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

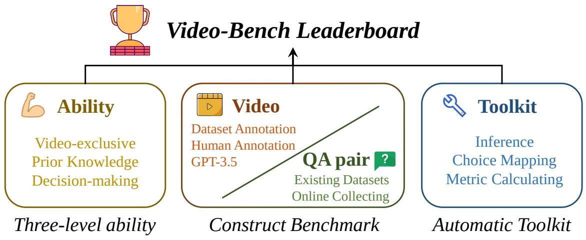


Figure 1: The illustrative pipeline for the intuition, construction and application of *Video-Bench*.

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 comprehensive 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), MSRVTT (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 Reasoning (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 pre-defined choices. SEED-Bench (Li et al., 2023b) introduces a series of temporal understanding tasks and establishes an automatic filtering. LVLM-eHub (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.

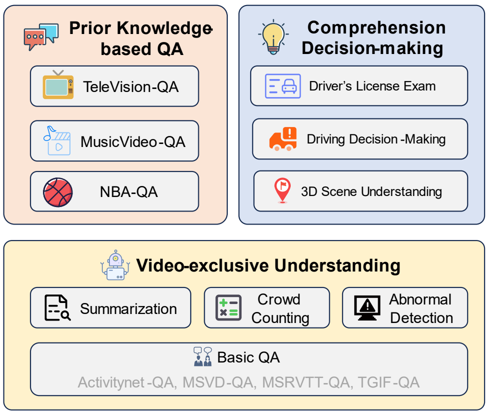
3 Video-Bench

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.,



(A) Three-level Ability Benchmark

Rank	Model	Avg. Acc			
1	Video-ChatGPT	**38.52 (+10.06)**	**39.82 (+13.97)**	29.24 (+4.77)	**46.50 (+11.45)**
2	Panda-GPT	37.52 (+9.07)	37.54 (+11.69)	31.99 (+7.51)	43.05 (+8.01)
3	Otter	37.47 (+9.02)	37.52 (+11.67)	**32.99 (+8.52)**	41.90 (+6.86)
4	VideoChat	35.41 (+6.96)	34.12 (+8.28)	29.61 (+5.14)	42.50 (+7.46)
5	Valley	33.96 (+5.50)	28.39 (+2.54)	29.21 (+4.74)	44.27 (+9.22)
6	mPLUG-Owl	33.15 (+4.69)	33.17 (+7.32)	26.40 (+1.93)	39.88 (+4.83)
7	Chat-UniVi	33.05 (+4.60)	33.08 (+7.23)	27.24 (+2.77)	38.85 (+3.80)
8	Video-LLaMA	32.83 (+4.38)	32.48 (+6.64)	27.80 (+3.33)	38.21 (+3.17)
	Random	28.45	25.85	24.47	35.04

(B) Ranking Table of Video-LLMs

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 MSRVTT-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.

(A) Video-exclusive Understanding

(B) Prior Knowledge-based Question-Answering

(C) Comprehension and Decision-making

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:

$$\text{Choice} = \arg \max_{i \in \{A,B,C,D,\ldots\}} P\left(\text{Token}_i \mid \text{Prompt}\right). \tag{1}$$

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.

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

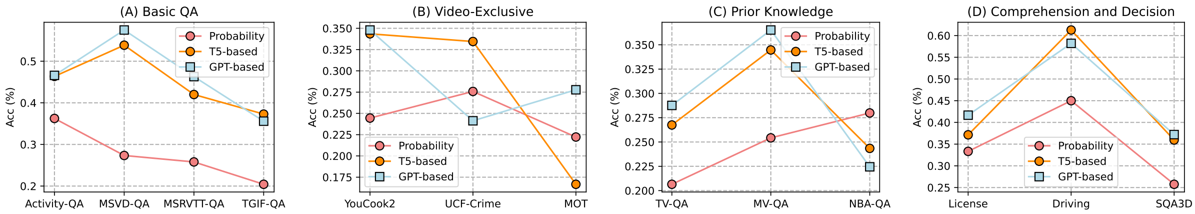


Figure 5: Comparison results of different metrics of Video-ChatGPT (Maaz et al., 2023) on all datasets.

