# Online Video Understanding: OVBench and VideoChat-Online

Zhenpeng Huang[1†], Xinhao Li[1,3†], Jiaqi Li[2†], Jing Wang[1], Xiangyu Zeng[1,3]
Cheng Liang[1], Tao Wu[1], Xi Chen[2], Liang Li[2], Limin Wang[1,3,✉]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]China Mobile Research Institute [3]OpenGVLab, Shanghai AI Laboratory
https://videochat-online.github.io/

## Abstract

*Multimodal Large Language Models (MLLMs) have significantly progressed in offline video understanding. However, applying these models to real-world scenarios, such as autonomous driving and human-computer interaction, presents unique challenges due to the need for real-time processing of continuous online video streams. To this end, this paper presents systematic efforts from three perspectives: evaluation benchmark, model architecture, and training strategy. First, we introduce* ***OVBench****, a comprehensive question-answering benchmark designed to evaluate models' ability to perceive, memorize, and reason within online video contexts. It features 6 core task types across three temporal contexts—past, current, and future—forming 16 subtasks from diverse datasets. Second, we propose a new* ***P****yramid* ***M****emory* ***B****ank (****PMB****) that effectively retains key spatiotemporal information in video streams. Third, we proposed an offline-to-online learning paradigm, designing an interleaved dialogue format for online video data and constructing an instruction-tuning dataset tailored for online video training. This framework led to the development of* ***VideoChat-Online****, a robust and efficient model for online video understanding. Despite the lower computational cost and higher efficiency, VideoChat-Online outperforms existing state-of-the-art offline and online models across popular offline video benchmarks and OVBench, demonstrating the effectiveness of our model architecture and training strategy.*

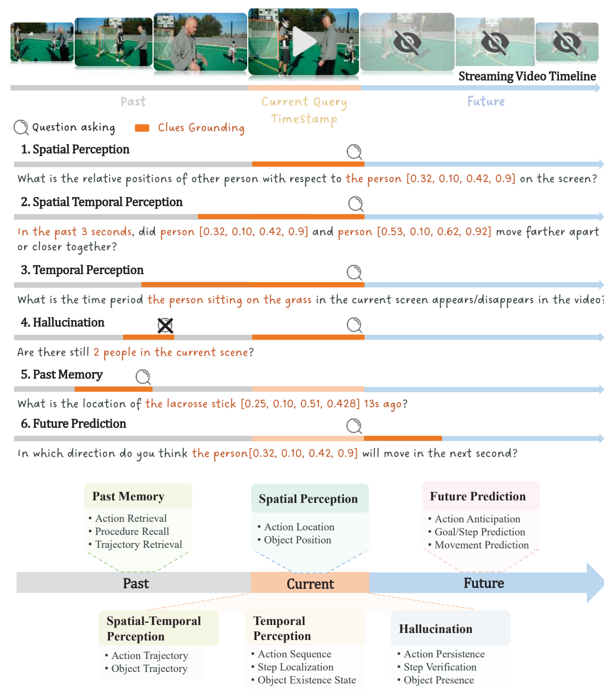


Figure 1. OVBench contains 6 core spatiotemporal understanding tasks in online scenarios, incorporating three primary temporal contexts—past, current, and future. Based on various interaction types, it is expanded into 16 subtasks in total.

## 1. Introduction

With the rapid development of Multimodal Large Language Models (MLLMs) [26, 35, 36, 51, 54] in recent years, these models have demonstrated impressive performance on video understanding benchmarks [27, 31, 37]. These advancements have laid the foundation for exploring real-time, online video scenarios, including autonomous driving, robotic assistants, and surveillance systems. Recent research, such as GPT-4o [36] and VideoLLM-Online [5], Flash-VStream [56], have further investigated online video understanding and model efficiency in streaming scenarios, highlighting the potential of MLLMs in understanding online video streams.

Despite these advances, applying MLLMs to real-world streaming scenarios presents unique challenges. Offline

† Equal contribution. ✉ Corresponding authors.

processing refers to models that analyze entire videos post-capture, rather than responding in real-time as frames are received. The unique characteristics of online video streams are not fully considered by the existing works as follows:

- **Online Temporal Perspective**: Based on the time when a user poses a question, the temporal perspective of online video streams can be distinctly defined as past, current, and future. In contrast to offline videos, it enables a finer temporal perspective (e.g., a few seconds ago, right now).
- **Time-dependent Contexts**: In offline video understanding, answers are derived from all prior frames, typically yielding a unique response. In streaming scenarios, as the temporal context evolves, the answer dynamically changes. Questions like "What is the person doing now?" may receive varying responses over time.
- **Real-time Spatio-temporal Interaction**: Applications like augmented reality (AR) glasses and autonomous driving systems require precise, real-time spatiotemporal interaction with the environment, where immediate responses to the environment (e.g. actions, objects, and events) are essential for functionality and safety.
- **Processing of infinitely long visual information**: Online video streams continuously introduce an infinite influx of new visual information. Therefore, designing online model architectures that can process and retain key information, akin to human cognition, is critical.

