

# Video-Bench: A Comprehensive Benchmark and Toolkit for Evaluating Video-based Large Language Models

Anonymous ACL submission

## Abstract

Video-adapted large language models (Video-LLMs) are pivotal for advancing artificial general intelligence (AGI) in video understanding. Despite progress, existing methods rarely undergo comprehensive assessment from an AGI construction perspective. We propose that an ideal video intelligence model should possess three essential abilities: (i) Video-exclusive Understanding, crucial for tasks like event summarization where direct video content analysis is paramount; (ii) Prior Knowledge-based Question-Answering, essential for applications needing contextual insights such as in-depth sports analysis or cultural understanding in music videos and television shows; (iii) Comprehension and Decision-making, vital for predictive tasks in complex environments like 3D scene navigation or autonomous vehicle guidance. To systematically evaluate these abilities, we introduce *Video-Bench*, an ability-oriented benchmark encompassing real-world video data and meticulously designed QA pairs, accompanied by an automated evaluation toolkit. Our analysis of 8 leading Video-LLMs show a significant gap in achieving human-like video understanding, underscoring the need for advancements in video comprehension AGI.

## 1 Introduction

Large language models (LLMs)(Radford et al., 2018, 2019; Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023a,b) have demonstrated strong capabilities in handling natural language processing (NLP) tasks, including comprehension, composition and reasoning, and achieved remarkable advancements on NLP benchmarks(Clark et al., 2018; Zellers et al., 2019; Hendrycks et al., 2020; Lin et al., 2021). This success has also inspired studies on Video-LLMs (Wang et al., 2022; Maaz et al., 2023; Li et al., 2023c,a; Su et al., 2023; Luo et al., 2023; Chen et al., 2023; Lyu et al., 2023; Wang et al., 2023), where models process video in-

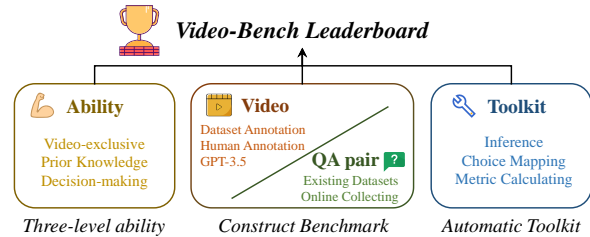


Figure 1: The illustrative pipeline for the intuition, construction and application of *Video-Bench*.

puts with textual prompts and generate corresponding answers, illuminating the prospective configuration of AGI for video understanding.

To gradually approach this goal, the establishment of an evaluation benchmark is indispensable for precisely measuring whether an artificial model possesses genuine video comprehension capabilities as humans. However, we find that existing benchmarks fall short of serving this purpose comprehensively. For instance, MMBench (Liu et al., 2023b) and LVLM-eHub (Xu et al., 2023) are concentrated on image understanding, ignoring the video understanding ability. SEED-Bench (Li et al., 2023b) includes several video tasks but is limited to temporal understanding. These benchmarks simply build some evaluation data and QA pairs in the traditional way, without measuring the limitations of existing Video-LLMs from the perspective of constructing video comprehension AGI.

With the ultimate goal of achieving AGI for video understanding, we assert that a truly intelligent video-language model should at least exhibit three distinct human-like capabilities: (i) Video-exclusive Understanding, i.e., performing well for questions whose answer can be extracted from the video itself; (ii) Prior Knowledge-based Question-Answering, i.e., answer questions that require the prior knowledge beyond the video, such as commentary on NBA games or providing background information on specific music videos; (iii) Comprehension and Decision-making, enabling a compre-

hensive understanding of scenarios, along with the ability to make predictions and informed decisions. Example applications encompass 3D scene understanding and decision-making for autonomous driving. To this end, we propose a new large-scale benchmark along with a toolkit, referred to as “*Video-Bench*”, to furnish a thorough evaluation of Video-LLMs, as depicted in Fig. 1.

In detail, aligning with our motivation, our *Video-Bench* encompasses tasks categorized into three distinct levels of capability: (i) For Video-exclusive Understanding, we begin by randomly selecting parts of traditional QA pairs (Xu et al., 2017; Yu et al., 2019; Jang et al., 2017), and proposing more challenging tasks to assess both temporal and contextual aspects of videos. Tasks include video summarization (Zhou et al., 2018), abnormal detection (Sultani et al., 2018), and crowd counting (Leal-Taixé et al., 2015); (ii) For Prior Knowledge-based Question-Answering, we evaluate the capability of model in understanding TV dramas (Lei et al., 2018), appreciating music videos, and providing information about players and games in NBA videos. (iii) For Comprehension and Decision-making, we employ two classical tasks: 3D indoor scene understanding (Ma et al., 2022) and auto-driving decision-making to assess the comprehension and decision-making abilities.

To streamline the evaluation process, we include another crucial component, i.e., the evaluation toolkit, along with the benchmarks. The toolkit automatically maps the long text outputs of Video-LLMs to corresponding answers with probability selection (Hendrycks et al., 2020) or LLM-based semantic understanding (Ouyang et al., 2022; Raffel et al., 2020). Subsequently, it calculates accuracy for each question and generates a final score, enhancing the efficiency of the evaluation workflow.

We evaluate eight representative Video-LLMs on *Video-Bench*: VideoChat (Li et al., 2023c), Video-ChatGPT (Maaz et al., 2023), Otter (Li et al., 2023a), Valley (Luo et al., 2023), PandaGPT (Su et al., 2023), mPLUG-Owl (Ye et al., 2023), Video-LLaMA (Zhang et al., 2023), and Chat-UniVi (Jin et al., 2023) with verified open-source model weights. The evaluation results reveal several interesting findings: (i) Most recent models can summarize the main content of videos but lack the capacity to detect details and temporal information. (ii) Due to the absence of domain-specific prior knowledge in the training data, these models encounter challenges in accurately comprehending

and responding to queries within a particular domain. (iii) Due to constraints in multimodal information extraction and the use of a weakened LLM backend (either 7B or 13B), the majority of tested models exhibit limited proficiency in comprehending and decision-making within complex scenarios. Our contributions can be summarized as follows:

- We suggest that there lacks specific measures for Video-LLMs, and propose the three-level ability assessment that systematically evaluates models in video-exclusive understanding, prior knowledge incorporation, and video-based decision-making abilities.
- We introduce *Video-Bench*, the first comprehensive evaluation benchmark for Video-LLMs, and provide a user-friendly evaluation toolkit. Accompanied by our datasets and QA pairs, the toolkit can streamline the performance assessment of Video-LLMs.
- We conduct extensive experiments to evaluate prominent Video-LLMs, summarizing their behaviors, analyzing main causes for observed limitations, and proposing future directions for improvement.

## 2 Related Work

**Video-LLMs.** A series Video-LLMs have emerged, building upon open-source LLMs (Touvron et al., 2023a,b; Chiang et al., 2023) or Image-LLMs (Alayrac et al., 2022; Awadalla et al., 2023; Liu et al., 2023a; Yuan et al., 2021). As outlined in Table 1, VideoChat (Li et al., 2023c) utilizes the Q-Former to map visual representations to Vicuna (Chiang et al., 2023), implementing a two-stage training process. Video-ChatGPT (Maaz et al., 2023) and Valley (Luo et al., 2023) originate from the LLaVA (Liu et al., 2023a) framework and introduce average pooling to enhance temporal sequence perception. Otter (Li et al., 2023a) proposes the MIMIC-IT dataset and fine-tunes Openflamingo (Awadalla et al., 2023) on their dataset. PandaGPT (Su et al., 2023) employs the ImageBind (Girdhar et al., 2023) as its backend for video comprehension. mPLUG-Owl (Ye et al., 2023) introduces an abstractor module to align image and text. Video-LLaMA (Zhang et al., 2023) incorporates a frame embedding layer and ImageBind to inject temporal and audio information into the LLM backend, while Chat-UniVi (Jin et al., 2023)

Table 1: Comparison between different Video-LLMs. ‘VE’, ‘TM’, ‘AE’, ‘LLM’, and ‘Adapt’ denote the visual encoder, temporal module, audio encoder, LLM backend and the adaptation module. The ‘CLIP (L)’ and ‘CLIP (O)’ represent the CLIP encoder pre-trained on LLaVA (Liu et al., 2023a) and OpenFlamingo (Awadalla et al., 2023). If the models are trained with two-stage, the training data of each stage is split by ‘/’. The ‘combined’ denotes the combination of typical V-L datasets including COCO (Chen et al., 2015), CC (Sharma et al., 2018), VG (Krishna et al., 2017), SBU (Ordonez et al., 2011) and LAION (Schuhmann et al., 2021).

Method	Model Configuration					Training Data	
	VE	TM	AE	LLM	Adapt	Source	Size
VideoChat (Li et al., 2023c)	BLIP-2	GMHRA	Whisper	Vicuna	Q-Former	Combined / Instruct-video	35M / 18K
Video-ChatGPT (Maaz et al., 2023)	CLIP (L)	AVG Pool	-	Vicuna	Linear	Instruct-video	100K
Otter (Li et al., 2023a)	CLIP (O)	-	-	LLaMA (O)	Linear	MIMIC-IT	2.8M
PandaGPT (Su et al., 2023)	ImageBind	-	ImageBind	Vicuna	Linear	LLAVA-mniGPT4	153.5K
Valley (Luo et al., 2023)	CLIP (L)	AVG Pool	-	Vicuna	Linear	WebVid / Instruct-video	702K / 47.8K
mPLUG-Owl (Ye et al., 2023)	CLIP	-	-	LLaMA	Abstractor	Combined / LLaVA	1100M / 150K
Video-LLaMA (Zhang et al., 2023)	BLIP-2	Frame Emb	ImageBind	Vicuna	Q-Former	WebVid / LLAVA-mniGPT4	2M / 153.5K
Chat-UniVi (Jin et al., 2023)	CLIP (L)	Cluster	-	Vicuna	Linear	Combined / Instruct-video	1.5M / 649K

merges visual tokens with similar semantic meanings using a clustering strategy. However, few of them try to address the challenges of temporal dimensions and audio modalities.

