwilling to let another passenger sit next to him.
D: One passenger is unwilling to let another passenger sit in the aisle.
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 16: An example QA of the Subtitle-Based Video Understanding task in MME-VideoOCR.


Text-Based Video Understanding
Multi-Hop Needle in A Haystack
0s
1s
5s
7s

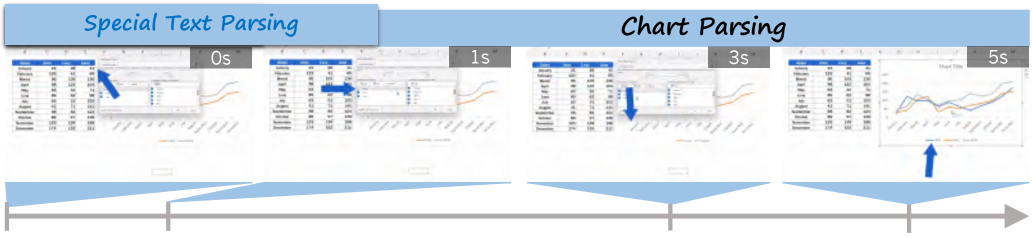
Question: Find the frame of a girl shows a banana to the camera. Locate the final frame based on the instructions in the image, and answer the following question based on the final frame: Which of the following objects or actions appears in the frame?
Option:
A: banana
B: giraffe
C: flower
D: bus
Answer: B
Evaluation Criteria: Multiple-Choice

Figure 17: An example QA of the Multi-Hop Needle in A Haystack task in MME-VideoOCR.



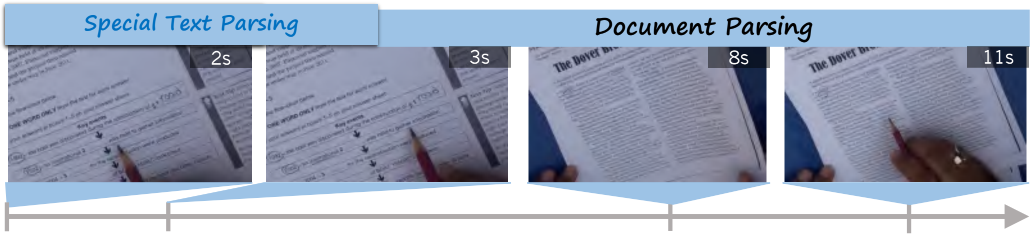
Question: How much bigger is the number for STU-006-Total than for STU-004-Total?
Option:
A: 5
B: 232
C: 227
D: Equal
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 18: An example QA of the Table Parsing task in MME-VideoOCR.



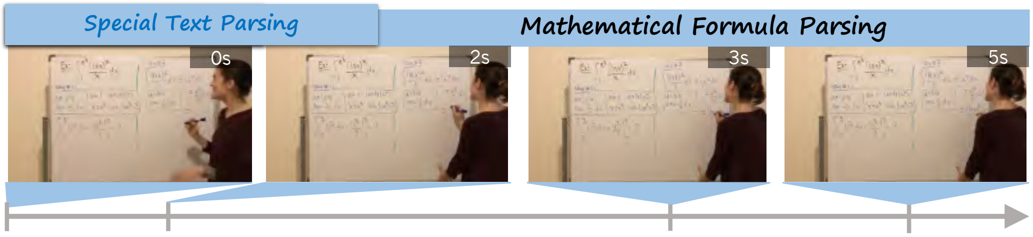
Question: What is the sales figure for March represented by the blue line in the line graph?
Option:
A: 136
B: 96
C: 105
D: 104
Answer: B
Evaluation Criteria: Multiple-Choice

Figure 19: An example QA of the Chart Parsing task in MME-VideoOCR.



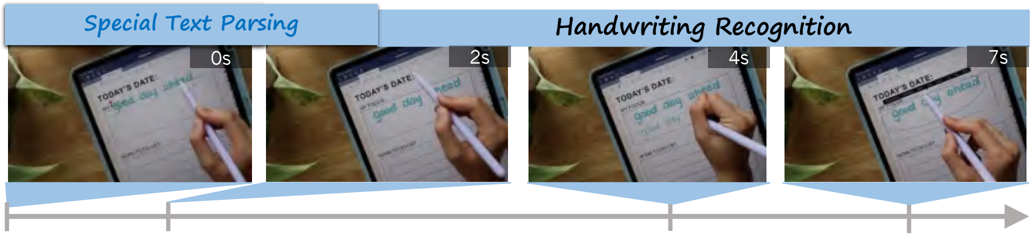
Question: According to the newspaper, which city does the road built in 1992 pass through?
Option:
A: Bradford
B: Canterbury
C: Dover
D: London
Answer: C
Evaluation Criteria: Multiple-Choice

Figure 20: An example QA of the Document Parsing task in MME-VideoOCR.