Given that most of the current video understanding benchmarks [13, 21, 27, 46, 50, 52, 55, 59] are conducted in offline mode, there is a pressing need to build a benchmark specifically tailored to online video streams, taking into account their unique spatiotemporal characteristics. To tackle these problems, we introduce the **O**nline **V**ideo Understanding **Bench**mark, **OVBench**. This benchmark aims to evaluate a model's capacity to understand and interpret spatiotemporal details in online scenarios. As shown in Fig. 1, we define the temporal context for streaming videos, and based on these temporal contexts, we design 6 task types encompassing a total of 16 subtasks. These tasks are based on seven datasets spanning 6 different domains (Movie, Instructional, Road, Scenes (Outdoor & Indoor), and Open-domain) to ensure a diverse range of task scenarios. To create high-quality annotations, our benchmark employs human annotators who generate $\sim$7,000 high-quality annotations that emphasize spatiotemporal details.

We evaluate leading MLLMs on OVBench, including offline image/video MLLMs (adapted to streaming via sliding windows) and online video MLLMs. Current models exhibit poor online spatiotemporal understanding, with online models lagging significantly behind MLLMs. This motivates us to develop a strong baseline for online video understanding with a novel architecture and training strategy. **(1) New Model Architecture**: Existing architectures [5, 41, 42, 56] struggle with fine-grained spatial details and long-range temporal dependencies as streaming contexts grow. We propose a Pyramid Memory Bank (PMB) to balance spatial and temporal understanding via progressive abstraction. PMB preserves recent-frame details while efficiently abstracting distant-frame information using adaptive frame eviction and resolution scaling, optimizing both comprehension and memory efficiency. **(2) New Training Strategy**: A key limitation of existing MLLMs is the lack of a tailored training strategy. We introduce an offline-to-online learning paradigm, constructing interleaved dialogue-style online video instruction tuning data. This, combined with offline video data, progressively enhances both offline and online video understanding.

Based on the aforementioned design principles, we have developed an efficient 4B-parameter online video MLLM, coined as **VideoChat-Online**, specifically designed for mobile deployment, thereby broadening the potential applications of online video understanding. On OVBench, our model outperforms the open-source offline MLLM Qwen2-VL [45] (7B parameters) by **4.19%** and the online Video MLLM Flash-Vstream [56] by **23.7%**, achieving this with a more efficient architecture. Furthermore, it demonstrates state-of-the-art performance on established offline video benchmarks, highlighting its robustness across both online and offline video understanding. All the models and data are publicly available. We hope that this work's benchmark, dataset, and model will inspire future research on online video understanding.

## 2. Related Work

**Online Video MLLMs.** The advent of large language models (LLMs) [9, 22, 32, 34, 44] has spurred substantial progress in multimodal understanding. Recent multimodal LLMs [8, 16, 17, 25, 26, 28, 30, 40, 45] have exhibited impressive capabilities in offline video comprehension by integrating visual encoders with LLMs. However, these models are inherently challenged in real-time applications due to their limited capacity for efficient streaming video frame compression, leading to increased computational demands and latency with accumulating input frames. Several strategies [10, 23, 41, 42] have been explored to mitigate computational burden through video redundancy reduction. However, most of these models lack a design specifically tailored for online video stream processing. Recent studies have introduced MLLMs specifically designed for online stream understanding. VideoLLM-Online [5] pioneers the development of general-purpose AI assistants for real-time video stream dialogue and multi-task execution. However, its performance is limited by restricted per-frame visual token input due to the lack of effective streaming context compression. VideoLLM-MOD [49] addresses this limitation by incorporating mixture of depth [39] for efficient visual token computation, enabling higher visual input resolu-