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

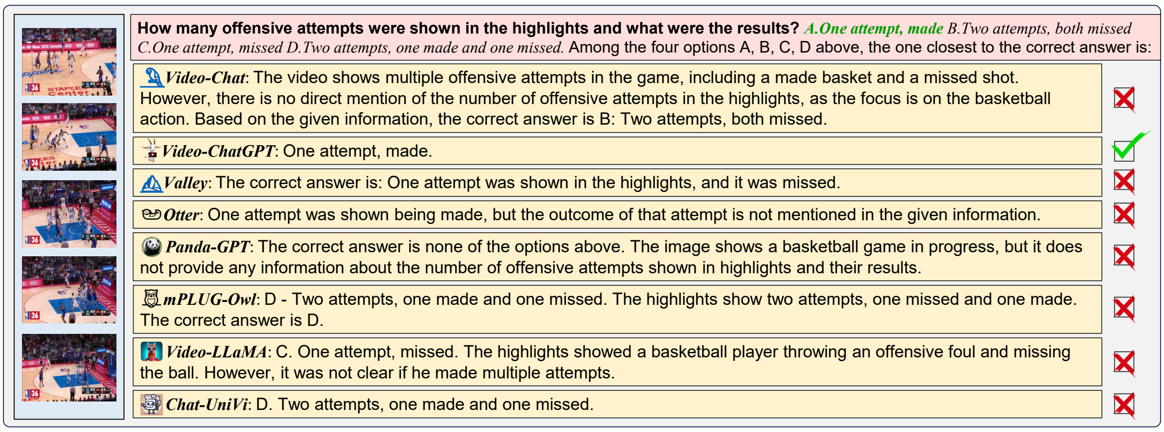


Figure 6: The illustrative sample of the generated responses from different Video-LLMs.

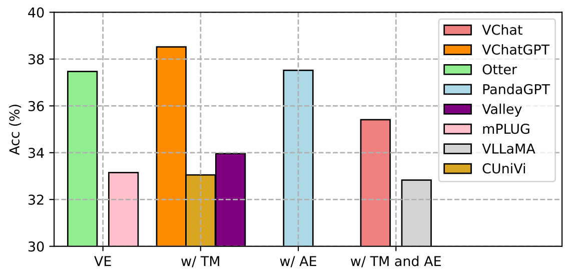


Figure 7: The impact of different module configuration. The w/ TM, w/ AE, and w/ TM and AE means the model contains Temporal Module (TM) or Audio Encoder (AE) or both of them.

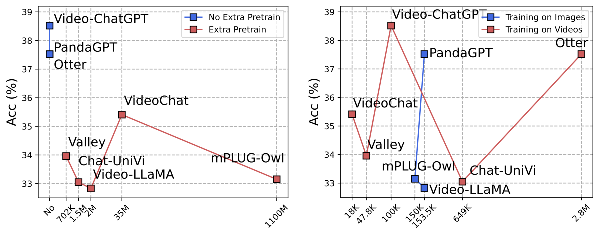


Figure 8: Impact of different datasize of pretrain data (left) or instruction tuning data (right).

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.

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.

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, 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 understanding. *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 visual 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 Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision 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 question 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, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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. *Advances in Neural Information Processing Systems*, 35:9287–9301.

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. *arXiv preprint arXiv:2305.06355*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. 2022. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE Transactions on Image Processing*, 31:6937–6950.

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? *arXiv preprint arXiv:2307.06281*.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.

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. *arXiv preprint arXiv:2306.09093*.

