OVBench: How Far is Your Video-LLMs from Real-World Online Video Understanding?

Anonymous CVPR submission

Paper ID 4962

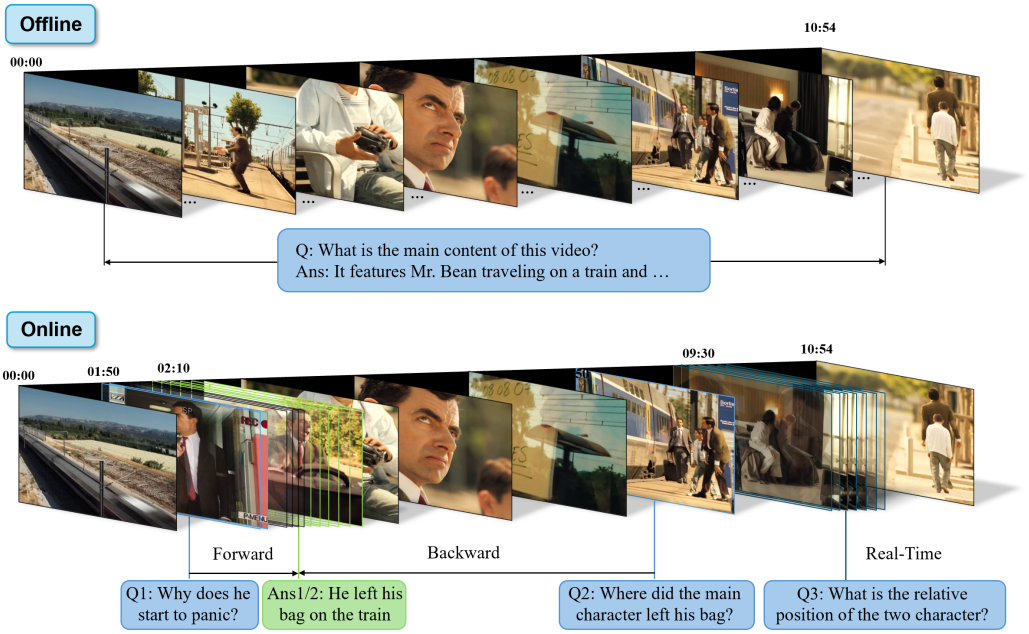

Figure 1. A demonstrative comparison between offline and online video understanding [6]. Offline video understanding focuses on answering questions based on the entirety of a video. In contrast, online video understanding involves posing queries about the context of a video at intermediate points, demanding the ability to trace back past information, perceive ongoing events, and adapt to continuous input.

Abstract

Integrating past information and adapting to continuous video input are pivotal for human-level video understanding. Current benchmarks, however, focus on coarse-grained, video-level question-answering in offline settings, limiting real-time processing and adaptability for practical applications.To this end, we introduce ***OVBench*** *(**O**nline-**V**ideo-**Bench**mark), which assesses online video understanding through three modes: (1)* ***Backward Tracing****, (2)* ***Real-Time Visual Perception****, and (3)* ***Forward Active Responding****. OVBench consists of 12 tasks, comprising about 2,800 meta-annotations with fine-grained, event-level timestamps paired with 858 videos across 10 domains,encompassing egocentric activities, virtual gaming worlds, and cinematic scenes. To minimize bias, we employ automated generation pipelines and human annotation for meticulous curation.We design an effective problem generation and evaluation pipeline based on these high-quality samples and densely query Video-LLMs across the video streaming timeline. Extensive evaluations of nine Video-LLMs reveal that despite rapid advancements and improving performance on traditional benchmarks, existing models struggle with online video understanding. Our compre-*

hensive evaluation reveals that the best-performing models still have a significant gap compared to human agents in online video understanding.We anticipate that OVBench will guide the development of Video-LLMs towards practical real-world applications and inspire future research in online video understanding. Our benchmark and code can be accessed at https://github.com/JoeLeelyf/OVBench.

1. Introduction

Large Vision Language Models (LVLMs) [29, 36, 48, 60] and Video-LLMs [25, 35, 58] have shown remarkable progress, achieving impressive scores on existing benchmarks [12, 13, 26]. Recent works like VideoLLM-online [6] and Flash-VStream [59] have taken initial steps toward J.A.R.V.I.S-like real-world video assistants by combining pre-trained vision encoders [41] and LLMs [10, 47]. However, a critical question remains: *How far are current state-of-the-art models from achieving human-level online video understanding?*

Despite the existence of dozens of evaluation benchmarks in video understanding, there remains a significant domain gap between these evaluations and real-world video understanding tasks. Early evaluations [19, 54, 56] primarily derive from video understanding and retrieval datasets [3, 55], and assess models' capability through coarse-grained QAs, such as *"Q: Who is dancing? A: Man"*. These QAs predominantly focus on short videos with fixed question types and lack temporal indispensability [12]. Subsequent works [13, 26, 62] attempt to address these limitations by extending video temporal length and incorporating more diverse tasks and video sources. E.T.Bench [32] advances this further by exploring inherent temporal information in videos and evaluating fine-grained temporal event detection capabilities. However, all works mentioned above are constrained to offline settings, where models can access all video frames when responding to queries. While models that perform well on these benchmarks demonstrate impressive capabilities in coarse-grained video understanding, a significant gap between their capability scope and the requirements for a real-world assistant or an independent agent still exists.