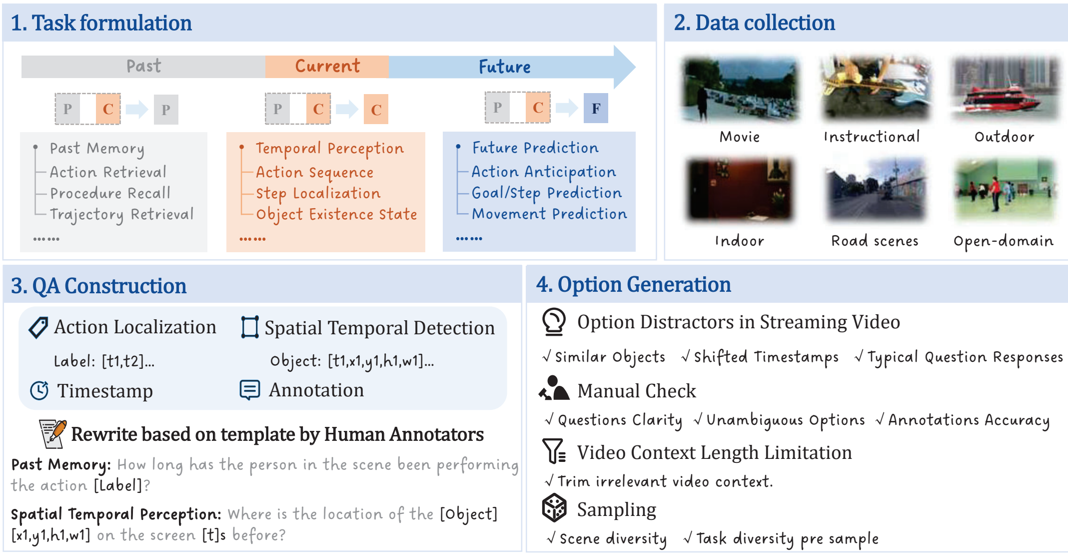


Figure 2. **Generation pipeline of OVBench.** We developed a method to ensure the quality of annotation based on the existing high-quality spatiotemporal data, including task definition, data collection, QA construction, and multiple-choice question generation suitable for streaming video scenarios. The details will be discussed in Section 3.

tion. Flash-Vstream [56] and VideoStreaming [38] achieve real-time comprehension through a learnable memory module for stream compression. However, prior work has often lacked well-reasoned architectural designs and training strategies, consequently struggling to achieve a balance between efficiency and performance. Our approach introduces novel designs in both architectural structure and training strategy, leading to the development of a more powerful online video MLLM.

**Online Video Benchmarks.** VideoLLM-online [5] evaluates the model as an online video assistant on the streaming narration task. MovieChat-1K [41] introduces a breakpoint mode, which requires the model to ask and answer questions at different time points during video playback. VStream-QA [56] represents the first benchmark to evaluate streaming multimodal video understanding. While it incorporates five types of timestamp-anchored questions generated through GPT-4 with human verification, its task paradigm largely mirrors offline scenarios. In contrast, OVBench emphasizes real-time spatiotemporal detail comprehension in streaming contexts, featuring a comprehensive task set tailored to streaming video characteristics. It builds upon high-quality spatiotemporal understanding datasets through targeted refinement to ensure benchmark integrity and reliability.

# 3. OVBench

In this section, we detail the development process of OVBench. Based on a foundational definition of temporal contexts, we first derive the task types in Figure 1 for online video streaming scenarios in section 3.1. We then introduce the detailed process of QA generation in Figure 2. Examples are listed in Table 1.

## 3.1. Task Formulation

**Basic Temporal Context Definition.** To systematically evaluate streaming video comprehension, we define three fundamental temporal contexts that characterize the relationship between a question's timestamp and the video timeline: **(1) Current:** The temporal window focuses specifically on the exact frame at which the question is posed, potentially including a small number of preceding frames necessary for understanding the current state. **(2) Past:** The sequence of frames preceding the question timestamp, containing historical information about actions, events, and object trajectories. **(3) Future:** The sequence of frames following the question timestamp, capturing the subsequent events based on current actions and trajectories.

**Task Formulation.** Based on three core temporal contexts — **Past (P)**, **Current (C)**, and **Future (F)** — we identify 6 essential capabilities for online models in streaming video scenarios, where "$\rightarrow$" represent the inference or verification process from one time period to another, "$\cup$" indicates a joint understanding of multiple periods:

- **Spatial Perception (C):** Identify and quantify discrete actions in the current frame, describe positions and spatial relationships of objects.
- **Temporal Perception ($\mathbf{C \rightarrow P \cup C}$):** Track sequences

| Task | Subtasks | Query Examples |
|---|---|---|
| Temporal Hallucination Verification (THV) | Action Persistence | Is the person in the [0.168, 0.193, 0.846, 0.996] location in the current frame performing walking? |
| | Step Verification | Is the person still installing the motherboard right now? |
| | Object Presence | How many markers are there on the screen 14.0 seconds before? Does the number increase or decrease compared with the past screen? |
| Past Memory (PM) | Action Retrieval | Where was the person currently performing the talk to (e.g., self, a person, a group) in the scene 8 seconds ago? |
| | Procedure Recall | Which step did the person perform for the longest duration in the last 60 seconds? |
| | Trajectory Retrieval | When does the sheep [0.491, 0.386, 0.584, 0.615] in the current screen first appear? Give the corresponding position when it first appears. |
| Future Prediction (FP) | Action Anticipation | What action is the person currently in the [0.328, 0.211, 0.436, 0.809] location likely to do next? |
| | Goal/Step Prediction | My goal is 'make flower crown'. What are the next steps I should take? |
| | Movement Prediction | What direction do you think the baby [0.0, 0.062, 0.526, 0.903] may move towards in the next second? |
| Spatial Perception (SP) | Action Location | What is the person at the location [0.024, 0.122, 0.624, 0.979] currently doing? |
| | Object Position | Which option most accurately describes the location of the blankets now? |
| Temporal Perception (TP) | Action Sequence | What is the sequence of actions the person in the scene has performed recently? |
| | Step Localization | How long has the person in the scene been performing the 'restore the fixed battery components and the back cover'? |
| | Object Existence State | What is the time period the turtle [0.459, 0.518, 0.501, 0.556] in the current screen appears in the video? And what is the time period in which it disappeared? |
| Spatio-Temporal Perception (STP) | Action Trajectory | What is the sequence of actions and the corresponding movement trajectory of the person currently in the [0.383, 0.304, 0.642, 0.991] location? |
| | Object Trajectory | What is the trajectory of the object among car [0.482, 0.518, 0.485, 0.531], car [0.561, 0.51, 0.616, 0.577] in the past 5 seconds, which moves the shortest distance? |