**Video Datasets.** Deep learning for video analysis relies on diverse datasets tailored to specific tasks. A notable task is human action recognition, featuring action classification datasets such as UCF-101 (Soomro et al., 2012), HMDB51 (Kuehne et al., 2011), and Kinetics (Kay et al., 2017), and action localization datasets like AVA (Gu et al., 2018) and Fineaction (Liu et al., 2022). Tasks involving anomaly detection in surveillance videos are addressed by datasets like UCSD-anomaly (Mahadevan et al., 2010) and UCF-crime (Sultani et al., 2018). Object identification and tracking in videos encompass multiple object tracking (MOT)(Leal-Taixé et al., 2015), video object segmentation (DAVIS)(Perazzi et al., 2016), and video instance segmentation (Youtube-VIS) (Yang et al., 2019). For multimodal tasks, video captioning datasets such as MSVD (Chen and Dolan, 2011), MSRVT (Xu et al., 2016), and Activitynet (Caba Heilbron et al., 2015) exist, along with their corresponding QA datasets (Xu et al., 2017,?; Yu et al., 2019). Scenario-specific datasets like MovieQA (Tapaswi et al., 2016) and TVQA (Lei et al., 2018) also contribute to the diversity of available datasets. However, these datasets often focus on specific tasks and lack the complexity to measure the comprehensive abilities of Video-LLMs.

**Vision Language Evaluation Benchmarks.** To evaluate the capabilities of LLMs, various benchmarks have been introduced, including AI2 Rea-

soning (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2020), and TruthfulQA (Lin et al., 2021). In the realm of multimodal LLMs, MMBench (Liu et al., 2023b) constructs a broad spectrum of evaluation for Vision-LLMs, and converts free-form predictions into predefined choices. SEED-Bench (Li et al., 2023b) introduces a series of temporal understanding tasks and establishes an automatic filtering. LVLMeHub (Xu et al., 2023) presents an online arena platform for user-level evaluation, simulating the real-world applications. ELEVATER (Li et al., 2022) focuses on evaluating the transferability of language-augmented visual models across multiple tasks. However, the aforementioned benchmarks are not tailored specifically for videos.

### 3 Video-Bench

In Fig.2, we show the overall structure of *Video-Bench* and the corresponding average results for existing Video-LLMs.

#### 3.1 Video-exclusive Summarization

As illustrated in Fig. 3 (A), we aim to measure the capacity of Video-LLMs to comprehend information from video itself, requiring no external prior knowledge or complex logic inference.

**Basic Understanding.** This task primarily evaluates the basic video recognition ability, such as responding to queries related to human actions in Activitynet-QA (Yu et al., 2019), providing answers related to objects, attributes, and actions corresponding to videos in MSVD-QA (Xu et al.,

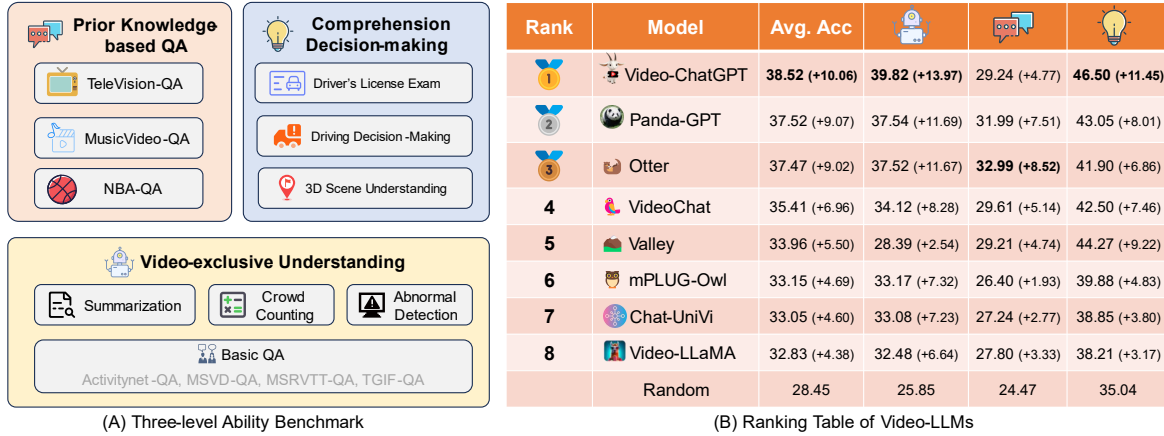


Figure 2: (A) part denotes the three-level evaluation of our *Video-Bench*. (B) part denotes the ranking of existing Video-LLMs. The reported results are accuracy (%) calculated with GPT-3.5. The number in parentheses represents the improvement over random results.

2017) and MSRVT-QA (Xu et al., 2017), and comprehending GIFs in TGIF-QA (Jang et al., 2017). **Summarization.** This task assesses the summarization ability of Video-LLMs when dealing with longer videos. Using the YouCook2 dataset (Zhou et al., 2018) with rich annotations and extended video duration, we generate a series of QA pairs to evaluate the ability to comprehend cooking information presented in the videos and audios. **Abnormal Detection.** This task evaluates the ability to review videos and identify anomalies. Leveraging the UCF-Crime dataset (Sultani et al., 2018), a collection of surveillance videos annotated with the type and timestamp of anomalies, we construct questions to assess the temporal comprehensive ability of Video-LLMs. **Crowd Counting.** This task primarily evaluates the ability to localize and count dense objects. Utilizing the MOT dataset (Leal-Taixé et al., 2015), which annotates all pedestrians, vehicles, and other targets in street or mall images, we test whether Video-LLMs can identify different pedestrians in different frames and provide the correct number.

### 3.2 Prior Knowledge-based Question-answering

As shown in Figure 3 (B), our objective is to evaluate Video-LLMs’ capacity to answer questions that necessitate prior knowledge. **TV-QA.** Utilizing the TVQA dataset (Lei et al., 2018), we transform image formats into videos, and incorporate audio and subtitles. This dataset allows us to evaluate the ability of Video-LLMs to integrate prior knowledge and information from video, audio, and text to answer questions related

to TV content.

**MV-QA.** Music videos pose a unique challenge due to their reliance on prior knowledge. In the absence of relevant existing datasets, we search for top music videos on YouTube and construct corresponding QA pairs based on authoritative wiki sources. This task assesses the ability of Video-LLMs to understand the song associated with the music video and provide answers regarding performers, background information, and relevant music theory knowledge.

**NBA-QA.** Understanding competitive sports videos also demands relevant prior knowledge to identify competing teams, players, technical actions, scores, or fouls within the video. We select top NBA plays from YouTube and manually annotate teams, players, and technical actions in each game, transforming them into question-answer pairs. These videos and questions serve as input to the model, expecting it to respond based on relevant prior knowledge.

### 3.3 Comprehension and Decision-making

As shown in Fig. 3 (C), to assess a similar capability in Video-LLMs, we propose evaluations in the realms of 3D scene understanding and autonomous-driving related tasks.

**3D Scene Comprehension.** Indoor scene comprehension and navigation hold significant practical implications. The complexity arises from the necessity for extensive knowledge-intensive reasoning to understand different situations (scenes and locations). The SQA3D dataset (Ma et al., 2022) is introduced to evaluate the 3D scene comprehension of Video-LLMs within the video modality.

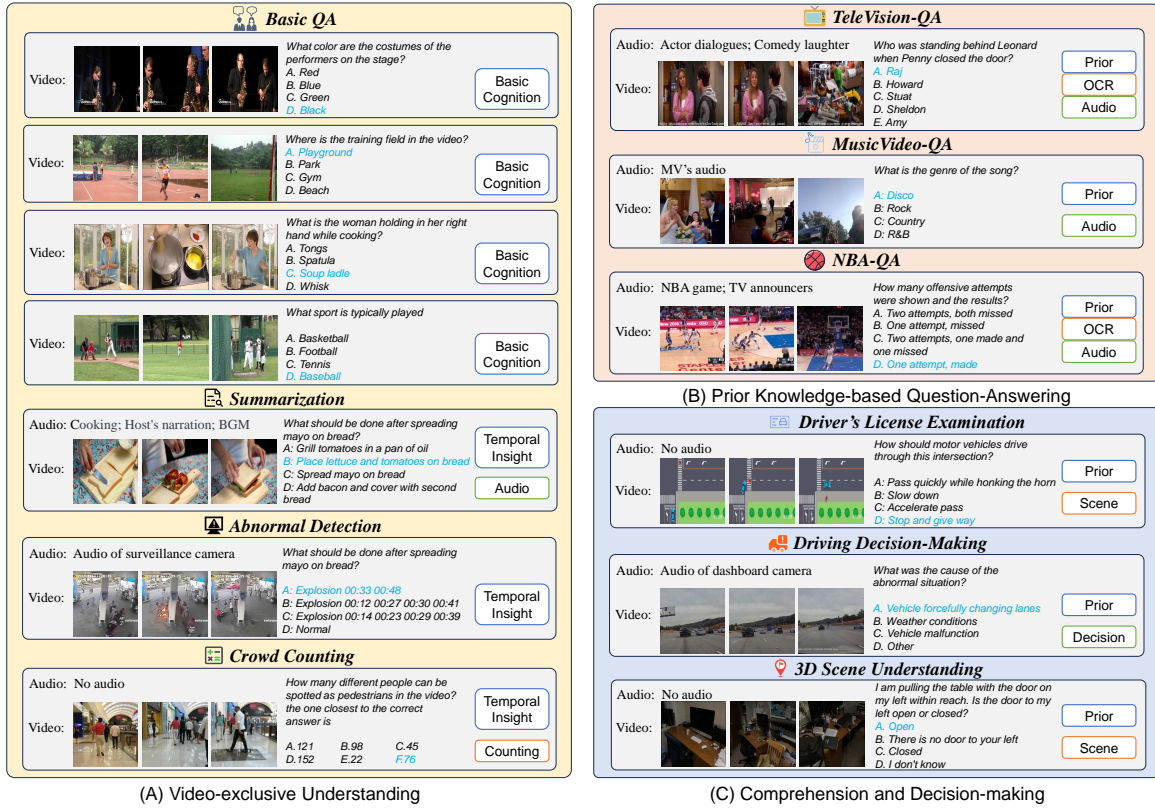


Figure 3: The detailed illustration of samples from each task and the corresponding ability required.

