A Representative Examples from MME-VideoOCR

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

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

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

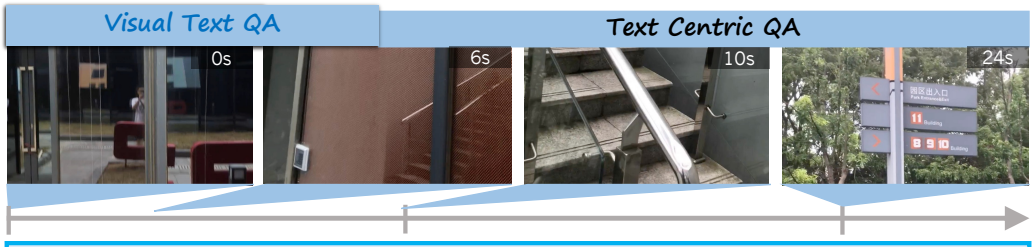
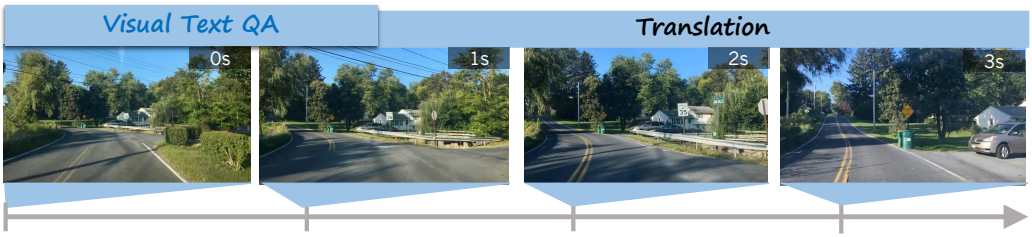
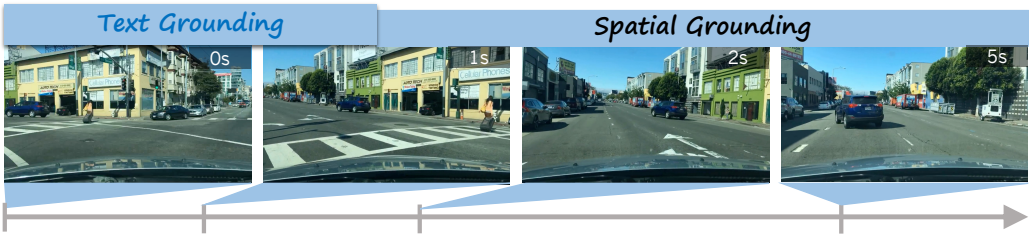


Figure 9: An example QA of the Text-Centric QA task in MME-VideoOCR.



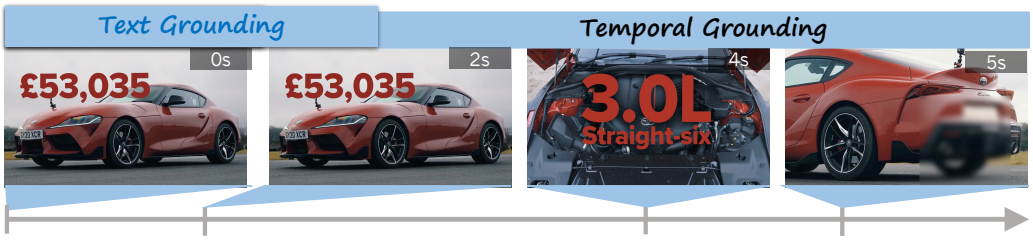
Question: Please translate the text above the number 35 on the white road sign into Chinese.
Answer: SPEED LIMIT
Evaluation Criteria: GPT-Assisted Scoring