Table 1. Task examples of OVBench. For simplicity, we selected only one question in each task's templates for the presentation. Complete template examples can be found in the Appendix.

of actions extending into the present moment, assess the duration of ongoing events and determine the existence status of objects over previous frames.

- **Spatio-temporal Perception ($\mathbf{C} \rightarrow \mathbf{P} \cup \mathbf{C}$):** Provide a comprehensive description of object motion trajectories, detailing displacements and relative positions for single or multiple targets.
- **Past Memory ($\mathbf{C} \rightarrow \mathbf{P}$; or $\mathbf{P}$):** Recall past events relevant to a given action, retrieve duration or goals achieved, or locate an object's past position and status when queried.
- **Temporal Hallucination Verification ($\mathbf{P} \leftrightarrow \mathbf{C}$):** Determine if an action observed in the past is still ongoing in the current frame, verify the state of events that have occurred, and analyze object location changes between past and current contexts.
- **Future Prediction ($\mathbf{P} \cup \mathbf{C} \rightarrow \mathbf{F}$):** Project likely upcoming actions based on observed motion patterns and current spatial-temporal configurations.

### 3.2. QA Generation

**Data Collection.** Unlike previous online question-answering benchmarks [56], which typically utilize LLMs to generate questions and answers, our task requires temporal and spatial detail understanding, where questions rely on specific timestamps and bounding box annotations for accurate spatiotemporal comprehension. To comprehensively capture the dynamics of streaming video, we curated 8 datasets across 6 varied domains in Figure 2, each dataset is selected to align closely with the real-time demands of streaming video comprehension. We only select from their validation and test sets to prevent potential data leakage.

**Option Generation for Streaming Video Scenes.** To ensure that answer options reflect the dynamic and shifting contexts in streaming video, we develop a multiple-choice generation process incorporating distractors that simulate real-world conditions. Distractors are selected from different timestamps within the same video, based on similar questions and objects or typical responses to such questions.

**Manual Check & Sampling. (1) Manual Check:** Our quality control involved several manual checks: question clarity and options ambiguity were assessed to avoid misinterpretations and annotation accuracy was verified by human annotators. **(2) Video Context Length Limitation:** Excessive video context length is trimmed according to the timestamp of the earliest relevant question. (i.e. The maximum time range that a question in the **Past Memory** task may be traced back to.) **(3) Sampling:** Further, to ensure question diversity and balance, question distribution is optimized by scaling question count proportionally to video duration, ensuring diverse scene coverage, and maintaining task type balance. The appendix provides a detailed QA generation methodology across tasks, ensuring the transparency of our approach.

## 4. Efficient Online Video Streams Modeling

### 4.1. Pyramid Memory Bank

For online scenarios, as the number of input frames increases, it becomes essential to compress the visual tokens of the video in order to maintain real-time performance while preserving key information. Achieving a balance between spatial and temporal details within limited visual to-

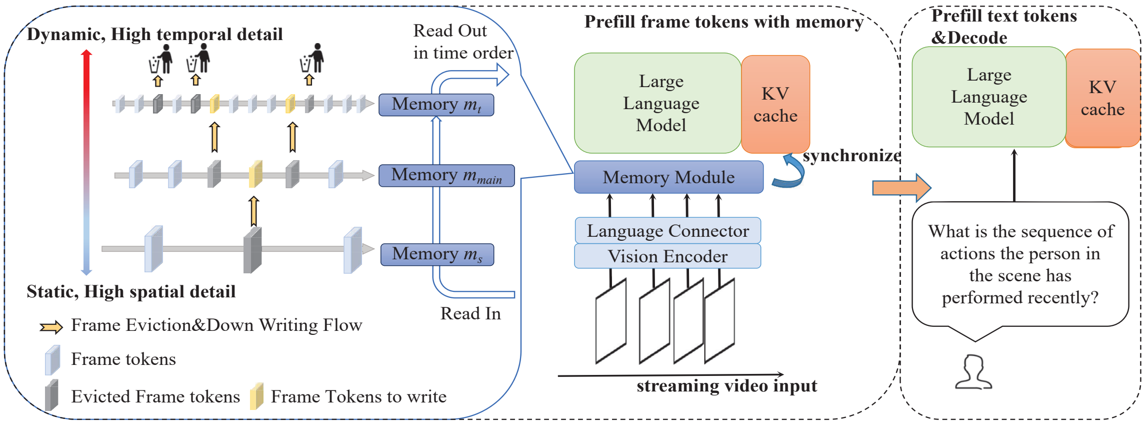


Figure 3. **Pyramid Memory Bank Architecture**: Illustrating the model's inference process with the pyramid memory bank structure. $m_{main}$ queues maintain balanced spatiotemporal information at different hierarchical levels, $m_t$ is a high-frequency sampling queue for enhanced temporal detail preservation, and $m_s$ queue is for spatial detail retention. The system supports simultaneous frame input to both the memory bank and KVCache, with synchronization mechanisms for maintaining consistency during memory modifications.