The models are tasked with understanding their environment and engaging in perception, reasoning, and action to accomplish the task.

**Driver's License Examination.** Video-based questions in driver's license examinations assess the ability of candidates to interpret simple animations depicting motor vehicle and driver status, requiring judgments of potential anomalies. In this task, we challenge Video-LLMs to comprehend scenarios and answer exam questions.

**Driving Decision-Making.** Making decisions for real-world driving scenarios is a more intricate task that demands a higher level of scene understanding and decision-making ability. For this task, we compile a diverse collection of YouTube driving videos depicting complex traffic situations and accidents. We conduct manual annotations for scene analysis and accident causes. Our expectation is that the model can effectively comprehend the origins of these complex traffic situations or accidents and make correct decisions to prevent their occurrence.

### 3.4 Automatic Evaluation Toolkit

LLMs are known for generating long-form text responses, often without adhering to a fixed format, making it challenging to quantify the correctness

of their answers. To address this, we propose an automatic evaluation toolkit to systematically assess the performance of Video-LLMs. Our toolkit provides three metrics to map the output of Video-LLMs to pre-defined answer choices and subsequently calculating the final scores. The first one is Probability (Hendrycks et al., 2020), a logits-based metric to acquire the probability of the next token following the prompt and treat the highest probability option as the prediction:

$$\text{Choice} = \arg \max_{i \in \{A, B, C, D, \dots\}} P(\text{Token}_i | \text{Prompt}). \quad (1)$$

The other two metrics are sentence-based, leveraging the natural language understanding capabilities of LLMs to obtain options. T5-based (Raffel et al., 2020) one calculates the textual similarities of generated sequences and options. GPT-3.5-based (Ouyang et al., 2022) transforms the sequences to a fixed format with prompt. All the above metrics can be implemented automatically with our toolkit, and users can analysis the ability of video-LLMs to comprehend video content and provide accurate responses to questions faithfully.

Table 2: Experiment results of tested Video-LLMs on various tasks. ‘\*’ denotes the QA-pairs are re-constructed or annotated by *Video-Bench*. ‘†’ denotes the tasks with fewer videos and multiplying the weight by 0.5 when calculating the final result. For each task, **blue** and **green** mark the **first** and **second** place respectively.

(A) Video-Exclusive Understanding										
Task	Dataset	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Basic QA	Activitynet-QA	34.58	44.55	46.60	44.30	44.96	38.10	41.47	39.85	<b>48.50</b>
	MSVD-QA*	26.22	42.15	<b>57.50</b>	54.95	50.43	32.03	42.45	41.15	38.15
	MSRVT-QA*	26.50	37.40	46.30	<b>46.95</b>	44.60	28.03	36.30	34.05	33.75
	TGIF-QA	22.37	33.74	<b>35.59</b>	34.27	29.66	31.41	31.66	31.28	33.48
	Average Score	25.85	34.12	<b>39.82</b>	37.52	37.54	28.39	33.17	32.48	33.08
Summarization	YouCook2*	25.00	27.66	<b>34.80</b>	32.65	33.02	29.05	27.05	28.90	29.00
Abnormal Detection	UCF-Cirme*	25.00	22.41	24.13	22.41	<b>33.01</b>	20.34	22.76	<b>27.59</b>	23.79
Crowd Counting	MOT*†	16.67	<b>27.78</b>	<b>27.78</b>	16.67	16.67	11.11	<b>27.78</b>	<b>16.67</b>	16.67
Average Score		25.85	34.12	<b>39.82</b>	37.52	37.54	28.39	33.17	32.48	33.08
(B) Prior Knowledge-based Question-Answering										
Task	Dataset	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Prior Knowledge	TV-QA*	20.00	26.15	<b>28.76</b>	27.65	27.85	23.70	23.95	24.75	22.20
	MV-QA*	26.15	34.11	36.52	<b>37.06</b>	<b>37.06</b>	32.59	30.17	32.41	34.29
	NBA-QA*	27.26	28.57	22.45	<b>34.26</b>	31.05	<b>31.34</b>	25.07	26.24	25.22
Average Score		24.47	29.61	29.24	<b>32.99</b>	<b>31.99</b>	29.21	26.40	27.80	27.24
(C) Comprehension and Decision-Making										
Task	Dataset	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Driving	License Exam*†	36.81	38.89	41.67	<b>52.78</b>	41.67	41.67	33.34	30.56	44.44
	Decision-Making*	44.21	55.38	<b>58.21</b>	48.72	56.03	56.54	51.03	49.10	47.95
3D Scene	SQA3D*	25.00	31.42	<b>37.20</b>	29.65	30.76	<b>33.30</b>	32.00	31.15	26.95
Average Score		35.04	42.50	<b>46.50</b>	41.90	43.05	44.27	39.88	38.21	38.85
(D) Final Result										
Average Score		28.45	35.41	<b>38.52</b>	37.47	37.52	33.96	33.15	32.83	33.05



Figure 4: The detailed amount of QA pairs of different tasks.

## 4 Experiment and Result

**Implementation details.** The detailed statistics of *Video-Bench* are listed in Fig. 4. To mitigate the impact of randomness, we multiply an additional weight of 0.5 for tasks with a smaller quantity of questions during the computation of the final average score. To ensure a fair comparison, we utilize the 7B LLM backend versions for all tested Video-LLMs during the inference process, thereby

mitigating language ability discrepancy stemming from different model sizes. The GPT-based metric of version *gpt-3.5-turbo-0613* are employed in the reported results by default.

**Results on Video-exclusive Understanding.** To evaluate the video-exclusive understanding ability, we validate Video-LLMs on the traditional basic QA tasks, summarization, abnormal detection and crowd counting tasks, as reported in Table. 2 (A). We have three observations. (i) Most Video-LLMs perform well on the four traditional QA datasets due to the simplicity of their questions, especially the Video-ChatGPT (Maaz et al., 2023) and Otter (Li et al., 2023a) with massive video instruction data, and the PandaGPT (Su et al., 2023) with a well-pretrained video encoder from ImageBind (Girdhar et al., 2023), which suggests extending the video data scale could be effective. (ii) Existing Video-LLMs are not temporal-sensitive. They cannot effectively summarize the order of each operation in YouCook2, and cannot respond effectively on the timestamp-related problems in UCF-Crime. (iii) These methods almost fail in the crowd counting task. These failure may come from the weak ability of precise locating and the temporal association.

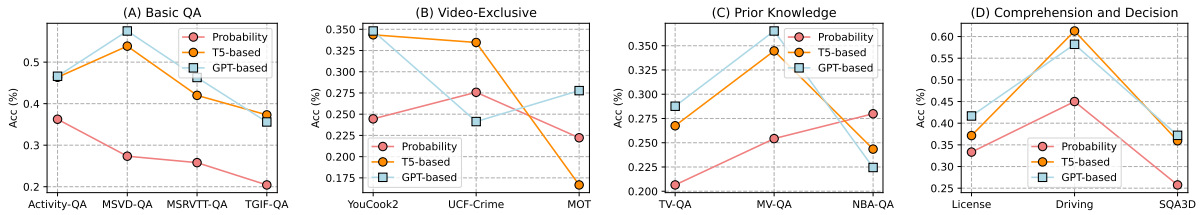


Figure 5: Comparison results of different metrics of Video-ChatGPT (Maaz et al., 2023) on all datasets.

**Results on Prior Knowledge-based QA.** Compared to enormous training data of LLMs, existing Video-LLMs are trained with limited instruction tuning data as Table. 1, resulting in the poor ability to recognize objects and information in specific domains. As shown in Table. 2 (B), we can have two observations. (i) Existing methods lack visual prior knowledge, which means they struggle to establish effective connection between the video and knowledge. For example, in NBA-QA task, even the players and technical actions are stored in the LLM backend, they cannot answer the questions when watching videos. Otter (Li et al., 2023a), which has the most instruction tuning data, achieves the best performance in this project, indicating that some prior knowledge is indeed contained in MIMIC-IT. (ii) Their poor performance on MV-QA indicates that they have limited audio understanding ability, since only some of the Video-LLMs possess audio modules. PandaGPT (Su et al., 2023) with the audio module of ImageBind shows the consistent results with the champion Otter (Li et al., 2023a) in MV-QA, proving that adding an audio encoder might improve this problem. In conclusion, existing Video-LLMs are requiring abundant prior knowledge pre-training for general domains on different modalities.

**Results on Comprehension and Decision-making.** The performance of existing Video-LLMs on 3D scene understanding and driving decision-making tasks is shown in Table. 2 (C). In these tasks, Video-ChatGPT (Maaz et al., 2023) continues to perform the best, thanks to its robust video instruction tuning. The followings are the Valley (Luo et al., 2023), which also possess powerful multi-modal understanding ability from vast instruct-tuning videos. To enhance the comprehensive and decision-making abilities, we suggest that future Video-LLMs must be trained with more prior knowledge and larger-scale data to cover more diverse domains. Besides, adopting Reinforcement Learning from Human Feedback (RLHF) and

larger model capability is also important for generalization and specific applications.

**Results on Different Metrics.** Our *Video-Bench* consists of a series of multiple-choice questions. Compared to open-ended questions, this test is relatively straightforward. However, due to the uncertainty and free form of LLM outputs, there is still room for designing more robust metrics. We evaluate the results of the best tested model, comparing the results with Probability (Hendrycks et al., 2020), T5-based and the GPT-based metrics. as shown in the Fig. 5. It can be seen that the result of Probability is overall low, because the output of Video-LLMs cannot effectively give a clear choice answer and the probability-based mapping may not faithfully reflect the correctness. Therefore, we recommend GPT as the metric to generate stable results.