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

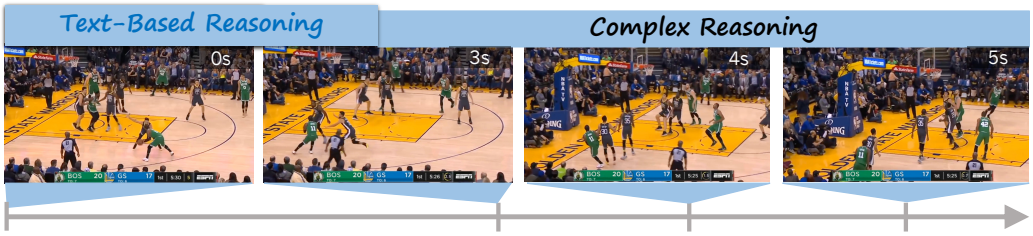
Figure 12: An example QA of the Temporal Grounding task in MME-VideoOCR.

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.

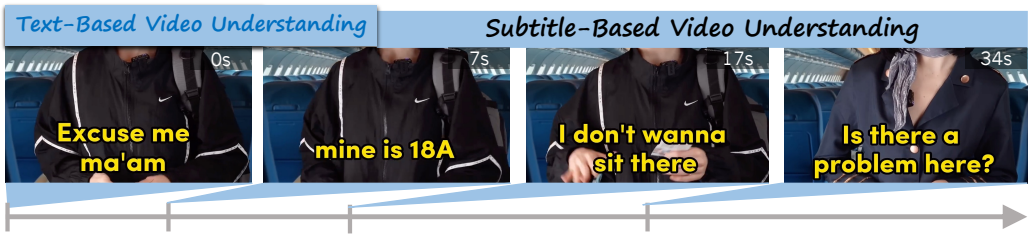
Question: How is the vehicle with the license plate number 5JVU366 moving?
Option:
A: Keeps going straight.
B: Moves into the left lane.
C: Moves into the right lane.
D: It stopped.
Answer: A
Evaluation Criteria: Multiple-Choice

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



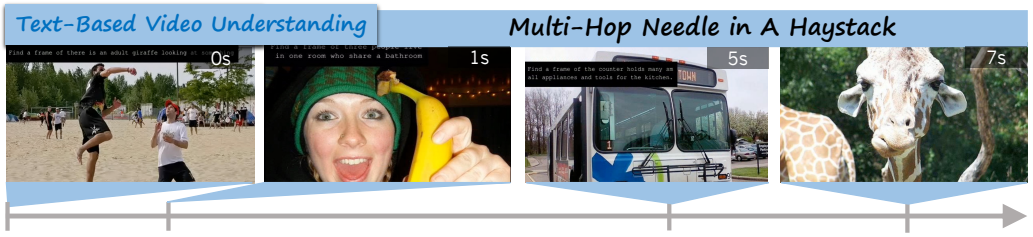
Question: How many points ahead will they be after making this shot?
Option:
A: 5
B: 4
C: 3
D: 6
Answer: A
Evaluation Criteria: Multiple-Choice

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



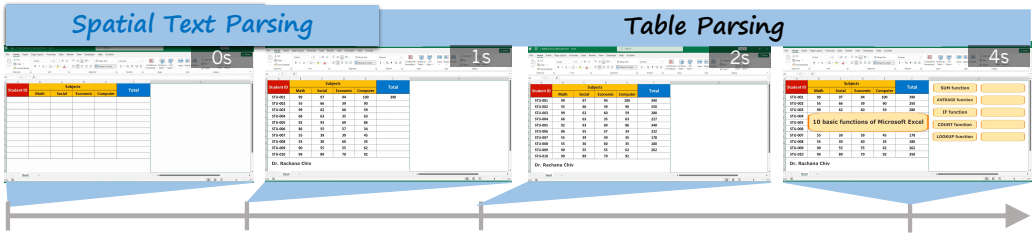
Question: What are the two passengers arguing about?
Option:
A: One passenger is unwilling to exchange seats with another passenger.
B: One passenger is unwilling to give up his seat to another passenger.
C: One passenger is unwilling to let another passenger sit next to him.
D: One passenger is unwilling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