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 *The Eleventh International Conference on Learning Representations*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1975–1981. IEEE.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

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. *Advances in Neural Information Processing Systems*, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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. *OpenAI blog*, 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. *The Journal of Machine Learning Research*, 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. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 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. *arXiv preprint arXiv:1212.0402*.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. 2022. Omnivl: One foundation model for image-language and video-language tasks. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A T5 evaluation

In our answer evaluation benchmark project, we explore two approaches: GPT-based metric and T5-based metric. T5-based metric serves as an auxiliary tool in the evaluation process, offering advantages in terms of cost, deployment, and performance. It provides a cost-effective solution by eliminating the need for ChatGPT API usage and allows for offline deployment on personal servers. As shown in Table 3, T5-based results demonstrate comparable performance to GPT-based in answer evaluation tasks, making it a valuable addition to our benchmark project for reliable and efficient assessment.

B Visualization Samples

In this part, we provide more samples of on all datasets concluded in *Video-Bench*, to illustrate the performance and behaviour of the tested Video-LLMs.

B.1 Video-exclusive Understanding

Activitynet-QA. The results of the Activitynet-QA is shown in Fig. 9. As mentioned in Sec 4, Video-LLMs perform well on these simple questions. The similar results are shown on the remaining three datasets of *Basic QA*.

MSVD-QA. The results of the MSVD-QA is shown in Fig. 10. As part of the *Basic QA*, the performance of Video-LLMs here are overall good.

MSRVTT-QA. The results of the MSRVTT-QA is shown in Fig. 11. The results shows a similar trend of the above.

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-".

(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	46.7	46.35	44.15	46.1	37.65	42.3	44.05	**47.95**
	MSVD-QA*	26.22	43.25	**53.85**	53.35	43.65	30.93	42.6	42.7	44.65
	MSRVTT-QA*	26.50	37.0	42.0	**42.9**	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	**42.30**
Summarization	YouCook2*	25.00	30.0	**34.35**	30.4	31.3	26.95	27.25	28.4	30.65
Abnormal Detection	UCF-Cirme*	16.67	18.62	**33.45**	26.21	24.83	13.45	18.62	22.07	30.86
Crowd Counting	MOT*†	16.67	22.22	16.67	**27.78**	5.56	11.11	11.10	16.67	11.11
Average Score		25.85	34.26	**39.33**	37.89	34.19	27.21	31.34	33.01	37.42

(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	**28.3**	26.75	25.0	27.55	22.15	24.25	25.45	23.05
	MV-QA*	26.15	30.26	34.47	32.41	**34.91**	27.13	29.01	27.84	33.48
	NBA-QA*	27.26	25.36	24.34	**32.51**	26.53	25.36	26.82	28.13	24.49
Average Score		24.47	27.97	28.52	**29.97**	29.66	24.88	26.69	27.14	27.01

(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	25.0	37.14	**55.56**	36.11	30.56	36.11	25.0	50.0
	Decision-Making*	44.21	60.77	61.28	47.44	**62.18**	56.28	53.21	49.49	49.74
3D Scene	SQA3D*	25.00	30.08	**35.95**	27.45	30.25	35.65	32.35	30.5	27.4
Average Score		35.04	41.34	**46.32**	41.07	44.19	42.88	41.45	37.00	40.86

(D) Final Result									
Task	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Average Score	28.45	34.53	**38.06**	36.31	36.02	31.66	33.16	32.38	35.10

(E) Comparison Result of GPT-based									
Task	Random	V-Chat	V-ChatGPT	Otter	PandaGPT	Valley	mPLUG	V-LLaMA	C-UniVi
Video-Exclusive Understanding	25.85	34.12	**39.82**	37.52	37.54	28.39	33.17	32.48	37.87
Prior Knowledge-based QA	24.47	29.61	29.24	**32.99**	31.99	29.21	26.40	27.80	27.43
Comprehension and Decision	35.04	42.50	**46.50**	41.90	43.05	44.27	39.88	38.21	40.64
Average Score	28.45	35.41	**38.52**	37.47	37.52	33.96	33.15	32.83	35.31

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

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

UCF-Crime. The results of the UCF-Crime is shown in Fig. 14. The poor performance illustrates the existing Video-LLMs lack the ability of temporal perception again.

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

B.2 Prior Knowledge-based Question-Answering

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

MV-QA. The results of the MV-QA is shown in Fig. 17. The poor performance may be also caused by the lack of prior knowledge and audio understanding ability.