Recent benchmarks VStream-QA [59] and StreamingBench [28] introduce streaming understanding to the community. While VStream-QA offers an initial exploration with limited video sources selected from Ego4d [23] and MovieNet [17], StreamingBench provides an evaluation of Video-LLMs' in streaming scenarios at a larger scale. However, StreamingBench's three major categories primarily focus on past visual inputs for real-time responses, offering an incomplete view of streaming perception. We propose that effective online video understanding requires simultaneous capabilities to **trace back past information**, **perceive the going-on**, and **adapt to the continuous input** simultaneously. Unlike StreamingBench's emphasis on streaming query format, our approach evaluates Video-LLMs' ability to find temporal visual clues from ongoing input, allowing models to wait for sufficient evidence before responding. We term this approach the **V**ideo **C**hain-**o**f-**T**ime thinking process (Figure 3), analogous to Chain-of-Thought reasoning in LLMs [50].

We introduce **OVBench** (**O**nline-**V**ideo-**Bench**mark) to evaluate Video-LLMs' online video understanding capabilities. The benchmark comprises 858 videos from diverse sources, including curated datasets and web videos, spanning 10 major domains (Sports, Video Games, 3D scanning, *etc.*) with durations ranging from minutes to half an hour. Using a hybrid approach combining semi-automated MLLM generation and human curation, we created 2814 high-quality samples (**Meta-Annotations**) with precise event timestamps. These Meta-Annotations are organized into 12 tasks across three categories: **Backward Tracing**, **Real-Time Visual Perception**, and **Forward Active Responding**, reflecting the human video understanding process illustrated in Fig. 3.

Building on the human-reviewed meta-annotations, we develop an evaluation pipeline that queries Video-LLMs densely along temporal axes to simulate continuous information processing. For **Backward Tracing** and **Real-Time Visual Perception**, we adopt multiple-choice evaluation, converting videos into segments from start to query time to accommodate offline models. With this approach, we explore the potential of explicitly leveraging state-of-the-art models in offline settings for online video understanding. We evaluated 9 Video-LLMs, including proprietary models GPT-4o [36] and Gemini-1.5-Pro [46], alongside six recent open-source MLLMs like QWen2-VL [48] and LLaVA-OneVision [24]. Despite their strong offline performance, these models struggle with online-style queries (e.g., *What is happening now?*), showing a significant gap from human performance.

Further experiments with streaming models like Flash-VStream [59] reveal even larger performance gaps, with Flash-VStream achieving only 47.2% accuracy on Real-Time Visual Perception compared to LLaVA-NeXT-Video's [61] 64.7% with equivalent parameters.

Our proposed **Forward Active Responding** setting introduces the first evaluation framework specifically designed for online video understanding, requiring models to adapt responses to ongoing visual input continuously. We evaluated the offline models using a novel multiple-triggering pipeline that enables comparative assessment of offline models' potential in online video understanding and implements score-based metrics for comprehensive evaluation.

2. Related Works

Video Large Language Models. Video Large Language Models (VLLMs) can process a video by treating it as a sequence of video frames. Projects like VideoChat [25], Video-LLaMA [58], and Video-ChatGPT [35] project the CLIP-ViT [42] embeddings of selected video frames through a Multi-Layer Perceptron (MLP) projector into the LLM embedding space, then concatenate these embeddings with text embeddings for enhanced video understanding. However, the context length of MLLMs limits their effectiveness in understanding long videos [25, 35], as longer videos require more frames and a longer context length. To address this limitation, two major approaches have been developed: compressing video features and selecting critical frames.

In the realm of feature compression, Chat-UniVi [22] merges similar visual tokens through clustering techniques. MovieChat [44] and MA-LLM [16] employ a memory bank to store a fixed number of video tokens by iteratively merging the most similar tokens. ST-LLM [31] and MovieChat [44] reduce video tokens to 32 using a pretrained Q-Former from BLIP2 [11]. LLaMA-VID [27] takes a more radical approach, compressing each frame into a content token and a context token.

On the other hand, frame selection methods aim to identify the most representative frames. VideoStreaming [40] utilizes a small LLM to select critical video clips, while FlashVstream [59] employs a clustering method to choose representative frames for high-resolution processing. LongVU [43] leverages question embeddings to select question-related frames, thereby enhancing video understanding.