kens is critical for effective spatiotemporal understanding.

To address this, we propose a **pyramid memory bank** structure that incrementally balances spatial and temporal details through progressive abstraction across multiple layers. As shown in Figure 3, the memory bank is divided into $n$ layers, denoted as $\{m_i \mid i = 1, 2, \ldots, n\}$. Each layer progressively reduces spatial details in favor of temporal patterns by adjusting two key properties:

**Sampling Rate** ($r_i$): Each layer $i$ samples frames from the input stream at a rate $r_i$, increasing progressively across layers to prioritize temporal continuity in deeper layers.

**Resolution** ($\text{Res}_i$): Each layer stores frames at a progressively lower resolution $\text{Res}_i$, ensuring that initial layers capture detailed spatial information, while deeper layers focus on temporal abstraction. The resolution for each layer is scaled as: $\text{Res}_i = \frac{\text{Res}_1}{\beta^{i-1}}$ where $\text{Res}_1$ is the input frame resolution in the first layer, and $\beta > 1$ is a down-scaling factor. In practice, we use $\beta = 2$.

Each memory layer $m_i$ performs 3 primary operations:

1. **Streaming Frame Writing**: The memory layer $m_i$ receives frames directly from the video stream, sampled according to $r_i$. These frames are stored in $m_i$ up to its capacity $C_i$. When the capacity $C_i$is full, perform the next operation.

2. **Frame Eviction&Down Writing**: the memory layer identifies the most similar adjacent frame pair $(f_a^i, f_b^i)$, where cosine similarity is calculated after applying average pooling to each frame separately. The older frame in the pair is evicted, and its spatial information is reduced to the corresponding spatial scale $\text{Res}_{i+1}$ through average pooling before being passed to the next layer $m_{i+1}$:

$$f_{\text{next}}^{i+1} = \text{AvgPool2d}(f_{\text{evicted}}^i, \text{Res}_{i+1}) \tag{1}$$

3. **Readout**: All stored frames across layers are read in temporal order when accessing memory banks.

**Compatibility with KVCache.** Existing memory-based compression methods, such as MovieChat [41, 42] and FlashVStream [56], as the memory updates with each additional input frame, the entire compressed memory must be processed as a single unit when a user inputs frames, suffer from a bottleneck in compression efficiency. This all-at-once processing leads to memory compression as computational overhead, limiting real-time performance. In contrast, our memory bank aligns closely with KVCache, allowing frame tokens to be precomputed and stored efficiently. During **Frame Eviction** (operation 2), tokens after the timestamps of frames $f_a$ and $f_b$ are erased to maintain synchronization, as follows:

$$\text{KVCache} \leftarrow \text{KVCache} \setminus \{t_i \mid t_i > \min(t_{f_a}, t_{f_b})\} \tag{2}$$

where $t_{f_a}$ and $t_{f_b}$ denote the timestamps of frames $f_a$ and $f_b$, respectively. By erasing tokens after these timestamps, we can optimize both memory usage and real-time processing efficiency.

## 4.2. Offline-to-Online Learning

**Data Collection.** To enhance the model's online spatiotemporal understanding capabilities, we prioritized datasets with rich spatiotemporal annotations. These datasets, featuring dense temporal annotations and spatial tracking information, inherently support multi-turn dialogue scenarios in streaming contexts.

- **Fine-grained Event Temporal Boundary Identification.** To capture temporal event evolution, we leverage TimeChat-IT [40] data, incorporating dense video cap-