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

B.3 Comprehension and Decision-Making

Driver's License Examination. The results of the Driver's License Examination is shown in Fig. 19. The poor performance validates the tested Video-LLMs have limited scene understanding and decision-making ability.

Driving Decision-Making. The results of the Driving Decision-Making is shown in Fig. 20, which demonstrates the tested Video-LLMs are difficult to understand the real driving environment.

SQA3D. The results of the SQA3D is shown in Fig. 21. The results show that they can only understand the simple environment and cannot understand the complex spatial relationship.

This is an appendix.

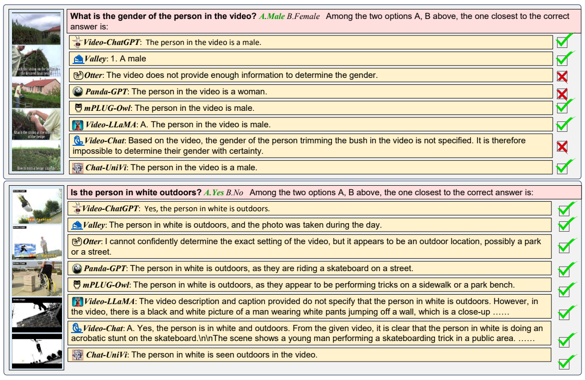


Figure 9: The illustrative sample of the generated responses from different Video-LLMs for Activitynet-QA.

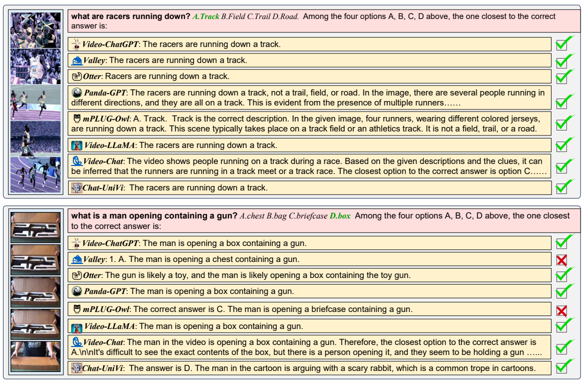


Figure 10: The illustrative sample of the generated responses from different Video-LLMs for MSVD-QA.

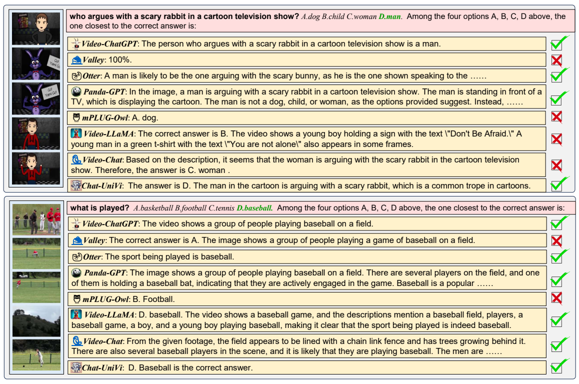


Figure 11: The illustrative sample of the generated responses from different Video-LLMs for MSRVTT-QA.

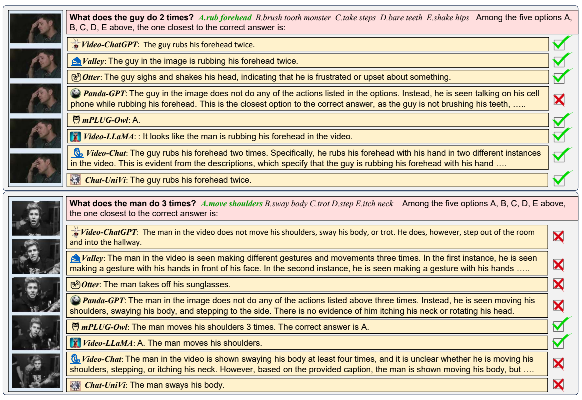


Figure 12: The illustrative sample of the generated responses from different Video-LLMs for TGIF-QA.