Benchmarks for Video Understanding. Traditional video benchmarks, *e.g.,* MSVD-QA [54], MSRVTT-QA [54], and ActivityNet-QA [56], predominantly consist of short videos, typically ranging from 1 to 2 minutes in duration. These datasets are meticulously annotated with corresponding questions and ground truth answers. GPT-4 [36] is employed to assess the accuracy of the answers by comparing them against the provided questions and ground truth responses. However, these benchmarks primarily focus on evaluating short, static video scenes. Hence, new benchmarks designed to test causal and temporal understanding, *e.g.,* NExT-QA [53], TemporalBench [4], and AutoEval-Video [8] are proposed.

To gauge the capabilities of models on long-duration videos, benchmarks like EgoSchema [34] covering over 5,000 egocentric videos with an average length of 3 minutes have been introduced. In contrast, Video-MME [13], LVBench [49], and LongVideoBench [52] feature videos spanning from 20 minutes to over an hour, evaluating a broad spectrum of video understanding capabilities. HourVideo [5] stands out with egocentric videos extending up to 2 hours, accompanied by more than 12,976 multiple-choice questions. Unlike these offline video benchmarks, our proposed **OVBench** is designed to evaluate online, interactive video understanding.

Online Video Understanding. Traditional offline video understanding methodologies primarily focus on accessing entire video sequences to facilitate prediction tasks. Conversely, online video understanding demands models to process video streams sequentially, making decisions based on current and past information. This approach is particularly well-suited for scenarios where future data is unavailable, such as in embodied intelligence, autonomous driving, and augmented reality applications. Among online video understanding methods, FlashVStream [59] employs a clustering method to select representative frames, enabling MLLMs for real-time interactions. LIVE [6] introduces a comprehensive framework for learning in video streams, which includes a training objective, data generation schema, and an inference pipeline tailored for online video understanding.

3. OVBench

In this section, we present the construction process of our OVBench. We start with a detailed introduction to the three different modes of online video understanding, followed by a comprehensive description of the data collection and annotation procedures. A statistical report of our proposed benchmark is displayed at the end of this section.

3.1. Online Video Understanding Mode Taxonomy

Online video understanding aims to equip real-world, always-on agents with the ability to receive and process video inputs continuously, which closely mimics the human visual perception process. We categorize online video understanding into three distinct problem-solving modes: **(1) Backward Tracing**, **(2) Real-Time Visual Perception**, and **(3) Forward Active Responding**. Given a user-provided text query Q_{t_0} at the current time t_0 and a streaming video input $X_{(-\infty,+\infty)}$, these modes are formally defined as follows:

1. **Backward Tracing**:

$$R_{t_0} = P(Q_{t_0}, X_{(-\infty,-T]})$$

2. **Real-Time Visual Perception**:

$$R_{t_0} = P(Q_{t_0}, X_{(-T,t_0]})$$

3. **Forward Active Responding**:

$$R_{(t_0,+\infty]} = P(Q_{t_0}, X_{(t_0,+\infty)})$$

in which T represents a threshold that defines the boundary for recent times, and R denotes the model's response. The first two modes, *Backward Tracing* and *Real-Time Visual Perception*, involve collecting visual information from

past and current timeframes respectively, and are expected to give immediate responses. In contrast, *Forward Active Responding* requires the model to withhold a response until sufficient future information becomes available to ensure a confident answer. Based on these distinctions, we have meticulously designed tasks tailored to each mode to effectively evaluate the performance of Video-LLMs across these diverse capabilities.

3.1.1. Backward Tracing

Memory, particularly long-term memory, is a crucial aspect of human intelligence. In video understanding systems, this capability involves recalling and reasoning about past events. We focus on the following three tasks to evaluate this capability:

1. **[EPM] Episodic Memory:** Backtrack and retrieve key moments from past video inputs.
2. **[ASI] Action Sequence Identification:** Identify the correct sequence of human actions in the video streams.
3. **[HLD] Hallucination Detection:** Ask questions irrelevant to existing video inputs.

3.1.2. Real-Time Visual Perception

Accurate real-time perception of visual content is crucial, as actions undertaken in the present shape future outcomes. In various real-world scenarios, immediate and precise understanding of ongoing visual inputs is essential. We propose six critical categories that constitute the foundational capabilities for effective real-time visual perception:

1. **[STU] Spatial Understanding.** Reason over the spatial relationships between objects occurring in nearby frames.
2. **[OJR] Object Recognition.** Recognize the objects appearing in the current frames.
3. **[ATR] Attribute Recognition.** Identify the characteristics or properties of objects, such as color, texture, and size.
4. **[ACR] Action Recognition.** Recognize and interpret the actions being performed by individuals in the current frame.
5. **[OCR] Optical Character Recognition.** Recognize and interpret characters that appear within the frame.
6. **[FTP] Future Prediction.** Forecast the most probable subsequent phase of the current scene, including changes in object states, actions, and other dynamic elements.