## 5 Visualization and Multi-Dimension Analysis

**Visualization.** Fig. 6 illustrates a set of typical responses from tested Video-LLMs. It can be observed that only Video-ChatGPT (Maaz et al., 2023) provides the correct response, while other models engage in discussions related to the video but fail to make the correct judgment after a lengthy discourse. This highlights the issue that the models struggle with questions with even the most fundamental prior knowledge. This situation reflects the current state of Video-LLMs, which can generate responses related to videos while lacking trustful reference value.

**Multi-dimension Analysis.** In Fig. 7, a comparative analysis of Video-LLMs with different modules is presented. We can conclude that with the current data and training setting, Video-LLMs lack tailored focus on the three-level ability of video comprehension. And the empirically proposed modules have not yielded significant improvements.

We also analysis the impact of different data









How many offensive attempts were shown in the highlights and what were the results? <i>A. One attempt, made B. Two attempts, both missed C. One attempt, missed D. Two attempts, one made and one missed.</i> Among the four options A, B, C, D above, the one closest to the correct answer is:	
 <b>Video-Chat:</b> The video shows multiple offensive attempts in the game, including a made basket and a missed shot. However, there is no direct mention of the number of offensive attempts in the highlights, as the focus is on the basketball action. Based on the given information, the correct answer is B: Two attempts, both missed.	✗
 <b>Video-ChatGPT:</b> One attempt, made.	✓
 <b>Valley:</b> The correct answer is: One attempt was shown in the highlights, and it was missed.	✗
 <b>Otter:</b> One attempt was shown being made, but the outcome of that attempt is not mentioned in the given information.	✗
 <b>Panda-GPT:</b> The correct answer is none of the options above. The image shows a basketball game in progress, but it does not provide any information about the number of offensive attempts shown in highlights and their results.	✗
 <b>mPLUG-Owl:</b> D - Two attempts, one made and one missed. The highlights show two attempts, one missed and one made. The correct answer is D.	✗
 <b>Video-LLaMA:</b> C. One attempt, missed. The highlights showed a basketball player throwing an offensive foul and missing the ball. However, it was not clear if he made multiple attempts.	✗
 <b>Chat-UniVi:</b> D. Two attempts, one made and one missed.	✗

Figure 6: The illustrative sample of the generated responses from different Video-LLMs.

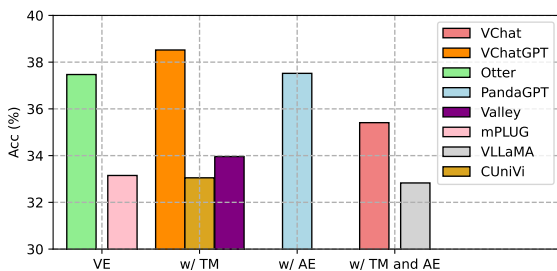


Figure 7: The impact of different module configuration. The w/ TM, w/ AE, and w/ TM and AE means the model contains Temporal Module (TM) or Audio Encoder (AE) or both of them.

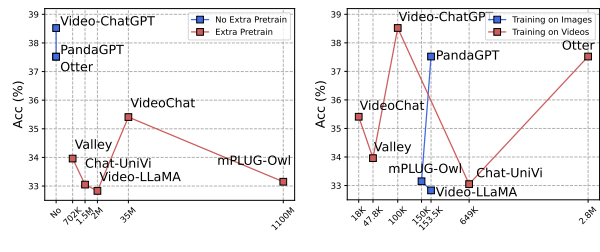


Figure 8: Impact of different datasize of pretrain data (left) or instruction tuning data (right).

## 6 Conclusion and Discussion

According to the above experimental results, we can conclude that the existing models are far from the truly intelligent Video-LLM and there are primarily three improvement directions.

**Vision Encoder with Temporal Awareness.** Existing methods process videos as frame clips, potentially missing crucial temporal information. Ideal Video-LLMs should understand the temporal sequence, possibly by selectively choosing keyframes or sampling frames to traverse the content efficiently.

**Domain-Specific Prior Knowledge Pre-training.** Lack of visual prior knowledge hinders accurate video comprehension. Incorporating domain-specific prior knowledge through pre-training can enhance domain expertise.

**Long Video Understanding.** One key differentiation point of Video-LLMs when compared to Image-LLMs should be the capability of processing long videos, which is highly neglected by existing research. Due to the memory and computation constraint, how to efficiently compress past frames and design an effective memory mechanism is crucial.

sizes in pre-training or instruction tuning process, as shown in Fig. 8. It can be observed that pre-training datasize may not necessarily play a decisive role, as the top-3 models, Video-ChatGPT (Maaz et al., 2023), PandaGPT (Su et al., 2023) and Otter (Li et al., 2023a), have no extra pretraining process. We suppose that the video encoders have received adequate training in multi-modal pre-training. In contrary, the influence of the instruction tuning datasize is notably evident, showing two trends: (i) The models trained on videos demonstrate overall better performance compared to those trained on images. This substantiates that native video data facilitates enhanced comprehension of video information by Video-LLMs. (ii) Model performance is positively correlated with the amount of video instruction tuning data. Video-ChatGPT (Maaz et al., 2023) and Otter (Li et al., 2023a) trained on large-scale video instruction tuning datasets are significantly better than other models.



516  
517  
518  
519  
520  
521  
522  
523  
524  
525

## **Limitations**

The scarcity of manually annotated data is the reason for this. However, the process of manual annotation actually provides us with an opportunity to cleverly integrate domain knowledge into the data. This not only enhances the authenticity and accuracy of the benchmark, but also makes it more professional and can better reflect the needs of practical applications. We will gradually enrich the dataset with more examples in our ongoing work.

## References

- 527 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc,  
528 Antoine Miech, Iain Barr, Yana Hasson, Karel  
529 Lenc, Arthur Mensch, Katherine Millican, Malcolm  
530 Reynolds, et al. 2022. Flamingo: a visual language  
531 model for few-shot learning. *Advances in Neural  
532 Information Processing Systems*, 35:23716–23736.
- 533 Anas Awadalla, Irena Gao, Josh Gardner, Jack Hes-  
534 sel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe,  
535 Yonatan Bitton, Samir Gadre, Shiori Sagawa, Je-  
536 nia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel  
537 Ilharco, Mitchell Wortsman, and Ludwig Schmidt.  
538 2023. Openflamingo: An open-source framework for  
539 training large autoregressive vision-language models.  
540 *arXiv preprint arXiv:2308.01390*.
- 541 Tom Brown, Benjamin Mann, Nick Ryder, Melanie  
542 Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind  
543 Neelakantan, Pranav Shyam, Girish Sastry, Amanda  
544 Askell, et al. 2020. Language models are few-shot  
545 learners. *Advances in neural information processing  
546 systems*, 33:1877–1901.
- 547 Fabian Caba Heilbron, Victor Escorcia, Bernard  
548 Ghanem, and Juan Carlos Niebles. 2015. Activitynet:  
549 A large-scale video benchmark for human activity  
550 understanding. In *Proceedings of the IEEE conference  
551 on computer vision and pattern recognition*, pages  
552 961–970.
- 553 David Chen and William B Dolan. 2011. Collecting  
554 highly parallel data for paraphrase evaluation. In  
555 *Proceedings of the 49th annual meeting of the associ-  
556 ation for computational linguistics: human language  
557 technologies*, pages 190–200.
- 558 Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang  
559 Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-  
560 llm: Bootstrapping advanced large language models  
561 by treating multi-modalities as foreign languages.  
562 *arXiv preprint arXiv:2305.04160*.
- 563 Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakr-  
564 ishna Vedantam, Saurabh Gupta, Piotr Dollár, and  
565 C Lawrence Zitnick. 2015. Microsoft coco captions:  
566 Data collection and evaluation server. *arXiv preprint  
567 arXiv:1504.00325*.
- 568 Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng,  
569 Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan  
570 Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al.  
571 2023. Vicuna: An open-source chatbot impressing  
572 gpt-4 with 90%\* chatgpt quality. See [https://vicuna.  
573 lmsys.org](https://vicuna.lmsys.org) (accessed 14 April 2023).
- 574 Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot,  
575 Ashish Sabharwal, Carissa Schoenick, and Oyvind  
576 Tafjord. 2018. Think you have solved question an-  
577 swering? try arc, the ai2 reasoning challenge. *arXiv  
578 preprint arXiv:1803.05457*.
- 579 Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Man-  
580 nat Singh, Kalyan Vasudev Alwala, Armand Joulin,  
and Ishan Misra. 2023. Imagebind: One embed-  
ding space to bind them all. In *Proceedings of the  
IEEE/CVF Conference on Computer Vision and Pat-  
tern Recognition*, pages 15180–15190.
- Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick,  
Caroline Pantofaru, Yeqing Li, Sudheendra Vijaya-  
narasimhan, George Toderici, Susanna Ricco, Rahul  
Sukthankar, et al. 2018. Ava: A video dataset of  
spatio-temporally localized atomic visual actions. In  
*Proceedings of the IEEE conference on computer  
vision and pattern recognition*, pages 6047–6056.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou,  
Mantas Mazeika, Dawn Song, and Jacob Steinhardt.  
2020. Measuring massive multitask language under-  
standing. *arXiv preprint arXiv:2009.03300*.
- Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim,  
and Gunhee Kim. 2017. Tgif-qa: Toward spatio-  
temporal reasoning in visual question answering. In  
*Proceedings of the IEEE conference on computer  
vision and pattern recognition*, pages 2758–2766.
- Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun  
Cao, and Li Yuan. 2023. Chat-univi: Unified vi-  
sual representation empowers large language models  
with image and video understanding. *arXiv preprint  
arXiv:2311.08046*.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang,  
Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio  
Viola, Tim Green, Trevor Back, Paul Natsev, et al.  
2017. The kinetics human action video dataset.  
*arXiv preprint arXiv:1705.06950*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin John-  
son, Kenji Hata, Joshua Kravitz, Stephanie Chen,  
Yannis Kalantidis, Li-Jia Li, David A Shamma, et al.  
2017. Visual genome: Connecting language and vi-  
sion using crowdsourced dense image annotations.  
*International journal of computer vision*, 123:32–73.
- Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote,  
Tomaso Poggio, and Thomas Serre. 2011. Hmdb: a  
large video database for human motion recognition.  
In *2011 International conference on computer vision*,  
pages 2556–2563. IEEE.
- Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth,  
and Konrad Schindler. 2015. Motchallenge 2015:  
Towards a benchmark for multi-target tracking. *arXiv  
preprint arXiv:1504.01942*.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg.  
2018. Tvqa: Localized, compositional video ques-  
tion answering. *arXiv preprint arXiv:1809.01696*.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang,  
Jingkang Yang, and Ziwei Liu. 2023a. Otter: A  
multi-modal model with in-context instruction tuning.  
*arXiv preprint arXiv:2305.03726*.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yix-  
iao Ge, and Ying Shan. 2023b. Seed-bench: Bench-  
marking multimodal llms with generative compre-  
hension. *arXiv preprint arXiv:2307.16125*.