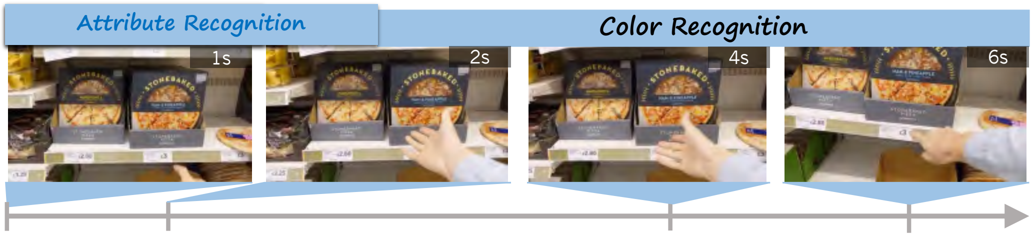
Question: What should be the denominator of the equation being written in the video?
Option:
A: 3
B: 3x
C: u^3
D: C
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 21: An example QA of the Mathematical Formula Parsing task in MME-VideoOCR.



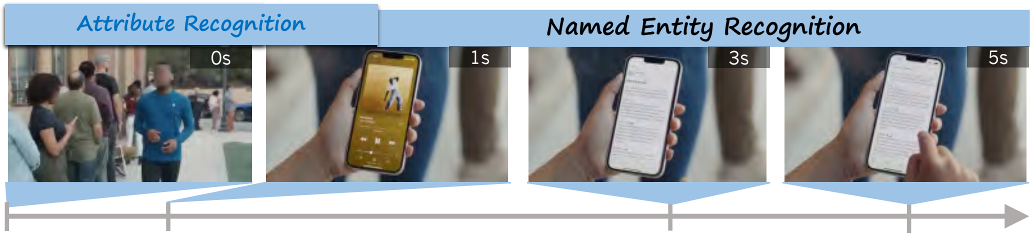
Question: What was written and then erased?
Answer: good day ahead
Evaluation Criteria: Containment Match

Figure 22: An example QA of the Handwriting Recognition task in MME-VideoOCR.



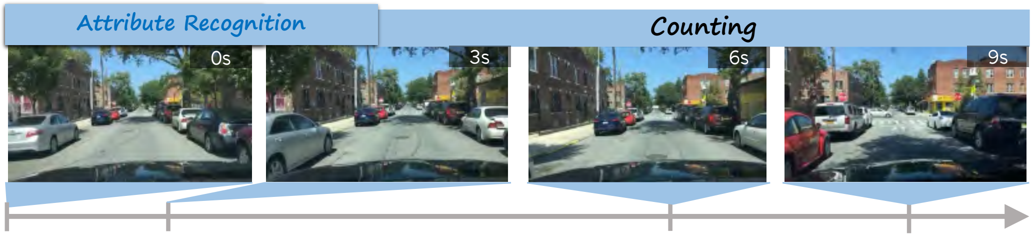
Question: What is the color of the text "HAM & PINEAPPLE"?
Option:
A: red
B: yellow
C: gray
D: blue
Answer: D
Evaluation Criteria: Multiple-Choice

Figure 23: An example QA of the Color Recognition task in MME-VideoOCR.



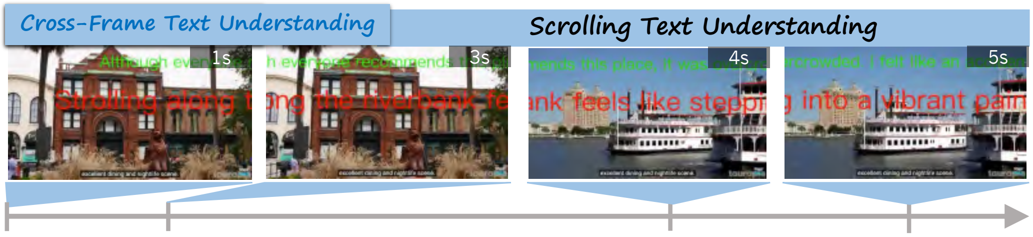
Question: How many Italian cities appeared in the Fall Trip plan shown in the video?
Option:
A: 2
B: 4
C: 3
D: 1
Answer: B
Evaluation Criteria: Multiple-Choice

Figure 24: An example QA of the Named Entity Recognition task in MME-VideoOCR.