3.1.3. Forward Active Responding

Transitioning from passive reception to active perception is essential for advanced video understanding systems. Existing benchmarks primarily focus on the aforementioned two understanding modes, where Video-LLMs are required to respond immediately based on available information. In contrast, we introduce the *Forward Active Responding* mode, which allows the model to adjust its responses based on forthcoming visual inputs. We devise four task dimensions to evaluate the models' active responding abilities:

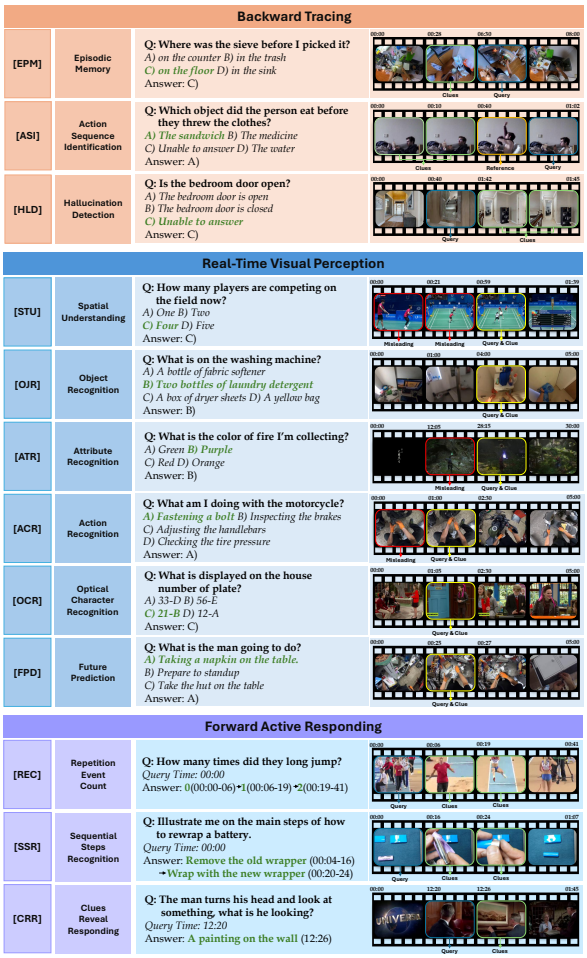

Figure 2. **Examples of each task in OVBench.** The 14 tasks are categorized into three different kinds of perceiving modes in online video understanding: **Backward Tracing**, **Real-Time Visual Perception**, and **Forward Active Responding**.

1. **[REC] Repetition Event Count.** Respond when a repetitive event occurs again, including both high-frequency repetitive actions over short durations and semantically long-term repetitive occurrences of certain events.
2. **[SSR] Sequential Steps Recognition.** Respond when a certain procedure or sequence of actions has transitioned to another stage.
3. **[CRR] Clues Reveal Responding.** Delay responding until sufficient information or clues are provided.

3.2. Benchmark Construction

Under the taxonomy guidelines above, we make our first step by collecting video data and annotations from existing datasets and crawling data from the web to increase diver-

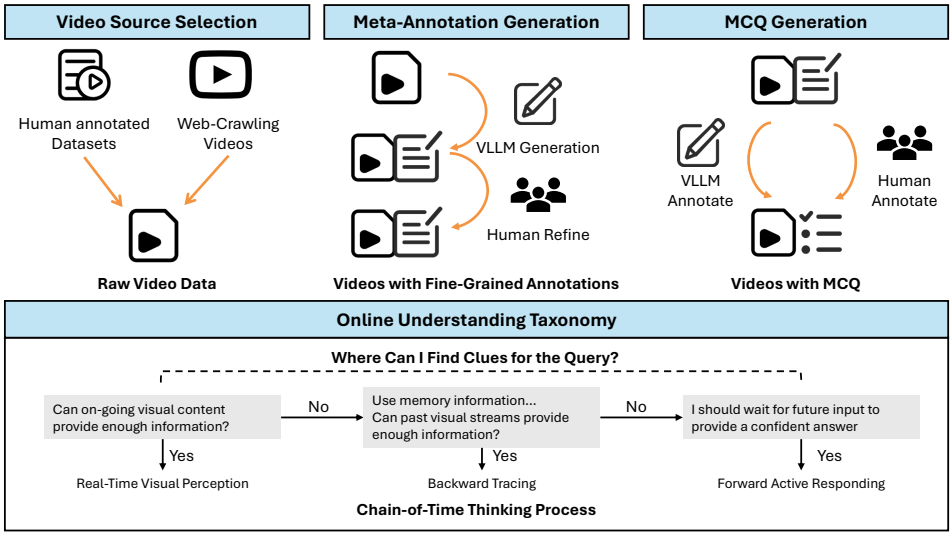

Figure 3. **Generation pipeline of OVBench.** Within public annotations, data is carefully filtered and relevant multiple-choice QAs are auto-generated. The effective system prompt and efficient answer prompt are employed to guide MLLMs toward precise outputs. The Video-LLMs we use to annotate videos are GPT-4o and Gemini-1.5 Pro.

sity. As our proposed evaluation pipeline highly relies on the accurate timestamp annotations of the referred events in the constructed prompt, the scarcity of event-level timestamps in existing datasets [51][33][39] promotes the design of our highly efficient meta-data generation pipeline 3. Raw annotations with coarse timestamps are then refined by humans to ensure accuracy. Our final questions and options for evaluation are constructed using our rule-based pipeline based on these human-refined meta-annotations. All QA samples undergo manual inspection before being included in the final test set.