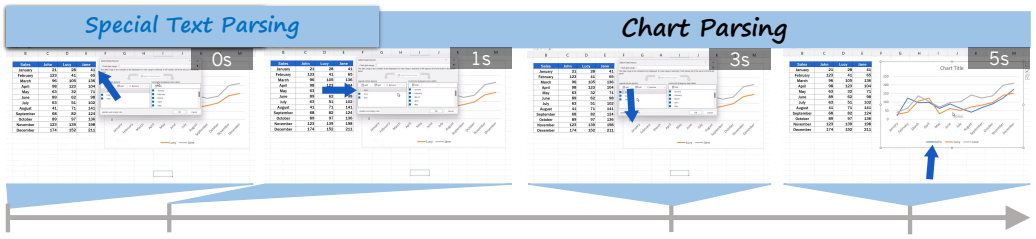
Question: Find the frame of a girl shows a banana to the camera. Locate the final frame based on the instructions in the image, and answer the following question based on the final frame: Which of the following objects or actions appears in the frame?
Option:
A: banana
B: giraffe
C: flower
D: bus
Answer: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

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



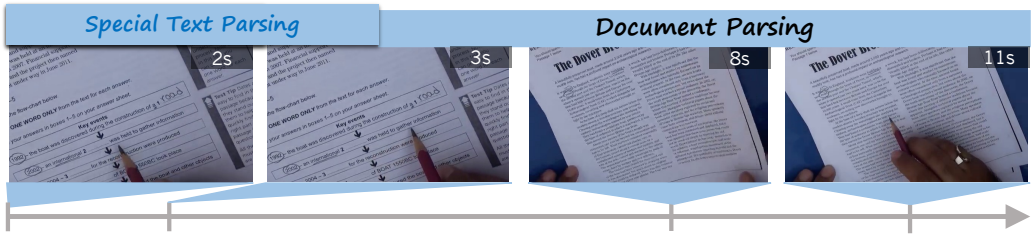
Question: How much bigger is the number for STU-006-Total than for STU-004-Total?
Option:
A: 5
B: 232
C: 227
D: Equal
Answer: A
Evaluation Criteria: Multiple-Choice

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



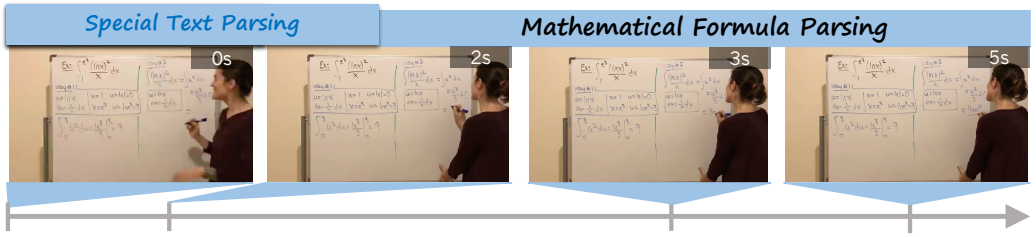
Question: What is the sales figure for March represented by the blue line in the line graph?
Option:
A: 136
B: 96
C: 105
D: 104
Answer: B
Evaluation Criteria: Multiple-Choice

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



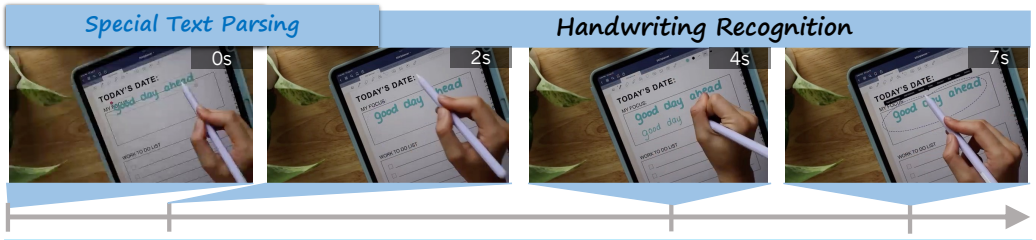
Question: According to the newspaper, which city does the road built in 1992 pass through?
Option:
A: Bradford
B: Canterbury
C: Dover
D: London
Answer: C
Evaluation Criteria: Multiple-Choice

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.