637	Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. 2022. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. <i>Advances in Neural Information Processing Systems</i> , 35:9287–9301.	691
638		692
639		693
640		694
641		695
642		696
643	KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. <i>arXiv preprint arXiv:2305.06355</i> .	697
644		698
645		699
646		700
647	Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. <i>arXiv preprint arXiv:2109.07958</i> .	701
648		702
649		
650	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. <i>arXiv preprint arXiv:2304.08485</i> .	703
651		704
652		705
653	Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. 2022. Fineaction: A fine-grained video dataset for temporal action localization. <i>IEEE Transactions on Image Processing</i> , 31:6937–6950.	706
654		707
655		708
656		709
657	Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? <i>arXiv preprint arXiv:2307.06281</i> .	710
658		711
659		712
660		713
661		714
662	Ruipu Luo, Ziwanng Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. <i>arXiv preprint arXiv:2306.07207</i> .	715
663		716
664		717
665		718
666		719
667	Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. <i>arXiv preprint arXiv:2306.09093</i> .	720
668		721
669		722
670		723
671		724
672	Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2022. Sqa3d: Situated question answering in 3d scenes. In <i>The Eleventh International Conference on Learning Representations</i> .	725
673		726
674		727
675		728
676		729
677	Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. <i>arXiv preprint arXiv:2306.05424</i> .	730
678		731
679		732
680		733
681		734
682	Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly detection in crowded scenes. In <i>2010 IEEE computer society conference on computer vision and pattern recognition</i> , pages 1975–1981. IEEE.	735
683		736
684		737
685		738
686		739
687	Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. <i>Advances in neural information processing systems</i> , 24.	740
688		741
689		742
690		743
		744
		745
	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. <i>Advances in Neural Information Processing Systems</i> , 35:27730–27744.	
	Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 724–732.	
	Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.	
	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. <i>OpenAI blog</i> , 1(8):9.	
	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. <i>The Journal of Machine Learning Research</i> , 21(1):5485–5551.	
	Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. <i>arXiv preprint arXiv:2111.02114</i> .	
	Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 2556–2565.	
	Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. <i>arXiv preprint arXiv:1212.0402</i> .	
	Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. <i>arXiv preprint arXiv:2305.16355</i> .	
	Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 6479–6488.	
	Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 4631–4640.	

746	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier	Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella,	801
747	Martinet, Marie-Anne Lachaux, Timothée Lacroix,	Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong	802
748	Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal	Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence:	803
749	Azhar, et al. 2023a. Llama: Open and effi-	A new foundation model for computer vision. <i>arXiv</i>	804
750	cient foundation language models. <i>arXiv preprint</i>	<i>preprint arXiv:2111.11432</i> .	805
751	<i>arXiv:2302.13971</i> .		
752	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali	806
753	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	Farhadi, and Yejin Choi. 2019. Hellaswag: Can a	807
754	Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti	machine really finish your sentence? <i>arXiv preprint</i>	808
755	Bhosale, et al. 2023b. Llama 2: Open founda-	<i>arXiv:1905.07830</i> .	809
756	tion and fine-tuned chat models. <i>arXiv preprint</i>	Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-	810
757	<i>arXiv:2307.09288</i> .	llama: An instruction-tuned audio-visual language	811
758	Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai,	model for video understanding. <i>arXiv preprint</i>	812
759	Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023.	<i>arXiv:2306.02858</i> .	813
760	Chatvideo: A tracklet-centric multimodal and ver-	Luowei Zhou, Chenliang Xu, and Jason Corso. 2018.	814
761	satile video understanding system. <i>arXiv preprint</i>	Towards automatic learning of procedures from web	815
762	<i>arXiv:2304.14407</i> .	instructional videos. In <i>Proceedings of the AAAI</i>	816
763	Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo,	<i>Conference on Artificial Intelligence</i> , volume 32.	817
764	Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-		
765	Gang Jiang, and Lu Yuan. 2022. Omnivl: One foun-	<b>A T5 evaluation</b>	818
766	dation model for image-language and video-language	In our answer evaluation benchmark project, we	819
767	tasks. In <i>Thirty-sixth Conference on Neural Informa-</i>	explore two approaches: GPT-based metric and	820
768	<i>tion Processing Systems (NeurIPS 2022)</i> .	T5-based metric. T5-based metric serves as an	821
769	Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang	auxiliary tool in the evaluation process, offering	822
770	Zhang, Xiangnan He, and Yueting Zhuang. 2017.	advantages in terms of cost, deployment, and per-	823
771	Video question answering via gradually refined atten-	formance. It provides a cost-effective solution by	824
772	tion over appearance and motion. In <i>Proceedings of</i>	eliminating the need for ChatGPT API usage and	825
773	<i>the 25th ACM international conference on Multime-</i>	allows for offline deployment on personal servers.	826
774	<i>dia</i> , pages 1645–1653.	As shown in Table 3, T5-based results demonstrate	827
775	Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-	comparable performance to GPT-based in answer	828
776	vtt: A large video description dataset for bridging	evaluation tasks, making it a valuable addition to	829
777	video and language. In <i>Proceedings of the IEEE con-</i>	our benchmark project for reliable and efficient	830
778	<i>ference on computer vision and pattern recognition</i> ,	assessment.	831
779	pages 5288–5296.		
780	Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao,	<b>B Visualization Samples</b>	832
781	Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang,	In this part, we provide more samples of on all	833
782	Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A compre-	datasets concluded in <i>Video-Bench</i> , to illustrate the	834
783	hensive evaluation benchmark for large vision-	performance and behaviour of the tested Video-	835
784	language models. <i>arXiv preprint arXiv:2306.09265</i> .	LLMs.	836
785	Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video	<b>B.1 Video-exclusive Understanding</b>	837
786	instance segmentation. In <i>Proceedings of the</i>	<b>Activitynet-QA.</b> The results of the Activitynet-QA	838
787	<i>IEEE/CVF International Conference on Computer</i>	is shown in Fig. 9. As mentioned in Sec 4, Video-	839
788	<i>Vision</i> , pages 5188–5197.	LLMs perform well on these simple questions. The	840
789	Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye,	similar results are shown on the remaining three	841
790	Ming Yan, Yiyang Zhou, Junyang Wang, An-	datasets of <i>Basic QA</i> .	842
791	wen Hu, Pengcheng Shi, Yaya Shi, et al. 2023.	<b>MSVD-QA.</b> The results of the MSVD-QA is	843
792	mplug-owl: Modularization empowers large lan-	shown in Fig. 10. As part of the <i>Basic QA</i> , the	844
793	guage models with multimodality. <i>arXiv preprint</i>	performance of Video-LLMs here are overall good.	845
794	<i>arXiv:2304.14178</i> .	<b>MSRVTT-QA.</b> The results of the MSRVTT-QA is	846
795	Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuet-	shown in Fig. 11. The results shows a similar trend	847
796	ing Zhuang, and Dacheng Tao. 2019. Activitynet-qa:	of the above.	848
797	A dataset for understanding complex web videos via		
798	question answering. In <i>Proceedings of the AAAI Con-</i>		
799	<i>ference on Artificial Intelligence</i> , volume 33, pages		
800	9127–9134.		

Table 3: **T5-large experiment results of tested video-LLMs on various tasks.** “\*” denotes the QA-pairs are re-constructed or annotated by *Video-Bench*. “†” denotes the tasks with fewer videos and multiplying the weight by 0.5 when calculating the final result. For each task, **blue** and **green** mark the **first** and **second** place respectively. All the reported results are accuracy (%) calculated with GPT-3.5-based (Ouyang et al., 2022) metric. The “Video-” and “Chat-” are abbreviated to “V-” and “C-”.

<b>(A) Video-Exclusive Understanding</b>										
Task	Dataset	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Basic QA	Activitynet-QA	34.58	46.7	46.35	44.15	46.1	37.65	42.3	44.05	<b>47.95</b>
	MSVD-QA*	26.22	43.25	<b>53.85</b>	53.35	43.65	30.93	42.6	42.7	44.65
	MSRVT-QA*	26.50	37.0	42.0	<b>42.9</b>	39.4	28.48	34.65	34.75	41.25
	TGIF-QA	22.37	36.02	37.29	35.41	34.17	33.84	32.75	34.27	<b>42.30</b>
Summarization	YouCook2*	25.00	30.0	<b>34.35</b>	30.4	31.3	26.95	27.25	28.4	30.65
Abnormal Detection	UCF-Cirme*	16.67	18.62	<b>33.45</b>	26.21	24.83	13.45	18.62	22.07	<b>30.86</b>
Crowd Counting	MOT*†	16.67	22.22	16.67	<b>27.78</b>	5.56	11.11	11.10	16.67	11.11
Average Score		25.85	34.26	<b>39.33</b>	37.89	34.19	27.21	31.34	33.01	37.42
<b>(B) Prior Knowledge-based Question-Answering</b>										
Task	Dataset	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Prior Knowledge	TV-QA*	20.00	<b>28.3</b>	26.75	25.0	27.55	22.15	24.25	25.45	23.05
	MV-QA*	26.15	30.26	34.47	32.41	<b>34.91</b>	27.13	29.01	27.84	<b>33.48</b>
	NBA-QA*	27.26	25.36	24.34	<b>32.51</b>	26.53	25.36	26.82	28.13	24.49
Average Score		24.47	27.97	28.52	<b>29.97</b>	29.66	24.88	26.69	27.14	27.01
<b>(C) Comprehension and Decision-Making</b>										
Task	Dataset	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Driving	License Exam*†	36.81	25.0	37.14	<b>55.56</b>	36.11	30.56	36.11	25.0	<b>50.0</b>
	Decision-Making*	44.21	60.77	61.28	47.44	<b>62.18</b>	56.28	53.21	49.49	49.74
3D Scene	SQA3D*	25.00	30.08	<b>35.95</b>	27.45	30.25	35.65	32.35	30.5	27.4
Average Score		35.04	41.34	<b>46.32</b>	41.07	44.19	42.88	41.45	37.00	40.86
<b>(D) Final Result</b>										
Task	Dataset	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Average Score		28.45	34.53	<b>38.06</b>	36.31	36.02	31.66	33.16	32.38	35.10
<b>(E) Comparison Result of GPT-based</b>										
Task	Dataset	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Video-Exclusive Understanding		25.85	34.12	<b>39.82</b>	37.52	37.54	28.39	33.17	32.48	37.87
Prior Knowledge-based QA		24.47	29.61	29.24	<b>32.99</b>	31.99	29.21	26.40	27.80	27.43
Comprehension and Decision		35.04	42.50	<b>46.50</b>	41.90	43.05	44.27	39.88	38.21	40.64
Average Score		28.45	35.41	<b>38.52</b>	37.47	37.52	33.96	33.15	32.83	35.31

849 **TGIF-QA.** The results of the TGIF-QA is shown  
850 in Fig. 12. Results prove that Video-LLMS can  
851 also understand simple GIFs.