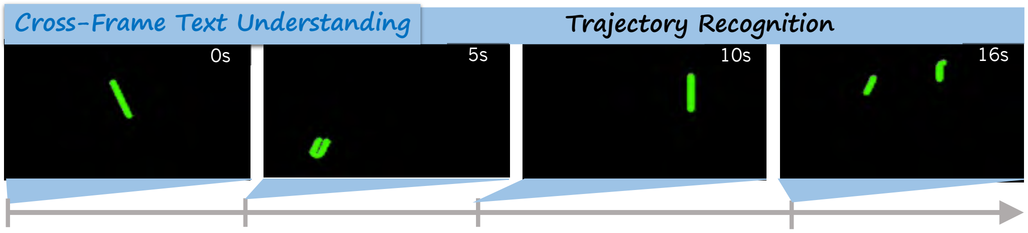
Question: How many stop signs with the word "STOP" can be seen in the video?
Option:
A: 2
B: 1
C: 3
D: 4
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 25: An example QA of the Counting task in MME-VideoOCR.



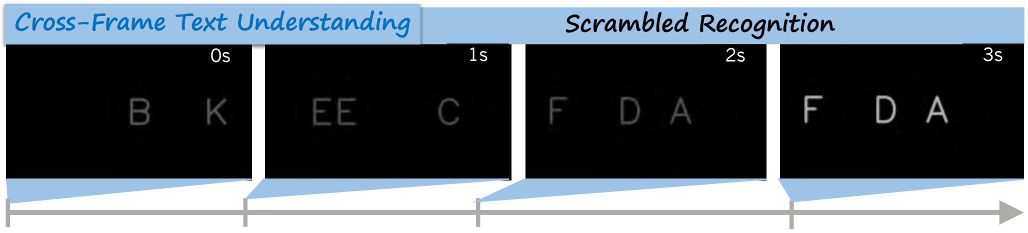
Question: What is the attitude of the green barrage towards this scenic spot?
Option:
A: excited
B: disgust
C: satisfied
D: interested
Answer: B
Evaluation Criteria: Multiple-Choice

Figure 26: An example QA of the Scrolling Text Understanding task in MME-VideoOCR.



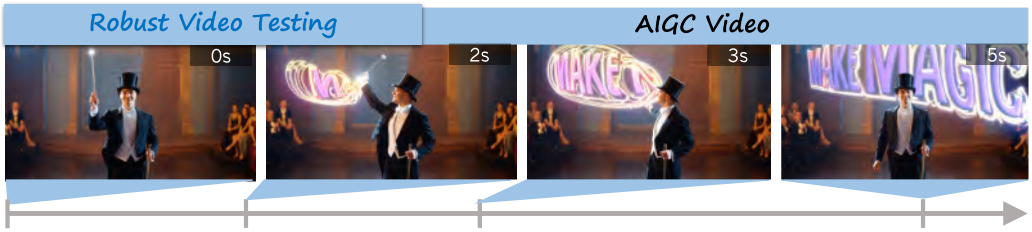
Question: What letter or combination of letters is formed by all the trajectories collectively?
Answer: AI
Evaluation Criteria: Containment Match

Figure 27: An example QA of the Trajectory Recognition task in MME-VideoOCR.



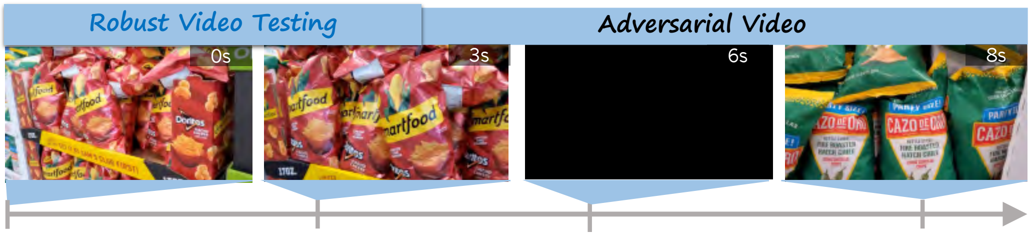
Question: What is formed by arranging these randomly appearing elements by position?
Answer: FEEDBACK
Evaluation Criteria: Containment Match

Figure 28: An example QA of the Scrambled Recognition task in MME-VideoOCR.



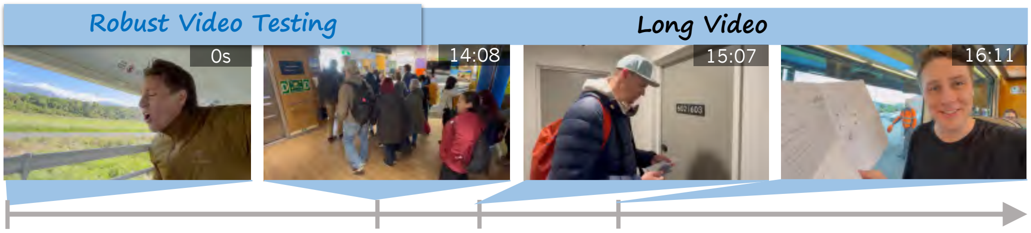
Question: What is the text revealed by the magician?
Answer: MAKE MAGIC
Evaluation Criteria: Containment Match

Figure 29: An example QA of the AIGC Video task in MME-VideoOCR.