3.2.1. Video and Annotation Collection

Video Source Selection. We follow existing benchmarks [32][26] by exploiting high-quality customized video datasets, and enrich our diversity by utilizing self-crawling videos from different domains. **(1) Human-annotated Video Dataset.** Our main consideration for utilizing organized datasets is to alleviate the labor-intensive source video collection process. Specifically, we include QA-Ego4D[2] and OpenEQA[1] for the **[EPM]** task, STAR[51], YouCook2[63], CrossTask[65], HiREST[57], and COIN[45] for the **[ASI]** task, Perception-Test[39] and Thumos[21][14] for the **[REC]** task, COIN[45] for the **[SSR]** task, MovieNet[17] for the **[CRR]** task, and Ego4D[23] for tasks under *Real-Time Visual Perception*. All samples are selected from val or test sets to avoid potential data leakage. **(2) Web-crawling Videos.** To further extend the diversity of our benchmark, we follow the existing practice [12][28] of crawling source videos from YouTube.

Meta-Annotations Collection. We employ three approaches to collect our meta-annotations which contain event-level timestamps: **(1) Existing Annotation Repurposing.** For human-annotated datasets with accurate event-level timestamps [2][45][23], we explicitly take advantage of these labels and reconstruct them to our final prompt. **(3) Semi-Automatic Generation.** For datasets that provide video-level QA pairs without complete temporal localization, including [33][51][39][21][14], we prompt temporal-sensitive Video-LLMs like Gemini-1.5[46] to provide coarse-grain timestamps which fit the event referred in question and answer. For tasks under the *Real-Time Visual Perception* scenario, timestamps are given during our automatic QA construction process, which will be illustrated in 3.2.2. We then perform meticulously inspect all collected source videos and the corresponding meta-annotations to ensure precision.

3.2.2. Prompt Generation

Question and Answer Generation. Besides carefully selecting QA pairs from existing datasets to fit into our proposed tasks, we also adopt a highly efficient automatic question and answer generation pipeline, particularly for the *Real-Time Visual Perception* scenario. We randomly sample short clips from original long-form videos and then leverage GPT-4o[18] to select potential candidates and construct questions and corresponding answers using human-refined prompts. Human-proposed questions are also adopted as a part of these tasks to alleviate possible LLM preferences. For the novel **[CRR]** task, even the strongest Video-LLMs/MLLMs like Geimini-1.5-Pro struggle to construct desired problems. Volunteers are then recruited to provide QA pairs under our guidance.

Options Generation and Selection. We adopt multiple-choice questions as testing forms for *Backward Tracing* and *Real-Time Visual Perception* scenarios. However, as re-

Figure 4. **Left** Queries Temporal Distribution in OVBench. **Center** Linguistic Characteristics of Text Queries. **Right** Video category distribution of OVBench.

vealed in [7], the naively designed options of a multi-choice form query can cause information leakage about answers. We propose to generate options using a carefully designed rule-based and visually grounded transformation of correct answers, bringing misleading information from original videos to increase difficulty. Specifically, we prompt Video-LLMs with original QA pairs and corresponding video clips to generate visual-related options. A careful human review is then conducted to further ensure the options' effectiveness. All options are shuffled after human review to avoid potential preference bias.

Prompting Offline-Models for Simulated Online Understanding. With the significant performance gap between main-streaming powerful offline Video-LLMs [46][37][48] and existing online models [6][59], one natural question is made: *Is it effective to prompt offline models directly for online video understanding?* For the *Real-Time Visual Perception* setting, we make human curation to the original question to include implies about the real-time query scenarios, for example, by using sentence patterns like *What is/What am I* or containing words like *Now/Currently*. We made another intuitive attempt to prompt offline models to solve tasks under our novel *Forward Active Responding* scenario, which asks for a continuous adapting capability. Specifically, we devise a multiple-triggering densely query and evaluation pipeline, allowing the model to decide whether existing information has provided enough clues for answering the user's query.

3.3. Datasets Statistics

OVBench consists of 858 videos spanning 10 major domains, including Sports, Video Games, and 3D Scanning, among others. The video durations range from a few minutes to half an hour, with the average query timepoint being 428.89 seconds. Figure 4 **Left** illustrates the duration distribution of the queries within OVBench. The benchmark includes 2,814 question-answer (QA) pairs, featuring a large number of multiple-choice questions and a smaller set of open-ended questions. The number of options for the multiple-choice questions varies between 2 and 5, rather than being fixed at four. The distribution of video category is visualized in Figure 4 **Right**.

4. Experiments

This section presents comprehensive experiments and in-depth analyses of OVBench.