852 **YouCook2.** The results of the YouCook2 is shown  
853 in Fig. 13. The poor results show that existing  
854 Video-LLMs possess limited temporal awareness,  
855 and they are difficult to summarize the sequence of  
856 action steps.

857 **UCF-Crime.** The results of the UCF-Crime is  
858 shown in Fig. 14. The poor performance illustrates  
859 the existing Video-LLMs lack the ability of tempo-  
860 ral perception again.

861 **MOT.** The results of the MOT is shown in Fig. 15.  
862 Existing Video-LLMs are proved to lack the ability  
863 to count accurately.

## 864 **B.2 Prior Knowledge-based** 865 **Question-Answering**

866 **TV-QA.** The results of the TV-QA is shown in  
867 Fig. 16, which demonstrate that existing Video-  
868 LLMs can hardly understand TV segments. This  
869 could be caused by the lack of prior knowledge and  
870 audio or subtitle understanding ability.

871 **MV-QA.** The results of the MV-QA is shown in  
872 Fig. 17. The poor performance may be also caused  
873 by the lack of prior knowledge and audio under-  
874 standing ability.

875 **NBA-QA.** The results of the NBA-QA is shown  
876 in Fig. 18, which illustrates that without vision-  
877 language pre-training for specific domains, the  
878 Video-LLMs can not connect the knowledge stored  
879 in LLM with visual content and response to corre-  
880 sponding questions.

## 881 **B.3 Comprehension and Decision-Making**

882 **Driver’s License Examination.** The results of  
883 the Driver’s License Examination is shown in  
884 Fig. 19. The poor performance validates the tested  
885 Video-LLMs have limited scene understanding and  
886 decision-making ability.

887 **Driving Decision-Making.** The results of the Driv-  
888 ing Decision-Making is shown in Fig. 20, which  
889 demonstrates the tested Video-LLMs are difficult  
890 to understand the real driving environment.

891 **SQA3D.** The results of the SQA3D is shown in  
892 Fig. 21. The results show that they can only un-  
893 derstand the simple environment and cannot under-  
894 stand the complex spatial relationship.

895 This is an appendix.

Image	Question	Options	Video-ChatGPT	Video-LLaMA	Video-Char	Video-GPT	Video-PLUG-Out	Video-Otter	Video-Valley	Video-ChatGPT
	What is the gender of the person in the video? A.Male B.Female Among the two options A, B above, the one closest to the correct answer is:		Video-ChatGPT: The person in the video is a male. ✓	Video-LLaMA: A. The person in the video is male. ✓	Video-Char: Based on the video, the gender of the person trimming the bush in the video is not specified. It is therefore impossible to determine their gender with certainty. ✗	Video-GPT: The person in the video is a woman. ✗	Video-PLUG-Out: The person in the video is male. ✓	Video-Otter: The video does not provide enough information to determine the gender. ✗	Video-Valley: 1. A male ✓	Video-ChatGPT: The person in the video is a male. ✓
	Is the person in white outdoors? A.Yes B.No Among the two options A, B above, the one closest to the correct answer is:		Video-ChatGPT: Yes, the person in white is outdoors. ✓	Video-LLaMA: The video description and caption provided do not specify that the person in white is outdoors. However, in the video, there is a black and white picture of a man wearing white pants jumping off a wall, which is a close-up. .... ✓	Video-Char: A. Yes, the person is in white and outdoors. From the given video, it is clear that the person in white is doing an acrobatic stunt on the skateboard. ✓	Video-GPT: The person in white is outdoors, as they appear to be performing tricks on a sidewalk or a park bench. ✓	Video-PLUG-Out: The person in white is outdoors, as they appear to be performing tricks on a sidewalk or a park bench. ✓	Video-Otter: I cannot confidently determine the exact setting of the video, but it appears to be an outdoor location, possibly a park or a street. ✓	Video-Valley: The person in white is outdoors, and the photo was taken during the day. ✓	Video-ChatGPT: The person in white is outdoors. ✓

Figure 9: The illustrative sample of the generated responses from different Video-LLMs for Activitynet-QA.

Image	Question	Options	Video-ChatGPT	Video-LLaMA	Video-Char	Video-GPT	Video-PLUG-Out	Video-Otter	Video-Valley	Video-ChatGPT
	What are racers running down? A.Track B.Field C.Trail D.Road Among the four options A, B, C, D above, the one closest to the correct answer is:		Video-ChatGPT: The racers are running down a track. ✓	Video-LLaMA: The racers are running down a track. ✓	Video-Char: The video shows people running on a track during a race. Based on the given descriptions and the clues, it can be inferred that the racers are running in a track meet or a track race. The closest option to the correct answer is option C. ✓	Video-GPT: The racers are running down a track, not a trail, field, or road. In the image, there are several people running in different directions, and they are all on a track. This is evident from the presence of multiple runners. .... ✓	Video-PLUG-Out: A. Track. Track is the correct description. In the given image, four runners, wearing different colored jerseys, are running down a track. This scene typically takes place on a track field or an athletics track. It is not a field, trail, or a road. ✓	Video-Otter: Racers are running down a track. ✓	Video-Valley: The racers are running down a track. ✓	Video-ChatGPT: The racers are running down a track. ✓
	What is a man opening containing a gun? A.chest B.bag C.briefcase D.box Among the four options A, B, C, D above, the one closest to the correct answer is:		Video-ChatGPT: The man is opening a box containing a gun. ✓	Video-LLaMA: The man is opening a box containing a gun. ✓	Video-Char: The man in the video is opening a box containing a gun. Therefore, the closest option to the correct answer is A. It's difficult to see the exact contents of the box, but there is a person opening it, and they seem to be holding a gun. .... ✓	Video-GPT: The man is likely a toy, and the man is likely opening a box containing the toy gun. ✓	Video-PLUG-Out: The correct answer is C. The man is opening a briefcase containing a gun. ✗	Video-Otter: The gun is likely a toy, and the man is likely opening a box containing the toy gun. ✓	Video-Valley: 1. A. The man is opening a chest containing a gun. ✗	Video-ChatGPT: The man is opening a box containing a gun. ✓

Figure 10: The illustrative sample of the generated responses from different Video-LLMs for MSVD-QA.

Image	Question	Options	Video-ChatGPT	Video-LLaMA	Video-Char	Video-GPT	Video-PLUG-Out	Video-Otter	Video-Valley	Video-ChatGPT
	Who argues with a scary rabbit in a cartoon television show? A.dog B.child C.woman D.man Among the four options A, B, C, D above, the one closest to the correct answer is:		Video-ChatGPT: The person who argues with a scary rabbit in a cartoon television show is a man. ✓	Video-LLaMA: The correct answer is B. The video shows a young boy holding a sign with the text "Don't Be Afraid." A young man in a green t-shirt with the text "You are not alone" also appears in some frames. ✗	Video-Char: Based on the description, it seems that the woman is arguing with the scary rabbit in the cartoon television show. Therefore, the answer is C. woman. ✓	Video-GPT: A man is likely to be the one arguing with the scary bunny, as he is the one shown speaking to the. ✓	Video-PLUG-Out: A. dog. ✗	Video-Otter: A man is likely to be the one arguing with the scary bunny, as he is the one shown speaking to the. ✓	Video-Valley: 100%. ✓	Video-ChatGPT: The person who argues with a scary rabbit in a cartoon television show is a man. ✓
	What is played? A.baseball B.football C.tennis D.baseball Among the four options A, B, C, D above, the one closest to the correct answer is:		Video-ChatGPT: The video shows a group of people playing baseball on a field. ✓	Video-LLaMA: D. baseball. The video shows a baseball game, and the descriptions mention a baseball field, players, a baseball game, a boy, and a young boy playing baseball, making it clear that the sport being played is indeed baseball. ✓	Video-Char: From the given footage, the field appears to be lined with a chain link fence and has trees growing behind it. There are also several baseball players in the scene, and it is likely that they are playing baseball. The men are. .... ✓	Video-GPT: The image shows a group of people playing baseball on a field. There are several players on the field, and one of them is holding a baseball bat, indicating that they are actively engaged in the game. Baseball is a popular. .... ✓	Video-PLUG-Out: B. Football. ✗	Video-Otter: The sport being played is baseball. ✓	Video-Valley: The correct answer is A. The image shows a group of people playing a game of baseball on a field. ✓	Video-ChatGPT: The video shows a group of people playing baseball on a field. ✓

Figure 11: The illustrative sample of the generated responses from different Video-LLMs for MSRVTT-QA.