Question: How much do the two packs of green snacks weigh?
Option:
A: 1246g
B: 6g
C: 320g
D: 623g
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 30: An example QA of the Adversarial Video task in MME-VideoOCR.



Question: Which room did they stay in the hotel?
Option:
A: 603
B: 602/603
C: 601
D: 602
Answer: A
Evaluation Criteria: Multiple-Choice

Figure 31: An example QA of the Long Video task in MME-VideoOCR.

B Benchmark Details

B.1 Task Definition

MME-VideoOCR collects 10 OCR task categories. Detailed definition of the taxonomy is depicted as below.

Text Recognition. Text Recognition is a fundamental OCR task that evaluates an MLLM's ability to perceive and interpret text. This category involves *Text Recognition at Designated Locations* and *Text Recognition Based on Specific Attributes*. These two subtasks can be flexibly combined to assess an MLLM's capacity for fine-grained text recognition. For instance, a query may require recognizing text specifically located on a license plate and written in a particular language or color, thereby evaluating both spatial awareness and attribute-based recognition within complex visual scenes.

Visual Text QA. Visual Text QA encompasses two tasks: *Text-Centric QA* and *Translation*. *Text-Centric QA* requires models to integrate textual content with relevant visual cues to answer context-dependent questions. *Translation* focuses on converting specific on-screen text into a designated target language. Both tasks challenge the model's ability to not only perceive but also comprehend multimodal information.

Text Grounding. Text Grounding involves *Spatial Grounding* and *Temporal Grounding*. *Spatial Grounding* concerns identifying the location of specified text based on visual context—such as recognizing that the text appears on a street sign or a product label—rather than relying on exact coordinates. *Temporal Grounding* centers on understanding the temporal properties of text, including when it appears, how long it remains visible, and the sequence in which it occurs. Together, these subtasks assess the model's ability to localize and interpret text across both spatial and temporal dimensions within dynamic visual scenes.

Attribute Recognition. This category is composed of three tasks: *Color Recognition*, where models are expected to identify the color of the text; *Named Entity Recognition*, which focuses on extracting and classifying named entities such as person names, organization names, and location names; and *Counting*, where models must accurately identify the number of textual elements that meet specified criteria.

Change Detection & Tracking. The task consists of two tasks: *Change Detection* and *Tracking*. Given the highly dynamic nature of text in video, *Change Detection* aims to accurately identify changes in textual content over time. *Tracking*, on the other hand, focuses on monitoring text elements as they change position across frames—for example, tracing the movement of a vehicle with a specified license plate number or identifying the player running with the ball based on their jersey number.

Special Text Parsing. Special Text Parsing includes five tasks: *Table Parsing*, *Chart Parsing*, *Document Parsing*, *Mathematical Formula Parsing*, and *Handwriting Recognition*. These tasks require models to accurately identify and understand text with either special structures or highly variable visual forms.

Cross-Frame Text Understanding. In video scenarios, relying on a single frame is often insufficient, as critical information may be distributed across multiple frames and closely interrelated. To address this, the task of Cross-Frame Text Understanding is introduced, which requires models to integrate information across multiple frames for coherent interpretation. It includes three subtasks: *Scrolling Text Understanding*, which focuses on recognizing dynamic text streams—such as on-screen bullet comments—that move across frames and may only be fully readable when aggregated over time; *Trajectory Recognition*, where the motion path of an object in the video forms a recognizable text, and the model must interpret this trajectory as the intended message; *Scrambled Recognition*, which involves identifying and reconstructing a complete text from characters that appear out of order across different positions in the video frames.

Text-Based Reasoning. Text-Based Reasoning, also referred to as *Complex Reasoning*, emphasizes advanced understanding of textual content, such as code analysis, mathematical operations, and logical reasoning. Unlike *Text-Centric QA*, which is a straightforward comprehension task centered on identifying explicit information, *Complex Reasoning* requires models to go beyond surface-level understanding by synthesizing dispersed textual cues, identifying implicit relationships, and resolving ambiguity or misleading content.

Text-Based Video Understanding. Current video understanding tasks are primarily based on visual dynamics, such as action recognition and video captioning. However, these tasks often overlook the textual information in videos, even though they are essential for video understanding in certain contexts. To address this gap, we introduce *Subtitle-Based Video Understanding*. In this task, the answer to a question is hidden in the subtitles, and MLLMs must combine subtitle information with visual content to answer correctly. This reflects real-world scenarios like conversations, tutorials, or news, where subtitles provide key information that visuals alone cannot capture. *Multi-Hop Needle in A Haystack* is a novel and effective task introduced in VideoChat-Flash [47], designed to test models' ability to retrieve information from videos based on subtitles that are spread across multiple frames. It requires reasoning over multiple pieces of subtitle content to find the correct answer.