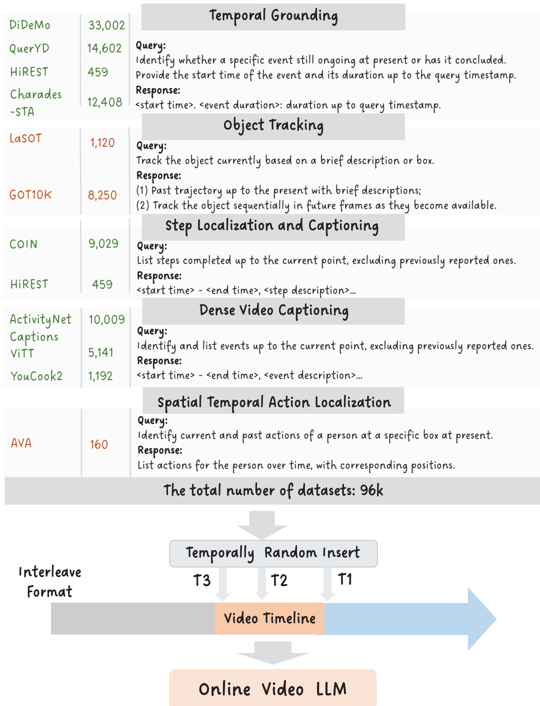


Figure 4. **Data Format Conversion Process for Online Spatiotemporal Instruction-Finetuning.** Our pipeline begins with 96K high-quality samples curated from 5 tasks across 12 datasets. The conversion process enhances online spatiotemporal understanding through template transformation. For each video sample, we strategically insert queries along the timeline in an organized interleaved format to facilitate temporal context differentiation.

tioning [19, 24, 60], step localization [43, 53], and temporal grounding datasets [3, 14, 33] for precise temporal boundary annotations.

- **Detailed Spatiotemporal Understanding.** We integrate object tracking [11, 20] and spatiotemporal action localization [15] annotations to enhance sequential object and action tracking capabilities, complementing the temporal information framework.

**Data Conversion.** To enhance multi-turn dialogue coherence and contextual awareness, we implement a structured temporal sampling strategy for question formulation, as shown in Figure 4. Questions are positioned at specific temporal intervals while maintaining natural dialogue progression. Each sample's queries maintain task-category consistency to facilitate cross-dialogue temporal reasoning. Following MVBench [27], we generate 5 diverse instructions per annotation task to ensure comprehensive interaction scenario coverage.

**Progressive Training.** Optimizing fine-grained spatiotemporal understanding while maintaining timestamp and bounding box prediction capabilities during online training presents significant challenges. Inspired by curriculum learning [4], we initially train the model on offline data to establish robust video understanding, followed by joint optimization with online data integration. An empirical analysis of this approach is presented in the ablation study section.

## 5. Experiments

### 5.1. Implementation Details

**Training Data.** To enhance the model's comprehensive video understanding, we supplemented the online training data with offline video data from VideoChat2-IT [28], STAR [47] and PerceptionTest [37], image data from ShareGPT4V [6], ShareGPT4o [8], as well as multi-image data from LLaVA-OneVision [25].

**Model Architecture.** We use InternVL2-4B [8] as a powerful baseline model for development, integrating InternViT-300M as the visual encoder and Phi-3 [1] as the language model. The input frames of all training processes are obtained by sampling at 1 fps, and the maximum input frame number is controlled at 64 frames through uniform sampling. For inference, the video sample rate for each memory in the hierarchical memory bank is $\{1, 2, 8\}$ with token per frame $\{256, 64, 16\}$, respectively. Maintain a consistent token ratio for each memory queue.

**Evaluation Settings.** As most multimodal large models currently cannot receive streaming video input, we adopted two distinct methods to evaluate the models effectively:

- **Sliding Window Setting:** We perform a sliding window evaluation on advanced MLLMs, capturing a 32-second time window before the question-asking time and extracting frames at 2 fps. This method allows for temporal context while evaluating responses.
- **Streaming Setting:** In this setting, we input all video frames from the beginning of the clip up to the question timestamp, sampled at 2 fps, to evaluate the model's real-time performance.

### 5.2. Main Results on OVBench

The main results on OVBench are shown in Table 2.

**Streaming Video LLMs Comparison.** In the streaming video setting, answering questions is more difficult because the model will be disturbed by more irrelevant context (see the FIFO and "w/o compression" of Table 5, when the entire video from the beginning to the end frame is input, the performance is reduced by about 4.62%). Nevertheless, current online models have a non-negligible gap in real-time performance with existing offline models, suggesting that accessing rich, task-specific online data is crucial to narrowing this gap. Although Flash-Stream [56] employs an online model architecture, training with offline data may introduce biases