Image	Question	Options	Video-ChatGPT	Video-LLaMA	Video-Char	Video-GPT	Video-PLUG-Out	Video-Otter	Video-Valley	Video-ChatGPT
	What does the guy do 2 times? A.rub forehead B.brush teeth C.shake head D.hair teeth E.shake hips Among the five options A, B, C, D, E above, the one closest to the correct answer is:		Video-ChatGPT: The guy rubs his forehead twice. ✓	Video-LLaMA: It looks like the man is rubbing his forehead in the video. ✓	Video-Char: The guy rubs his forehead two times. Specifically, he rubs his forehead with his hand in two different instances in the video. This is evident from the descriptions, which specify that the guy is rubbing his forehead with his hand. .... ✓	Video-GPT: The guy sighs and shakes his head, indicating that he is frustrated or upset about something. ✓	Video-PLUG-Out: A. ✓	Video-Otter: The guy sighs and shakes his head, indicating that he is frustrated or upset about something. ✓	Video-Valley: The guy in the image is rubbing his forehead twice. ✓	Video-ChatGPT: The guy rubs his forehead twice. ✓
	What does the man do 3 times? A.move shoulders B sway body C.rise D.step E.tick neck Among the five options A, B, C, D, E above, the one closest to the correct answer is:		Video-ChatGPT: The man in the video does not move his shoulders, sway his body, or trot. He does, however, step out of the room and into the hallway. ✗	Video-LLaMA: A. The man moves his shoulders. ✓	Video-Char: The man in the video is shown swaying his body at least four times, and it is unclear whether he is moving his shoulders, stepping, or kicking his neck. However, based on the provided caption, the man is shown moving his body, but. .... ✗	Video-GPT: The man in the image does not do any of the actions listed above three times. Instead, he is seen moving his shoulders, swaying his body, and stepping to the side. There is no evidence of him kicking his neck or rotating his head. ✗	Video-PLUG-Out: The man moves his shoulders 3 times. The correct answer is A. ✓	Video-Otter: The man takes off his sunglasses. ✗	Video-Valley: The man in the video is seen making different gestures and movements three times. In the first instance, he is seen making a gesture with his hands in front of his face. In the second instance, he is seen making a gesture with his hands. .... ✗	Video-ChatGPT: The man in the video does not move his shoulders, sway his body, or trot. He does, however, step out of the room and into the hallway. ✗

Figure 12: The illustrative sample of the generated responses from different Video-LLMs for TGIF-QA.

Video	Question	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT
	What should be done after cooking the outer layer of the meat? A. Place them in a refrigerator. B. Brush some soy sauce on top. C. Sprinkle some salt and pepper. D. Flip them over. Among the four options A, B, C, D above, the one closest to the correct answer is:	Video-ChatGPT: After cooking the outer layer of the meat, it should be flipped over and cooked for a few more minutes.	Video-ChatGPT: The correct action after cooking the outer layer of the meat is to place them in a refrigerator. This is because the meat should be preserved in a cold place to prevent spoilage and bacterial growth, and to keep it fresh for...	Video-ChatGPT: A, B, and C are incorrect. D is the correct answer. After cooking the outer layer of the meat, one should flip the steak to ensure that it is evenly cooked. Additionally, brushing some soy sauce or other sauce on top of the meat after...	Video-ChatGPT: The correct answer is C, which is the correct spelling of the word "brush".	Video-ChatGPT: After cooking the outer layer of the meat, the next step should be to flip them over. This ensures that the meat is cooked evenly and reaches the desired level of doneness. Once the outer layer is cooked, the meat can be removed.....	Video-ChatGPT: A) Place them in a refrigerator after cooking the outer layer.	Video-ChatGPT: After cutting the beef into square patties, the next step is to cook them on a grill or in a pan. The video shows a man cooking the patties on a grill.	Video-ChatGPT: The correct answer is B. Toast the buns. After cutting the beef into square patties, the next step is to toast...

Figure 13: The illustrative sample of the generated responses from different Video-LLMs for YouCook2.

Video	Question	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT
	Known common crime types are: "Shooting", "Arson", "Arrest", "Fighting", "RoadAccidents", "Stealing", "Burglary", "Explosion", "Robbery", "Vandalism", "Shoplifting", "Abuse", "Assault". Each option below is the crime type and the start and end time of each criminal behavior video clip (if the video has no crime type and is normal, there is no corresponding start and end time of the video clip). A. Arson 00:24 00:37 B. Arson 02:57 03:15 03:28 03:44 C. Explosion 00:14 00:27 03:07 03:27 D. Normal. Please choose the correct answer by selecting the corresponding letter (A, B, C, or D) in the options provided.	Video-ChatGPT: The correct answer is D. Normal. The video shows a group of people working at a desk in an office, with no visible signs of any criminal activity or unusual behavior.	Video-ChatGPT: The one closest to the correct answer is option D, which represents a normal situation without any criminal activity.	Video-ChatGPT: The correct answer is A. Arson. The crime type is arson, and the start and end times of the associated crime video clip are 00:24 to 00:37.	Video-ChatGPT: The correct answer is D. Normal. The video shows a group of people working at a desk in an office, with no visible signs of any criminal activity or unusual behavior.	Video-ChatGPT: The correct answer is D. Normal. The video shows a group of people working at a desk in an office, with no visible signs of any criminal activity or unusual behavior.	Video-ChatGPT: The correct answer is D. Normal. The video shows a group of people working at a desk in an office, with no visible signs of any criminal activity or unusual behavior.	Video-ChatGPT: The correct answer is D. Normal. The video shows a group of people working at a desk in an office, with no visible signs of any criminal activity or unusual behavior.	Video-ChatGPT: The correct answer is D. Normal. The video shows a group of people working at a desk in an office, with no visible signs of any criminal activity or unusual behavior.

Figure 14: The illustrative sample of the generated responses from different Video-LLMs for UCF-Crime.

Video	Question	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT	Video-ChatGPT
	In the video, how many pedestrians can be recognized as different individuals? A. 114 B. 176 C. 52 D. 26 E. 88 F. 140 Among the six options A, B, C, D, E, F above, the one closest to the correct answer is:	Video-ChatGPT: The correct answer is 176.	Video-ChatGPT: The correct answer is A. 114 pedestrians can be recognized as different individuals in the video.	Video-ChatGPT: The correct answer is A. A total of 114 different individuals can be recognized in the video.	Video-ChatGPT: The correct answer is D. 26 pedestrians can be recognized as different individuals in the video.	Video-ChatGPT: The correct answer is A. There are 176 individuals walking in the video.	Video-ChatGPT: The correct answer is A. There are 176 individuals walking in the video.	Video-ChatGPT: The correct answer is A. There are 176 individuals walking in the video.	Video-ChatGPT: The correct answer is A. There are 176 individuals walking in the video.

Figure 15: The illustrative sample of the generated responses from different Video-LLMs for MOT.



	<b>Why does Howard say they're late after walking in ?</b> A. They got lost. B. They went out to dinner. C. There was a lot of traffic. D. They had to swing by the lab. E. Bernie took forever getting ready. Among the four options A, B, C, D, E, the one closest to the correct answer is:	
	<b>Video-ChatGPT:</b> Howard says they are late because they had to swing by the lab.	✗
	<b>Valley:</b> 1. They got lost.	✗
	<b>Other:</b> It is unclear why they are late, as there is no context provided in the video.	✗
	<b>Panda-GPT:</b> The correct answer is none of the options above. Howard says they are late because Bernie took forever getting ready. This implies that Bernie took a longer time than expected to get dressed, which caused them to arrive .....	✗
	<b>mPLUG-Owl:</b> They got lost.	✗
	<b>Video-LLaMA:</b> A. They got lost.	✗
	<b>Video-Char:</b> The option closest to the correct answer would be D. There was a lot of traffic. In the video, a woman wearing a blue jacket and a man are sitting in a kitchen, and Howard says "We're late!" as he is greeting them.....	✗
	<b>Chat-LLaV:</b> B. They went out to dinner.	✗

	<b>How does Emily react when Raj tells her about Lucy ?</b> A. She gets up and walks out. B. She says he does it need to tell her who she is seeing. C. She throws her drink in his face. D. She tells him that she does it feel good. E. She runs to the bathroom and climbs out the window. Among the five options A, B, C, D, E above, the one closest to the correct answer is:	
	<b>Video-ChatGPT:</b> Emily reacts by telling Raj that she doesn't feel good and runs to the bathroom and climbs out the window.	✗
	<b>Valley:</b> A. She gets up and walks out.	✗
	<b>Other:</b> The woman in the video appears to be upset, but there is no physical altercation or aggression shown towards the man. It is possible that she is simply expressing her disappointment or frustration with his actions.	✓
	<b>Panda-GPT:</b> None of the options presented in the question is the correct answer. The image shows a woman sitting at a table with a glass of wine, and a man is telling her about another woman he is seeing. The woman does not react in .....	✗
	<b>mPLUG-Owl:</b> A. Emily gets up and walks out.	✗
	<b>Video-LLaMA:</b> According to the video descriptions, when Raj tells Emily about Lucy, Emily gets up and walks out. This implies that Emily is not interested in hearing more of Raj's conversation with Lucy and decides to leave the table .....	✗
	<b>Video-Char:</b> Based on the given dialogue and description, the most likely reaction from Emily when Raj tells her about Lucy is that she will get up and walk out. In the video, Raj says "You can still say what you want with her. You know ....."	✗
	<b>Chat-LLaV:</b> Emily gets up and walks out when Raj tells her about Lucy.	✗

Figure 16: The illustrative sample of the generated responses from different Video-LLMs for TV-QA.