Robust Video Testing. To evaluate model effectiveness and robustness across diverse scenarios, we introduce three specialized video types: *AIGC Videos*, *Long Videos*, and *Adversarial Videos*. *AIGC Videos*, generated by AI systems [42], assess model adaptability to increasingly common synthetic content. *Long Videos* test the ability to extract relevant information from lengthy sequences with substantial redundancy. Since existing MLLMs primarily process videos by extracting frames, we construct a set of *Adversarial Videos* by strategically inserting all-black frames into normal videos. While these adversarial samples have minimal impact on human perception, they can easily mislead the model, rendering it virtually "blind".

B.2 Task Distribution

Table 5: **Number of QA Pairs per task in MME-VideoOCR**.

Task Category	**Task**	**#QA**
Text Recognition	Text Recognition at Designated Locations	200
	Text Recognition Based on Specific Attributes	100
Visual Text QA	Text-Centric QA	250
	Translation	50
Text Grounding	Spatial Grounding	100
	Temporal Grounding	100
Attribute Recognition	Color Recognition	50
	Named Entity Recognition	50
	Counting	50
Change Detection & Tracking	Change Detection	100
	Tracking	100
Special Text Parsing	Table Parsing	50
	Chart Parsing	50
	Document Parsing	50
	Mathematical Formula Parsing	50
	Handwriting Recognition	50
Cross-Frame Text Understanding	Scrolling Text Understanding	50
	Trajectory Recognition	50
	Scrambled Recognition	50
Text-Based Reasoning	Complex Reasoning	150
Text-Based Video Understanding	Subtitle-Based Video Understanding	100
	Multi-Hop Needle in a Haystack	100
Robust Video Testing	AIGC Videos	50
	Long Videos	50
	Adversarial Videos	50
Total	-	2,000

Given the diverse range of task types included in MME-VideoOCR, which assess a broad spectrum of model capabilities, we carefully allocate the number of QA pairs across different tasks. Table 5

Table 6: **Evaluation prompt setting of MME-VideoOCR (Containment Match)**.

`[Video]` Based on the video and the question below, directly answer the content that needs to be recognized in plain text. Do not include any additional explanations, formatting changes, or extra information. Question: `[Question]` The answer is:

Table 7: **Evaluation prompt setting of MME-VideoOCR (GPT-Assisted Scoring)**.

`[Video]` Based on the video and the question below, directly provide the answer in plain text. Do not include any additional explanations, formatting changes, or extra information. Question: `[Question]` The answer is:

presents the specific number of QA pairs for each task. This allocation ensures a balanced distribution among perception, understanding, and reasoning tasks, thereby supporting a comprehensive and equitable evaluation of model capabilities.

B.3 Evaluation Prompt

The prompt settings for Containment Match, GPT-Assisted Scoring and Multiple-Choice are shown in Table 6, Table 7 and Table 8. For GPT-Assisted Scoring (designed for the Translation task), after obtaining the model's response using the prompt shown in Table 7, we subsequently utilize GPT-4o-0806 to evaluate the response. The corresponding evaluation prompt is provided in Table 9.

C Experiment Details

C.1 Evaluated Models

We evaluate a total of 18 mainstream MLLMs, including 3 leading proprietary models and 15 high-performing open-source models.

For proprietary models, we evaluate GPT-4o [56], Gemini-2.5 Pro [57] and Gemini-1.5 Pro [5].

- *GPT-4o* is the latest multimodal large language model developed by OpenAI, offering fast and cost-effective performance across text, image, and audio modalities. It achieves state-of-the-art results on a variety of benchmarks, with notable improvements in visual reasoning, OCR, and multilingual understanding. GPT-4o features a unified architecture that enables seamless cross-modal interaction, making it highly efficient and versatile for real-world multimodal applications.
- *Gemini-2.5 Pro* is one of the latest Multimodal Large Language Models released by Google DeepMind. It features improved visual and video understanding capabilities, with support for extended context lengths and more efficient cross-modal alignment. Gemini-2.5 Pro demonstrates strong performance across a wide range of tasks, including video captioning, image reasoning, and OCR-based understanding. Its enhanced architecture and training scale make it particularly competitive in complex multimodal benchmarks.
- *Gemini-1.5 Pro*, an earlier version in the Gemini series, also supports multimodal input and is optimized for high-quality text generation and basic vision-language tasks. While it delivers reliable performance on standard image-based benchmarks, its video comprehension ability—especially in tasks requiring temporal reasoning and dense visual-textual alignment—is more limited compared to its successor. Nevertheless, it remains a strong baseline among proprietary models.