4.1. Models and Evaluation Strategies

We evaluate four existing types of models: (1) Offline Multimodal Models, including GPT-4o [38], Gemini-1.5-Pro [46], Qwen2-VL [48], LLaVA-NeXT-Video [30], LLaVA-OneVision [29], InternVL-V2 [9] and LongVU [43], (2) Online Multimodal Models, including Flash-VStream-7B [59] and Videollm-Online[6] (3) Blind LLMs, including GPT-4-turbo [36]. (4) Human Agents. To ensure a fair comparison of model performance, we adhere to the principle of consistency by maintaining the same number of frames or frames per second (fps) across all models.

Considering the limitations on input video length for existing offline Video-LLMs, we adopt specialized video input methods tailored to such models. Specifically, we segment the video into clips based on the timestamps of the questions.For instance, for a question Q_i posed at timestamp t_i, we extract the video clip Video$[0 : t_i]$ as the visual input. This approach simulates a streaming question-answering scenario in online video understanding.

4.2. Main Results

Table 1 reports the performance of eleven models under different settings on OVBench, including the *Real-Time Visual Perception*, *Backward Tracing*, and *Forward Active Responding*. Our evaluation brings several important findings, as follows:

Offline Video-LLMs' video understanding capabilities can be effectively transferred to real-time video understanding. The results demonstrate that offline Video-LLMs, despite being designed for offline processing, perform competitively in *Real-Time Visual Perception* tasks. This suggests that the advanced video comprehension abilities developed in offline settings are transferable and can enhance performance in certain online scenarios, thereby

Model	# Frames	Real-Time Visual Perception							Backward Tracing				Forward Active Responding				Overall Avg.
		OCR	ACR	ATR	STU	FTP	OJR	Avg.	EPM	ASI	HLD	Avg.	REC	SSR	CRR	Avg.	Overall Avg.
Human																	
Human Agents	-	93.96	92.57	94.83	92.70	91.09	94.02	93.20	92.59	93.02	91.37	92.33	95.48	89.67	93.56	92.90	92.81
Blind LLMs																	
GPT-4-turbo	-	28.86	24.77	25.67	33.76	27.72	26.63	27.90	42.76	48.65	70.05	53.82	-	-	-	-	-
Proprietary Multimodal Models-Offline																	
Gemini 1.5 Pro	1fps	**87.23**	65.12	**78.48**	**57.18**	68.32	**66.66**	**70.83**	**57.58**	71.62	**58.34**	**62.51**	**37.48**	70.37	**73.61**	**60.49**	**66.67**
GPT-4o	64	73.15	**68.81**	75.78	56.18	**77.23**	62.51	68.94	55.22	**75.68**	40.65	57.15	29.45	**77.21**	71.53	59.40	63.05
Open-source Multimodal Models-Offline																	
Qwen2-VL-72B	64	**77.18**	**61.47**	**76.72**	**51.68**	**76.24**	64.62	**67.98**	52.19	**75.68**	**68.02**	**65.30**	**36.41**	**75.50**	63.19	**58.37**	**63.88**
Qwen2-VL-7B	64	70.74	52.29	66.38	43.26	71.29	63.23	61.20	47.14	65.54	38.57	50.42	30.92	60.97	51.39	47.76	53.13
InternVL-V2-8B	64	73.15	58.72	66.15	50.00	70.30	60.53	63.81	48.48	60.14	25.88	44.83	31.32	60.11	53.47	48.30	52.31
LLaVA-NeXT-Video-7B	64	74.50	60.55	68.11	**51.68**	72.28	**64.68**	65.63	55.89	65.54	6.09	42.51	30.52	69.80	45.14	48.49	52.21
LLaVA-OneVision-7B	64	70.74	59.63	73.28	47.19	71.29	63.64	64.63	**56.23**	56.08	16.74	43.02	27.58	63.82	38.19	43.20	50.28
LongVU-7B	1fps	55.70	54.13	62.07	42.98	68.32	60.14	57.89	50.17	65.31	3.55	39.68	19.14	60.11	**63.89**	47.71	48.43
Open-source Multimodal Models-Online																	
Flash-VStream-7B	1fps	40.94	51.38	44.90	38.52	62.38	44.97	47.18	42.76	38.51	3.05	28.11	15.8	40.15	65.97	40.64	38.64
VideoLLM-online-8B	2fps	8.05	23.85	12.07	14.04	45.54	21.20	20.79	22.22	18.80	12.18	17.73	-	-	-	-	-

Table 1. **Detailed evaluation results on OVBench.** † refers to using *"low"* resolution.To enhance the challenge of the questions by increasing the time interval between the question and the clues, the question time for **[EPM]** and **[ASI]** in the table is uniformly placed at the end of the video. **[HLD]** are evaluated in the standard manner following the streaming mode.For **Forward Active Responding**, accuracy-based evaluation metrics are utilized in this table.More high-quality **[ASI]** tasks are supplemented, further enhancing differentiation.

partially bridging the gap between offline and online video understanding.