| Task Name | Size | FP | | | THV | | | PM | | | SP | | STP | | TP | | | AVG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Subset Name | | AA | GSP | MP | AP | SV | OP | AR | PR | TR | AL | OP | AT | OT | AS | SL | OES | |
| *sliding window size=32s fps=2* | | | | | | | | | | | | | | | | | | |
| Gemini-1.5-Flash [2] | - | 71.4 | 53.6 | 21.9 | 56.5 | 60.8 | 40.6 | 36.7 | 47.9 | 62.5 | 32.3 | 37.5 | 87.0 | 50.0 | 83.3 | 22.3 | 46.9 | 50.7 |
| InternVL2 [8] | 7B | 52.6 | 60.2 | 27.6 | 57.5 | 52.0 | 58.5 | 38.8 | **67.1** | 58.3 | 38.1 | **31.3** | 87.4 | 37.0 | 75.4 | 31.4 | 5.9 | 48.7 |
| InternVL2 [8] | 4B | 57.7 | 57.0 | 14.4 | 59.2 | 49.4 | 60.0 | 30.3 | 61.8 | 46.3 | 30.9 | 20.1 | 83.0 | 32.3 | 70.7 | 29.4 | 3.4 | 44.1 |
| LLaMA-VID [10] | 7B | 43.6 | 50.9 | 19.6 | 64.0 | 47.5 | 46.8 | 29.4 | 48.9 | 51.2 | 31.9 | 11.2 | 75.7 | 24.8 | 59.1 | 26.0 | 40.0 | 41.9 |
| LLaVA-Onevision [25] | 7B | **68.0** | 62.7 | **35.9** | 58.4 | 50.3 | 46.5 | 29.4 | 60.7 | 58.0 | **43.1** | 14.2 | 86.5 | 49.7 | 70.7 | 28.1 | 30.2 | 49.5 |
| LongVA [57] | 7B | 64.1 | 56.5 | 29.5 | 54.9 | 51.9 | 34.8 | 35.3 | 55.6 | 57.7 | 31.6 | 3.4 | 67.4 | 44.7 | **80.0** | 26.7 | 4.0 | 43.6 |
| MiniCPM-V2.6 [16] | 7B | 33.3 | 35.9 | 15.0 | 59.2 | 50.8 | 55.1 | 25.0 | 37.4 | 41.7 | 26.6 | 11.8 | **98.3** | 36.3 | 66.1 | 26.4 | 6.2 | 39.1 |
| Qwen2-VL [45] | 7B | 60.3 | **66.1** | 22.1 | 54.9 | 51.5 | 51.1 | 37.8 | 64.4 | 69.3 | 35.3 | 28.5 | 97.0 | 49.4 | 65.1 | 30.8 | 11.7 | 49.7 |
| LITA [18] | 7B | 19.2 | 24.5 | 19.9 | 40.8 | 48.9 | 24.9 | 3.1 | 27.3 | 6.4 | 6.9 | 14.6 | 35.2 | 23.9 | 27.4 | 0.5 | 3.4 | 20.4 |
| TimeChat [40] | 7B | 7.7 | 15.3 | 18.7 | 20.6 | 15.7 | 11.7 | 9.1 | 14.7 | 9.8 | 7.5 | 19.5 | 13.9 | 10.3 | 9.3 | 10.1 | 10.8 | 12.8 |
| VTimeLLM [17] | 7B | 37.2 | 23.4 | 15.0 | **64.8** | 43.8 | 53.2 | 25.9 | 38.8 | 32.5 | 25.9 | 20.4 | 40.9 | 6.8 | 48.4 | **43.5** | 8.6 | 33.1 |
| VideoChat-Online (Ours) | 4B | 56.4 | 63.0 | 15.6 | 57.1 | **57.9** | **61.9** | **39.1** | 54.2 | **73.9** | 41.3 | 29.7 | 92.2 | **53.1** | 69.8 | 27.3 | **69.9** | **53.9** |
| *Streaming video input at 2 fps.* | | | | | | | | | | | | | | | | | | |
| VideoLLM-Online [5] | 7B | 0 | 1.8 | 20.9 | 5.2 | 5.9 | 32.6 | 0 | 2.3 | 26.7 | 0.6 | 26.6 | 0.9 | 19.9 | 0.9 | 1.7 | 8.3 | 9.6 |
| MovieChat [41] | 7B | 23.1 | 27.5 | 23.6 | 58.4 | 43.9 | 40.3 | 25.6 | 31.1 | 23.9 | 26.9 | **39.6** | 24.4 | 28.9 | 29.3 | 25.5 | 21.9 | 30.9 |
| Flash-Vstream [56] | 7B | 26.9 | 37.6 | **23.9** | 60.1 | 41.9 | 40.0 | 23.4 | 35.3 | 26.1 | 24.7 | 28.8 | 27.0 | 21.4 | 29.8 | 25.6 | 26.8 | 31.2 |
| VideoChat-Online (Ours) | 4B | **64.1** | **59.7** | 16.6 | **63.1** | **58.3** | **62.8** | **42.2** | **54.4** | **70.6** | **54.1** | 24.8 | **88.7** | **48.5** | **73.0** | **25.9** | **71.7** | **54.9** |

Table 2. Evaluations results on OVBench. Our 4B-parameter model demonstrates substantial performance advantages in two key comparisons: a 23.7% improvement over existing streaming-capable models, and a 4.2% enhancement compared to advanced offline MLLMs while maintaining deployment flexibility. For VideoLLM-Online, we modify the official script for evaluation on the OVBench. However, it cannot follow instructions accurately and generates either nothing or redundant information, see the appendix for more details.