For open-source models, we select Qwen2.5-VL [30], LLaVA-Video [29], LLaVA-OneVision [58], VideoLLaMA 3 [61], VideoChat-Flash [47], Oryx-1.5 [60], Slowfast-MLLM [62], InternVL3 [64], VITA-1.5 [2] and Kimi-VL [59]. Among them, for Oryx-1.5, Qwen2.5-VL, and InternVL3, we include versions with different parameter scales in our experiments.

Table 8: **Evaluation prompt setting of MME-VideoOCR (Multiple-Choice)**.

[Video]
Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option.
Question: [Question]
Option:
A. [Option A]
B. [Option B]
C. [Option C]
D. [Option D]
The best answer is:

Table 9: **Evaluation prompt setting of the Translation task**.

You are a professional bilingual translation evaluator.

Here are two sentences: one in Chinese and one in English.
Sentence 1: [Ground Truth]
Sentence 2: [MLLM's Response]

Please evaluate whether the two sentences convey the same meaning and can be considered accurate translations of each other.

If the meanings are equivalent and the translation is accurate, respond with "correct".
If there are significant differences in meaning or inaccuracies in translation, respond with "wrong".

You must only respond with one word: "correct" or "wrong". Do not provide any explanations, comments, or additional text.
Focus solely on semantic equivalence, not grammar or style. Ignore minor differences as long as the meaning is preserved.

- *Qwen2.5-VL* is a vision-language model that introduces two key innovations: native dynamic-resolution processing and Multi-scale Rotary Position Embedding (MRoPE). The dynamic-resolution capability allows the model to process images and videos of varying resolutions and frame rates efficiently, extending to the temporal dimension through dynamic FPS sampling. This enables precise temporal event localization in long videos. MRoPE enhances the model's ability to capture multi-scale positional information, improving its performance in tasks requiring fine-grained spatial and temporal understanding .
- *LLaVA-Video* extends the LLaVA framework to video understanding by unifying visual representations into the language feature space. This alignment before projection enables the model to perform visual reasoning on both images and videos simultaneously. By training on a mixed dataset of images and videos, LLaVA-Video leverages mutual enhancement between modalities, achieving superior performance across various visual-language tasks .
- *LLaVA-OneVision* is designed for versatile visual task transfer across single-image, multi-image, and video scenarios. It employs a SigLIP vision encoder and a Qwen2 language backbone, processing images with the Anyres technique to handle high-resolution inputs effectively. Videos are processed with a fixed token length per frame for memory efficiency. This architecture enables LLaVA-OneVision to excel in diverse visual-language tasks without task-specific fine-tuning.
- *VideoLLaMA 3* is a vision-centric multimodal foundation model that advances image and video understanding. It utilizes Any-resolution Vision Tokenization (AVT) to process images and videos of varying resolutions dynamically. The model's training paradigm emphasizes high-quality image-text data to enhance video understanding capabilities. VideoLLaMA 3 achieves state-of-the-art performance on multiple benchmarks by integrating vision-centric training and framework designs.
- *VideoChat-Flash* is a long-context video-language model that introduces a Hierarchical visual token Compression (HiCo) method, effectively reducing redundancy in long videos by compressing visual tokens from the clip-level to the video-level. This approach enables high-fidelity representation while significantly lowering computational costs. Coupled with

a multi-stage short-to-long learning scheme and training on the LongVid dataset, VideoChat-Flash achieves state-of-the-art performance on both long and short video benchmarks.

- *Oryx-1.5* presents a unified multimodal architecture designed for on-demand spatial-temporal understanding of images, videos, and multi-view 3D scenes. It features a dynamic compressor module that performs token compression and adaptive positional embedding, allowing the model to efficiently process visual inputs with arbitrary spatial sizes and temporal lengths. This flexibility enables Oryx-1.5 to seamlessly handle diverse visual inputs across various modalities.
- *Slowfast-MLLM* integrates the SlowFast dual-pathway architecture with a multimodal large language model to explicitly capture both coarse and fine-grained temporal dynamics. The slow branch models long-term context, while the fast branch focuses on short-term changes, enabling rich motion representation. This design enhances temporal alignment and supports detailed video-text interaction in tasks such as action question answering and event tracking.
- *InternVL3* is a powerful vision-language model that unifies visual grounding, dense captioning, and temporal understanding via a cross-modality fusion backbone. It introduces region-level supervision and multi-frame alignment strategies, significantly improving its spatial-temporal grounding capabilities. InternVL3 demonstrates superior performance across a wide range of multimodal tasks, benefiting from its native multimodal pre-training paradigm and advanced post-training techniques.
- *VITA-1.5* is a multimodal large language model designed to achieve real-time vision and speech interaction. It pioneers a meticulously crafted three-stage training strategy to effectively integrate vision, language, and speech modalities. This strategy systematically introduces visual and auditory data, mitigating conflicts between modalities while preserving robust multimodal capabilities. This methodology empowers VITA-1.5 to process and understand both visual and speech inputs and to generate fluent, end-to-end speech outputs, thereby enabling more natural and seamless interactive multimodal conversations.
- *Kimi-VL* is a state-of-the-art vision-language model developed by Moonshot AI, based on the Kimi series of large language models. Designed to handle complex multimodal tasks, Kimi-VL integrates high-resolution visual encoders with large-scale language understanding to enable robust performance in image captioning, visual question answering, and document understanding. It adopts a Mixture-of-Experts (MoE) architecture to improve inference efficiency, dynamically activating a subset of experts for each input. This design allows Kimi-VL to scale effectively while maintaining strong generalization across diverse visual-language benchmarks.