Current Video-LLMs lack temporal prioritization when handling VQA tasks. Existing Video-LLMs do not prioritize real-time temporal information when answering questions, leading to an inability to accurately locate the correct scene when multiple misleading scenes matching the question appear in the video stream, as shown in Fig2. Even the best current proprietary models achieve only 57.18% and 68.81% on **[STU]** and **[ACR]** tasks, respectively, which represents a significant gap compared to Human Agents.

A powerful LLM backbone is the key to achieving high-performance video understanding. As shown in Table 1, the Qwen2-VL-72B model significantly outperforms its smaller counterpart, Qwen2-VL-7B, across all evaluated metrics. The larger model's superior architecture allows it to excel in *Real-Time Visual Perception* with an average score of 67.98%, compared to 61.20% for the 7B variant. Additionally, Qwen2-VL-72B demonstrates enhanced capability in *Backward Tracing* tasks, achieving a score of 68.02%, notably higher than the 54.15% scored by Qwen2-VL-7B. These results highlight the importance of a robust and powerful LLM backbone in effectively processing and understanding complex video data, underscoring the need for more powerful architectures to achieve high performance in video understanding tasks.

Hallucinations are prevalent in Video-LLMs. The **[HLD]** in Table 1 measures hallucinations in Video-LLMs [20], indicating that hallucinations are a significant issue, particularly in open-source and online models. Proprietary models like Gemini 1.5 Pro perform better in managing hallucinations, yet there remains a notable gap compared to human performance(58.34% vs. 91.37%). This problem arises due to the models' inability to fully comprehend complex visual and temporal contexts, leading to errors in interpretation and response. Addressing hallucinations is crucial for improving the reliability and accuracy of Video-LLMs in real-world applications.

4.3. Comparison between online Video-LLMs and offline Video-LLMs

Models like **Gemini 1.5 Pro** and **Qwen2-VL-72B**, representative of offline Video-LLMs, demonstrate strong performance across various tasks, as shown in Fig7. Specifically, Gemini 1.5 Pro achieves the highest average score among these models. This superior performance suggests that offline models, despite not being designed for online or real-time processing, can effectively comprehend and process complex visual information when provided with sufficient computational resources and pre-processing time. Their architectures typically allow for processing the entire video sequence holistically, leveraging global context and detailed temporal information, which enhances their temporal understanding and reasoning capabilities.

In contrast, **Flash-VStream-7B**, representing online Video-LLMs, shows comparatively lower performance in real-time perception tasks compared to offline models. This model is designed to process video in a streaming manner, handling inputs frame by frame with strict latency constraints to achieve real-time responsiveness. The performance gap highlights a potential trade-off between real-time processing capabilities and the depth of visual understanding.

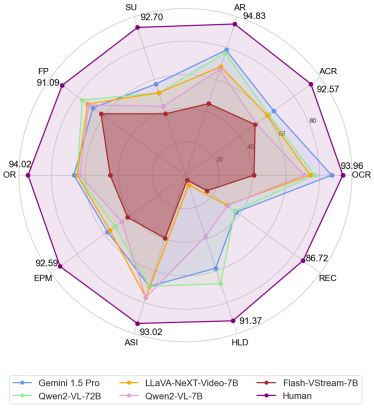

Figure 5. **Performance comparison between online Vedio-LLMs and offline Vedio-LLMs.** The figure illustrates the average scores of different models on the OVbench in real-time visual perception tasks.

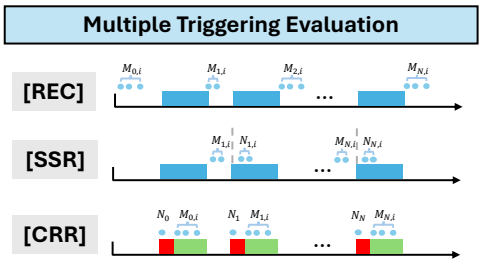

Figure 6. **Multiple triggering evaluation pipeline of prompt offline models for online video understanding.** Offline Video-LLMs are densely queried along the temporal axes to make independent decisions of whether existing visual content provide enough clues for answering.

4.4. Forward Active Responding

We include our evaluation pipeline design for our proposed *Forward Active Responding*. While our high-quality human-annotated queries and clues lay an ideal testbed for future real-world online understanding models, existing naively designed online video models usually collapse in our evaluation process. We made our initial attempts to leverage our multiple-triggering query pipeline to prompt offline VideoLLMs to perform online video understanding thinking schema and further explore their potential in always-on visual perception.

Evaluation Pipeline and Metrics. As illustrated in Fig.6, We propose to query the Video-LLMs densely along the temporal axes, particularly around the interested events'. Our main concerns are twofold: 1) Encourage models' timely finding of the right clues, and 2) Avoid any possible hallucination before the right clue appears. For the

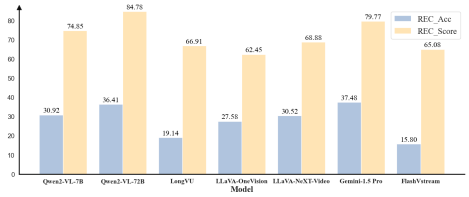

Figure 7. **Performance of models in Repetition Event Count and Clue Reveal Responding task.** Evaluation results are given using the average accuracy and our proposed metrics.