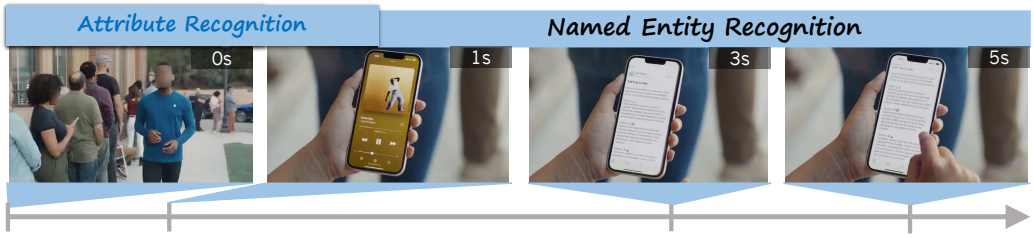


Question: What was written and then erased?
Answer: good day ahead
Evaluation Criteria: Containment Match

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

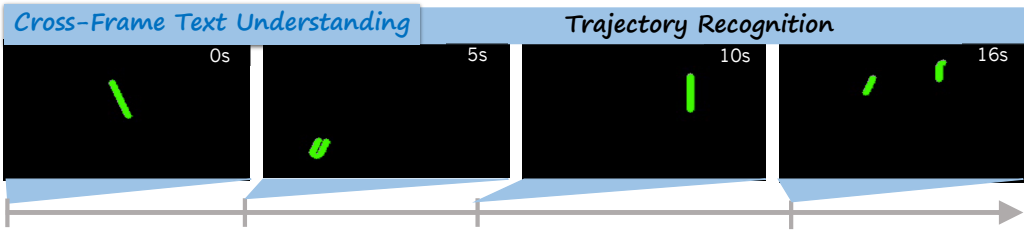
Figure 24: An example QA of the Named Entity Recognition task in MME-VideoOCR.

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.

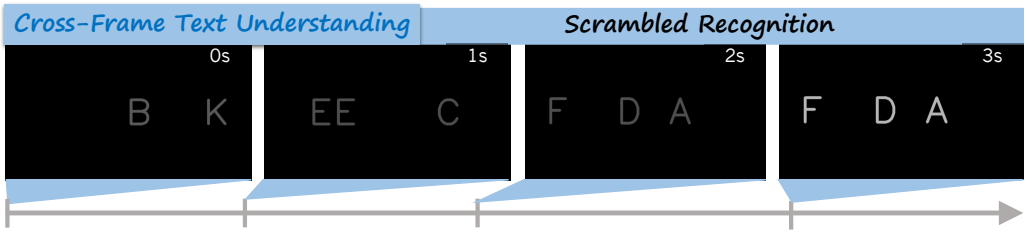
Question: What is the attitude of the green barrage towards this scenic spot?
Option:
A: excited
B: disgust
C: satisfied
D: interested
Answer: B
Evaluation Criteria: Multiple-Choice

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.

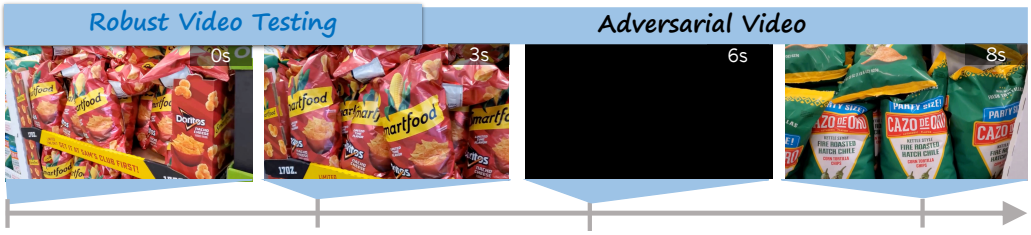


Question: What is formed by arranging these randomly appearing elements by position?
Answer: FEEDBACK
Evaluation Criteria: Containment Match

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%