	<b>Who produced the song?</b> A. Sia B. All of the above C. Sean Paul Henriques D. Greg Kurstin. Among the four options A, B, C, D above, the one closest to the correct answer is:	
	<b>Video-ChatGPT:</b> The song was produced by Sean Paul Henriques.	✗
	<b>Valley:</b> A. Sia	✗
	<b>Other:</b> The song was co-written by Sia, Sean Paul, and Greg Kurstin.	✗
	<b>Panda-GPT:</b> The song was produced by Greg Kurstin.	✓
	<b>mPLUG-Owl:</b> A. Sia	✗
	<b>Video-LLaMA:</b> A. Sean Paul Henriques. Sia is not credited for this song. Greg Kurstin and Sean Paul Henriques are also not credited.	✗
	<b>Video-Char:</b> The closest option to the correct answer is "all of the above". The song "I'm Gonna Getcha Good!" was co-written by Sean Paul Henriques and produced by him and Greg Kurstin.....	✗
	<b>Chat-LLaV:</b> the song was produced by Greg Kurstin.	✓

	<b>What is one use of the song in media?</b> A. Theme song for wrestling events. B. Commercial for Omega 51 featuring Rory McIlroy C. Theme song for a Disney Princess series. D. All of the above. Among the four options A, B, C, D above, the one closest to the correct answer is:	
	<b>Video-ChatGPT:</b> The correct answer is A. The song is used as a theme song for wrestling events.	✗
	<b>Valley:</b> 100	✗
	<b>Other:</b> One use of music in media is to create a mood or atmosphere, as demonstrated in the video when the woman begins to dance with the man while the song plays in the background.	✗
	<b>Panda-GPT:</b> The correct answer is none of the above. The song is not being used for any of ....	✗
	<b>mPLUG-Owl:</b> The correct answer among the four options is D. The song is a theme song for ....	✓
	<b>Video-LLaMA:</b> The song in media can be used in several contexts, including music production, advertising, and film soundtracks. In the given video, there is no clear indication of .....	✗
	<b>Video-Char:</b> Based on the lyrics of the song, it could be used in media as a theme song for a sports-related event or a commercial for a brand. The lyrics mention sports and competition, .....	✗
	<b>Chat-LLaV:</b> The correct answer is A. The song is used as the theme song for wrestling events.	✗

Figure 17: The illustrative sample of the generated responses from different Video-LLMs for MV-QA.

	<b>Who was the defender on the scored basket?</b> A. Johnson B. Jordan C. Anthony D. Smith Among the four options A, B, C, D above, the one closest to the correct answer is:	
	<b>Video-ChatGPT:</b> The defender on the scored basket was Anthony.	✓
	<b>Valley:</b> A. John Shotton	✗
	<b>Other:</b> The answer is not clear from the given information. However, it is possible that the answer is either A or B, as they are the only players mentioned in the sentence.	✗
	<b>Panda-GPT:</b> The correct answer is none of the options above. The image shows a basketball game in progress, but it does not provide any information about the number of offensive attempts shown in highlights and their results.	✗
	<b>mPLUG-Owl:</b> The correct answer is A. Johnson. The Miami Heat's Dwyane Wade made the scored basket against the Chicago Bulls, and Johnson was the defender on him when the shot was made.	✗
	<b>Video-LLaMA:</b> A. Johnson, since he was the player defending the scored basket.	✗
	<b>Video-Char:</b> As there is no clear information in the video about the defender on the scored basket, it is difficult to determine who was the defender. However, the closest possible option to the correct answer might be option C - LeBron James.....	✗
	<b>Chat-LLaV:</b> The defender on the scored basket was A. Johnson.	✗

	<b>Which game is this?</b> A. 2014 NBA All-Star Game. B. 2014 NBA Finals G6. C. 2014 NBA Western Conference First Round G6. D. 2014 NBA Eastern Conference Finals G6. Among the four options A, B, C, D above, the one closest to the correct answer is:	
	<b>Video-ChatGPT:</b> The game shown in the video is the 2014 NBA Western Conference First Round G6.	✓
	<b>Valley:</b> "100"	✗
	<b>Other:</b> The correct answer appears to be option C, which states that the man is playing basketball in a stadium.	✓
	<b>Panda-GPT:</b> The closest option to the correct answer is A. 2014 NBA All-Star Game. The image shows a basketball game with a crowd of people watching, and the players are wearing their respective team uniforms. The other options .....	✗
	<b>mPLUG-Owl:</b> The image depicts a basketball game, which is likely an NBA All-Star Game, given the presence of multiple NBA players, the court setup, and the overall atmosphere of the event. However, it is not a game of the NBA Finals, .....	✗
	<b>Video-LLaMA:</b> The video is a 2014 NBA All-Star Game.	✗
	<b>Video-Char:</b> The basketball game depicted in the video is likely the 2014 NBA Finals Game 6, as the players and the environment are consistent with a professional basketball game. Specifically, the basketball players are wearing .....	✗
	<b>Chat-LLaV:</b> The correct answer is D. 2014 NBA Eastern Conference Finals G6.	✗

Figure 18: The illustrative sample of the generated responses from different Video-LLMs for NBA-QA.

Model	Response	Correct
Video-ChatGPT	There is one illegal act in the video, which is a red car speeding down the road.	✗
Valley	There are four illegal acts in the video.	✗
Qwen	I am not certain about the exact number of illegal acts, but it is mentioned that there are multiple instances of cars running the red light in the video.	✗
Panda-GPT	The correct answer is none. There are no illegal acts in the video.	✗
mPLUG-Owl	The one closest to the correct answer is B: Three illegal acts. The video shows...	✗
Video-LLaMA	The correct answer is option A. There are three illegal acts in the video - one is driving without a license, another is using a handheld device, and a third is crossing...	✗
Video-Chat	Based on the given video, there are no illegal acts visible in the video. There are cars driving on a road, but no cars are speeding, driving on the wrong side of the road, or violating...	✗
Chat-UaIvF	D. An illegal act.	✗

Model	Response	Correct
Video-ChatGPT	The driving behavior of car A is incorrect when driving on the highway.	✓
Valley	The driving behavior of car A is correct when the motor vehicle is driving on the highway.	✗
Qwen	It is not clear from the information provided whether the behavior of Car A is correct or not.	✗
Panda-GPT	In the video, car A is driving on a highway and appears to be changing lanes. However, the correct answer cannot be determined solely based on the video. It is essential to consider...	✗
mPLUG-Owl	As shown in the video, car A is driving in the correct lane and has the appropriate speed for a highway, which is a correct driving behavior on a highway.	✗
Video-LLaMA	The correct answer is option A. There are three illegal acts in the video - one is driving without a license, another is using a handheld device, and a third is crossing...	✗
Video-Chat	Based on the video, there are several cars driving on the highway, including a yellow car and a white car. From the context, it is difficult to determine whether the driving behavior of car...	✗
Chat-UaIvF	B. error. The video shows that the car A is driving on the left lane of the highway, but the driving behavior of car A is not correct when the motor vehicle is driving on the highway.	✓

Figure 19: The illustrative sample of the generated responses from different Video-LLMs for Driver's License Examination.

Model	Response	Correct
Video-ChatGPT	Yes, there was an accident in the video. A man was hit by a car while crossing the street.	✓
Valley	Yes, there was an accident in the video.	✓
Qwen	There is no mention of an actual accident happening, but the man does swerve to avoid something on the road.	✗
Panda-GPT	Yes, there was an accident in the video. The accident involved a car and a motorcycle, and it occurred on a busy city street. The video shows a car and a motorcycle colliding, and the motorcycle rider is seen lying on the ground. ....	✓
mPLUG-Owl	Yes, there was an accident in the video.	✓
Video-LLaMA	Yes, there was an accident in the video. The cars in the video were involved in a minor accident.	✓
Video-Chat	Based on the video, it is not clear if there was an accident in the video. While there is a motorcycle on the ground, it is not clear if it is the cause of any accidents or incidents. From the video, it appears that there are some .....	✗
Chat-UaIvF	No, there was no accident in the video. The video shows a car following a man on a motorcycle, and the car...	✗

Model	Response	Correct
Video-ChatGPT	No.	✓
Valley	Yes, there were traffic signals or signs present in the video.	✗
Qwen	Yes, there were traffic signals and signs present in the video.	✗
Panda-GPT	The image shows a nighttime view of a city street with traffic lights and signs. There are multiple traffic lights visible, indicating that the street is well-regulated for traffic flow. However, there are no visible signs in the image. ....	✗
mPLUG-Owl	No, there were no traffic signals or signs present in the nighttime driving scene in the video.	✓
Video-LLaMA	Yes, there was an accident in the video. The cars in the video were involved in a minor accident.	✗
Video-Chat	Based on the video, it is not clear if there was an accident in the video. While there is a motorcycle on the ground, it is not clear if it is the cause of any accidents or incidents. From the video, it appears that there are .....	✗
Chat-UaIvF	B. Yes, there were traffic signals and signs present in the video.	✗

Figure 20: The illustrative sample of the generated responses from different Video-LLMs for Driving Decision-Making.

Model	Response	Correct
Video-ChatGPT	The correct answer is Pillow.	✓
Valley	The correct answer is option B, which is a cushion.	✗
Qwen	The cushion is located at the 12:00 position on the circular couch.	✗
Panda-GPT	There is no object on top of the couch that is at 1 o'clock. The couch is empty.	✗
mPLUG-Owl	The correct answer is A - a blanket is on top of the couch at 1 o'clock.	✗
Video-LLaMA	A. Blanket. The blanket is the item found on top of the couch that is at 1 o'clock.	✗
Video-Chat	It appears that there is a pillow on top of the couch at 1 o'clock. You can see that it is lying on the couch in the corner near the arm of the couch.	✓
Chat-UaIvF	The correct answer is Pillow.	✓

Model	Response	Correct
Video-ChatGPT	To see the mirror in the bathroom, you should look up.	✓
Valley	A. Look up.	✗
Qwen	The direction you should look to find the mirror is up, as shown in the video.	✓
Panda-GPT	When trying to see the mirror in the bathroom, you should look down. In the image, the mirror is located below the sink, which is why looking up would not be effective in seeing the mirror. ....	✗
mPLUG-Owl	The correct answer is A: look up. In the bathroom, you should look up to the mirror to see your reflection.	✗
Video-LLaMA	A: look up. If you are in a restroom and you need to see your own face, it would be best to look up.	✓
Video-Chat	Looking up is the correct way to see the mirror in the bathroom. This is because the mirror is mounted on the wall above the sink and you would need to look up to see it. Looking down or looking left or right would not .....	✓
Chat-UaIvF	The correct answer is B.	✗

Figure 21: The illustrative sample of the generated responses from different Video-LLMs for SQA3D.