[REC] task, larger counting numbers are awarded. Based on this, we proposed our designed scoring metrics for the three tasks in the *Forward Active Responding*. Given N the total times of events, and *i/j* the corresponding trigger times, our evaluation metrics is formed as below:

1. **Repetition Event Count**:

$$Score = \sum_{i}^{N} e^{n/5} \cdot \sum_{j=0}^{3} acc_j^{\frac{1}{i+1}}$$

2. **Sequential Steps Recognition**:

$$Score = Acc \cdot \sum_{i=0}^{2} p_i \cdot 2^{1-i}$$

3. **Clues Reveal Responding**:

$$Score = Acc \cdot \sum_{i=0}^{3} p_i \cdot 2^{1-i}$$

Offline Models for Online Video Understanding. Despite their promising performance on the *Backward-Tracing* and *Real-Time Visual Perception*, in which the models are given full information for making confident responses, our preliminary results show that even state-of-the-art offline models like Geimini-1.5-Pro, fails to capture the linguistic information of ongoing querying, showing limited understanding of online video content.

5. Conclusion and Future Work

In this work, we introduced OVBench, a comprehensive benchmark designed to assess online video understanding capabilities of Video-LLMs across three critical modes: *Backward Tracing*, *Real-Time Visual Perception*, and *Forward Active Responding*.We anticipate that OVBench will serve as a valuable resource for the research community, guiding the development of Video-LLMs toward practical, real-world applications. By highlighting current limitations and providing a platform for rigorous evaluation, we hope to inspire future research dedicated to advancing online video understanding and achieving human-level comprehension in artificial intelligence systems.

References

[1] Xiaohan Zhang Pranav Putta Sriram Yenamandra Mikael Henaff Sneha Silwal Paul Mcvay Oleksandr Maksymets Sergio Arnaud Karmesh Yadav Qiyang Li Ben Newman Mohit Sharma Vincent Berges Shiqi Zhang Pulkit Agrawal Yonatan Bisk Dhruv Batra Mrinal Kalakrishnan Franziska Meier Chris Paxton Sasha Sax Aravind Rajeswaran Arjun Majumdar, Anurag Ajay. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *CVPR*, 2024. 5, 4

[2] Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1560–1568, 2022. 5, 4

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 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, 2015. 2

[4] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models, 2024. 3

[5] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *arXiv preprint arXiv:2411.04998*, 2024. 3

[6] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *CVPR*, 2024. 1, 2, 3, 6

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 6

[8] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering, 2024. 3

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 6

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 2

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 3

[12] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024. 2, 5

[13] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2, 3

[14] Alex Gorban, Haroon Idrees, Yu-Gang Jiang, A Roshan Zamir, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2015. 5, 4

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 4

[16] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding supplementary material. 3

[17] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 2, 5, 4

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

[19] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 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, 2017. 2

[20] Shaoxiang Chen Jingjing Chen Yu-Gang Jiang Jiacheng Zhang, Yang Jiao. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv: 2409.16597*, 2024. 7

[21] Yu-Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014. 5, 4

[22] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023. 3

[23] Eugene Byrne Zachary Chavis Antonino Furnari Rohit Girdhar Jackson Hamburger Hao Jiang Miao Liu Xingyu Liu et al. Kristen Grauman, Andrew Westbury. Ego4d: Around

the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 2, 5

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[25] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2, 3

[26] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2, 5

[27] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. 2024. 3

[28] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024. 2, 5

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 6

[30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6

[31] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025. 3

[32] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Chang Wen Chen, and Ying Shan. E.t. bench: Towards open-ended event-level video-language understanding. In *Neural Information Processing Systems (NeurIPS)*, 2024. 2, 5, 4

[33] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498, 2024. 5

[34] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. 3

[35] Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*, 2023. 2, 3

[36] OpenAI. Gpt-4 technical report, 2023. Technical report. 2, 3, 6

[37] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o, 2024. 6

[38] OpenAI. Hello gpt-4o, 2024. https://openai.com/index/hello-gpt-4o. 6

[39] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 4

[40] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *arXiv preprint arXiv:2405.16009*, 2024. 3

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3

[43] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 3, 6

[44] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding, 2023. 3

[45] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 5, 4

[46] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2, 5, 6

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[48] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 6

[49] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 3

[50] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny

Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 2

[51] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 5, 4

[52] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. 3

[53] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 3

[54] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 2, 3

[55] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 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, 2016. 2

[56] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 2, 3

[57] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23056–23065, 2023. 5, 4

[58] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2, 3

[59] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams, 2024. 2, 3, 6

[60] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023. 2

[61] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 2

[62] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 2

[63] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 5

[64] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 4

[65] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 5, 4