C.2 Experimental Setup

For proprietary models, we used the `gpt-4o-2024-08-06`, `gemini-2.5-pro-preview-05-06` and `gemini-1.5-pro-002` APIs, respectively.

In the MME-VideoOCR evaluation, most models were configured with a maximum input frame count of 64. GPT-4o was limited to 50 input frames due to API token constraints, while VITA-1.5 was restricted to 16 frames because of context length limitations. All other settings followed default or recommended configurations.
During the comparative experiments described in Section 4.2, the number of input frames was fixed at 32 when varying the resolution, while the default resolution settings were applied to all models when varying the number of input frames.

C.3 Experiment Results

Table 10, Table 11 and Table 12 present the complete results of evaluated models across all tasks in MME-VideoOCR.

D Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Table 10: **Accuracy of evaluated MLLMs on each task of MME-VideoOCR**.

Task Category	**Task**	**Gemini 1.5 Pro**	**Qwen2.5-VL 32B**	**InternVL 8B**	**Qwen2.5-VL 7B**	**Kimi-VL**
Text Recognition	Text Recognition at Designated Locations	80.0%	55.0%	64.0%	70.0%	54.5%
	Text Recognition Based on Specific Attributes	70.0%	65.0%	56.0%	71.0%	55.0%
Visual Text QA	Text-Centric QA	83.0%	81.5%	75.5%	76.0%	68.5%
	Translation	56.0%	60.0%	58.0%	46.0%	58.0%
Text Grounding	Spatial Grounding	78.0%	73.0%	77.0%	77.0%	71.0%
	Temporal Grounding	45.0%	52.0%	43.0%	39.0%	47.0%
Attribute Recognition	Color Recognition	62.0%	78.0%	80.0%	78.0%	70.0%
	Named Entity Recognition	80.0%	78.0%	72.0%	76.0%	70.0%
	Counting	52.0%	50.0%	56.0%	52.0%	48.0%
Change Detection & Tracking	Change Detection	43.0%	40.0%	49.0%	40.0%	33.0%
	Tracking	67.0%	64.0%	64.0%	57.0%	63.0%
Special Text Parsing	Table Parsing	72.0%	66.0%	56.0%	58.0%	54.0%
	Chart Parsing	74.0%	60.0%	60.0%	68.0%	48.0%
	Document Parsing	80.0%	90.0%	72.0%	86.0%	74.0%
	Mathematical Formula Parsing	76.0%	76.0%	64.0%	60.0%	60.0%
	Handwriting Recognition	68.0%	60.0%	60.0%	60.0%	52.0%
Cross-Frame Text Understanding	Scrolling Text Understanding	72.0%	52.0%	70.0%	48.0%	70.0%
	Trajectory Recognition	0.0%	0.0%	0.0%	0.0%	0.0%
	Scrambled Recognition	22.0%	16.0%	0.0%	4.0%	0.0%
Text-Based Reasoning	Complex Reasoning	68.7%	68.7%	57.3%	49.3%	56.7%
Text-Based Video Understanding	Subtitle-Based Video Understanding	90.0%	93.0%	96.0%	90.0%	95.0%
	Multi-Hop Needle in A Haystack	17.0%	16.0%	14.0%	16.0%	20.0%
Robust Video Testing	AIGC Videos	86.0%	66.0%	86.0%	78.0%	82.0%
	Long Videos	42.0%	46.0%	50.0%	56.0%	54.0%
	Adversarial Videos	76.0%	84.0%	78.0%	80.0%	78.0%
Total	-	64.9%	61.0%	59.8%	59.1%	56.2%

Table 11: **Accuracy of evaluated MLLMs on each task of MME-VideoOCR**.