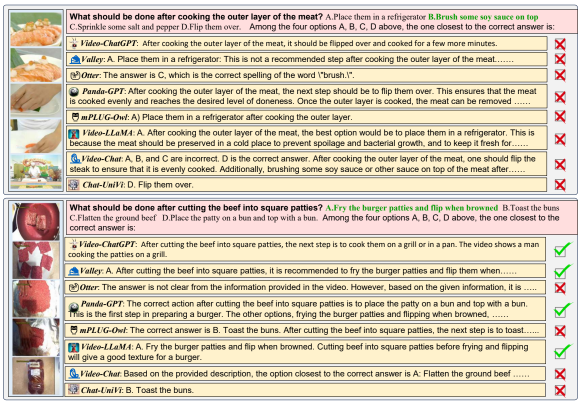


Figure 13: The illustrative sample of the generated responses from different Video-LLMs for YouCook2.

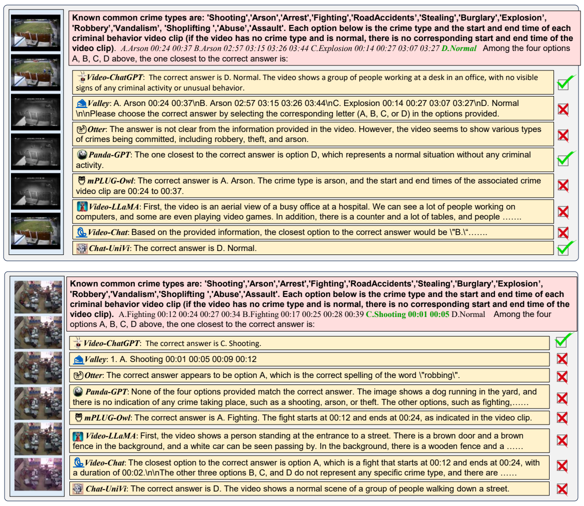


Figure 14: The illustrative sample of the generated responses from different Video-LLMs for UCF-Crime.

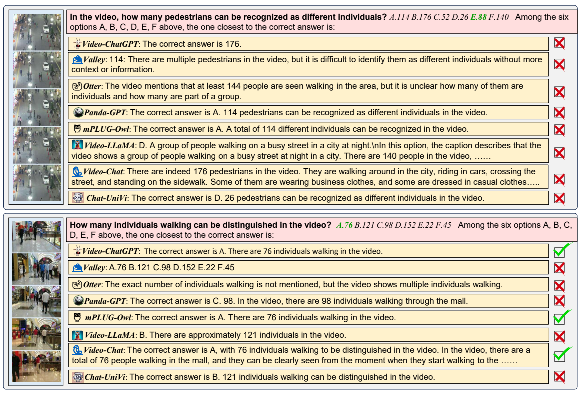


Figure 15: The illustrative sample of the generated responses from different Video-LLMs for MOT.

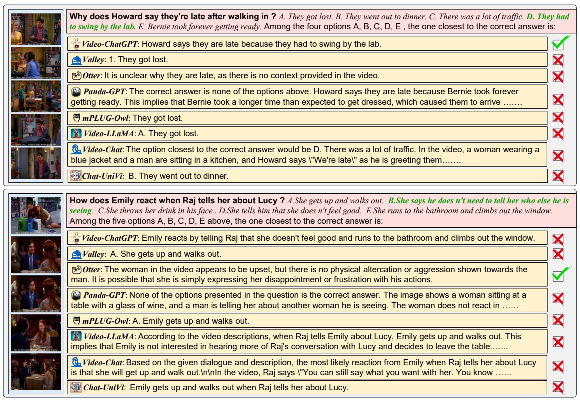


Figure 16: The illustrative sample of the generated responses from different Video-LLMs for TV-QA.