Task Category	**Task**	**Oryx-1.5 32B**	**Video-LLaMA 3**	**LLaVA Video-7B**	**Oryx-1.5 7B**
Text Recognition	Text Recognition at Designated Locations	52.5%	47.5%	49.0%	53.0%
	Text Recognition Based on Specific Attributes	46.0%	47.0%	43.0%	49.0%
Visual Text QA	Text-Centric QA	67.0%	63.5%	67.0%	62.0%
	Translation	32.0%	34.0%	28.0%	22.0%
Text Grounding	Spatial Grounding	73.0%	65.0%	70.0%	59.0%
	Temporal Grounding	54.0%	71.0%	52.0%	42.0%
Attribute Recognition	Color Recognition	66.0%	76.0%	84.0%	64.0%
	Named Entity Recognition	68.0%	66.0%	66.0%	64.0%
	Counting	54.0%	52.0%	56.0%	36.0%
Change Detection & Tracking	Change Detection	37.0%	39.0%	40.0%	35.0%
	Tracking	55.0%	61.0%	57.0%	54.0%
Special Text Parsing	Table Parsing	52.0%	44.0%	44.0%	50.0%
	Chart Parsing	46.0%	50.0%	42.0%	44.0%
	Document Parsing	76.0%	68.0%	64.0%	70.0%
	Mathematical Formula Parsing	74.0%	64.0%	56.0%	58.0%
	Handwriting Recognition	54.0%	44.0%	44.0%	42.0%
Cross-Frame Text Understanding	Scrolling Text Understanding	64.0%	60.0%	60.0%	68.0%
	Trajectory Recognition	0.0%	0.0%	0.0%	0.0%
	Scrambled Recognition	0.0%	4.0%	4.0%	2.0%
Text-Based Reasoning	Complex Reasoning	54.7%	48.7%	47.3%	48.7%
Text-Based Video Understanding	Subtitle-Based Video Understanding	86.0%	91.0%	93.0%	78.0%
	Multi-Hop Needle in A Haystack	36.0%	19.0%	20.0%	16.0%
Robust Video Testing	AIGC Videos	80.0%	78.0%	86.0%	80.0%
	Long Videos	52.0%	56.0%	54.0%	40.0%
	Adversarial Videos	72.0%	68.0%	66.0%	72.0%
Total	-	55.2%	53.5%	52.8%	49.6%

Table 12: **Accuracy of evaluated MLLMs on each task of MME-VideoOCR**.

Task Category	**Task**	**VITA-1.5**	**Slow-fast MLLM**	**Videochat-Flash-7B**	**LLaVA OneVision-7B**
Text Recognition	Text Recognition at Designated Locations	48.0%	46.0%	37.5%	42.0%
	Text Recognition Based on Specific Attributes	51.0%	46.0%	35.0%	42.0%
Visual Text QA	Text-Centric QA	63.0%	61.5%	55.5%	57.0%
	Translation	40.0%	28.0%	18.0%	22.0%
Text Grounding	Spatial Grounding	53.0%	61.0%	61.0%	58.0%
	Temporal Grounding	33.0%	43.0%	59.0%	40.0%
Attribute Recognition	Color Recognition	66.0%	66.0%	64.0%	66.0%
	Named Entity Recognition	58.0%	70.0%	66.0%	62.0%
	Counting	60.0%	44.0%	50.0%	34.0%
Change Detection & Tracking	Change Detection	37.0%	44.0%	43.0%	36.0%
	Tracking	61.0%	50.0%	55.0%	46.0%
Special Text Parsing	Table Parsing	44.0%	42.0%	32.0%	40.0%
	Chart Parsing	44.0%	42.0%	40.0%	40.0%
	Document Parsing	72.0%	64.0%	56.0%	56.0%
	Mathematical Formula Parsing	64.0%	60.0%	58.0%	56.0%
	Handwriting Recognition	42.0%	32.0%	44.0%	40.0%
Cross-Frame Text Understanding	Scrolling Text Understanding	60.0%	58.0%	58.0%	58.0%
	Trajectory Recognition	0.0%	0.0%	0.0%	0.0%
	Scrambled Recognition	0.0%	2.0%	0.0%	2.0%
Text-Based Reasoning	Complex Reasoning	51.3%	43.3%	50.0%	45.3%
Text-Based Video Understanding	Subtitle-Based Video Understanding	83.0%	83.0%	88.0%	86.0%
	Multi-Hop Needle in A Haystack	11.0%	14.0%	20.0%	18.0%
Robust Video Testing	AIGC Videos	68.0%	58.0%	78.0%	78.0%
	Long Videos	42.0%	38.0%	44.0%	36.0%
	Adversarial Videos	66.0%	66.0%	60.0%	66.0%
Total	-	49.5%	47.8%	47.8%	46.0%