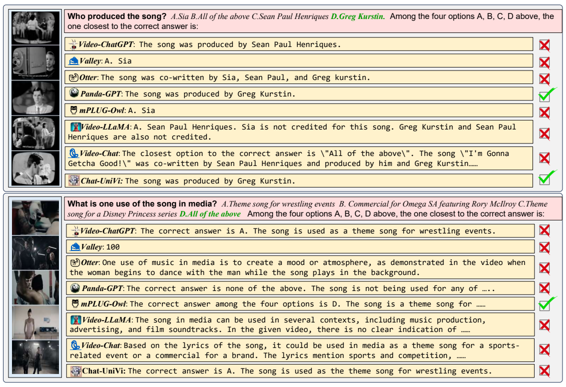


Figure 17: The illustrative sample of the generated responses from different Video-LLMs for MV-QA.

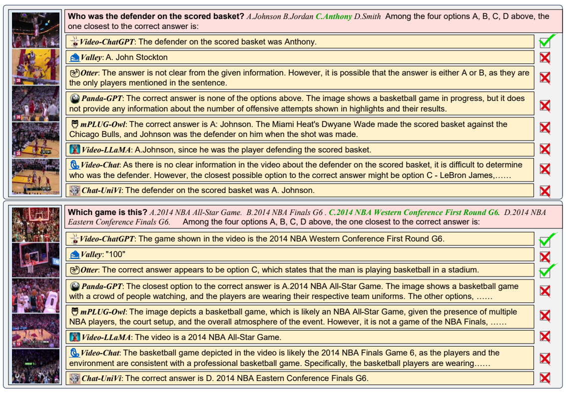


Figure 18: The illustrative sample of the generated responses from different Video-LLMs for NBA-QA.

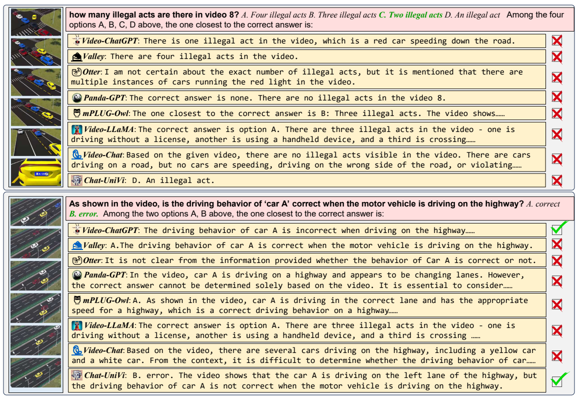


Figure 19: The illustrative sample of the generated responses from different Video-LLMs for Driver's License Examination.

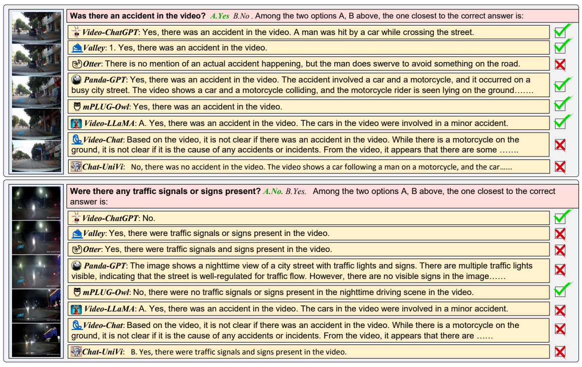


Figure 20: The illustrative sample of the generated responses from different Video-LLMs for Driving Decision-Making.

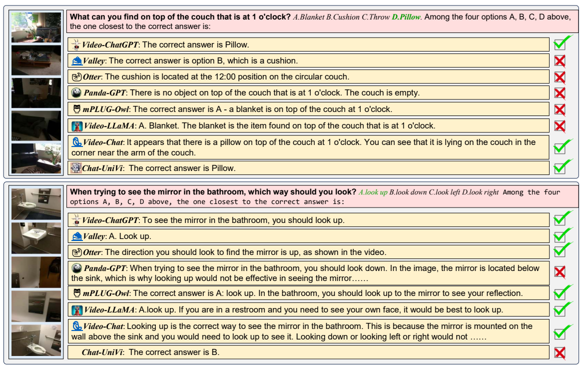


Figure 21: The illustrative sample of the generated responses from different Video-LLMs for SQA3D.