| Model | Size | EgoSchema | MLVU | VideoMME | | MVBench | LongVideo Bench |
|---|---|---|---|---|---|---|---|
| | | | | Overall | Long | | |
| Video-LLaVA [29] | 7B | 38.4 | 47.3 | 39.9 | 36.2 | - | 39.1 |
| Chat-UniVi [23] | 7B | - | - | 40.6 | 35.8 | - | - |
| LLaMA-VID [10] | 7B | 38.5 | 33.2 | - | - | 41.9 | - |
| TimeChat [40] | 7B | 33.0 | 30.9 | 34.7 | 32.3 | 38.5 | - |
| MovieChat [41] | 7B | 53.5 | 25.8 | 38.2 | 33.4 | 55.1 | - |
| Video-LLaMA2 [54] | 7B | 51.7 | 48.5 | 47.9 | | 54.6 | - |
| LLaVA-Next-Video [58] | 7B | 43.9 | - | 46.6 | | | 43.5 |
| ShareGPT4Video [7] | 8B | - | 46.4 | 39.9 | 35.0 | 51.2 | 39.7 |
| VideoChat2 [27] | 7B | 54.4 | 47.9 | 39.5 | 33.2 | 60.4 | 39.3 |
| LongVA [57] | 7B | - | 56.3 | 52.6 | 46.2 | - | - |
| Video-CCAM [12] | 9B | - | 58.5 | 50.3 | 39.6 | 64.6 | - |
| Video-CCAM [12] | 4B | - | 56.5 | 50.1 | 40.9 | 62.8 | - |
| VideoChat-Online (Ours) | 4B | **54.7** | **60.8** | **54.4** | **47.1** | **65.2** | **54.1** |

Table 3. VideoChat-Online's results on other offline long and short video benchmarks show comparable or better overall performance.

| Memory Bank | Memory Capacity | | | OVBench | | | |
|---|---|---|---|---|---|---|---|
| | $m_t$ | $m_{main}$ | $m_s$ | SP | TP | STP | Overall |
| w/o $m_t$ | 0 | 5 | 2 | 38.9 | 51.1 | 68.5 | 54.2 |
| w/o $m_{main}$ | 20 | 0 | 2 | 35.4 | 55.7 | 69.2 | 54.3 |
| w/o $m_s$ | 12 | 10 | 0 | 36.9 | **56.9** | **69.3** | 53.7 |
| Ours | 12 | 2 | 2 | **39.4** | **56.9** | 68.6 | **54.9** |

Table 4. **Memory Structure Ablations.** We remove each memory module for evaluation but keep the fixed number of visual tokens.

that impact real-time adaptability. VideoLLM-Online [5], trained on streaming narration data and free-form dialogue data from a first-person perspective, faces challenges in generalizing across diverse contexts and applications. Comparisons of computational costs and online scenarios cases with other models can be found in the appendix.

**Offline MLLMs in a Sliding Window Setting.** Our results show that offline MLLMs are more effectively generalized to online tasks, as they can be seen as a special case of broader online scenarios (e.g., using fixed question timestamps at the end of the video rather than dynamically across the video). **Effective Knowledge Transfer from offline to online:** Offline models have demonstrated superior performance than native online models in online scenarios, highlighting that knowledge from offline models can successfully transfer to online applications. Developing and transforming the streaming video model architecture based on existing advanced MLLMs emerges as a superior choice. Furthermore, by incorporating limited online scene data, our model achieves optimal results with a small 4B LLM, outperforming Qwen2-VL by 4.19%.

### 5.3. Offline Benchmark Results

We conduct experiments on offline video understanding benchmarks to comprehensively evaluate our model. The test videos range from a few seconds to one hour in length. As shown in Table 3, our model outperforms previous state-of-the-art approaches, achieving the highest scores across multiple benchmarks, including 54.7% on Egoschema [31], 60.8% on MLVU [59], 54.4% on VideoMME [13] (Overall), 47.1% on VideoMME (Long), 65.2% on MVBench [27], and 54.1% on LongVideoBench [48]. This consistent performance across both online and offline scenarios demonstrates VideoChat-Online's robust generalization capabilities.

### 5.4. Ablations of VideoChat-Online

**Memory Bank Design.** We conduct comprehensive ablation studies on the memory component across two dimen-

| Update Policy | OVBench | VideoMME-long |
|---|---|---|
| Token Merge [41] | 51.5 | 43.9 |
| First In First Out | 54.0 | 41.3 |
| Uniform Sample | 52.1 | 45.0 |
| w/o Compression | 49.3 | *oom* |
| Ours | **54.9** | **47.1** |

Table 5. **Memory Update Policies Ablations.** FIFO prioritizes recent data for real-time queries, while Uniform Sample is applied to offline models. The "w/o Compression" method directly inputs raw video data without further processing. Our method outperforms existing approaches in both online and offline benchmarks.

| Training Strategy | | | AVG |
|---|---|---|---|
| Offline Training | Online Training | | |
| | Progressive | Interleaved | |
| | | | 44.12 |
| ✓ | | | 45.23 (+1.11) |
| ✓ | ✓ | | 52.42 (+8.30) |
| ✓ | | ✓ | 51.84 (+7.72) |
| ✓ | ✓ | ✓ | **53.89 (+9.77)** |

Table 6. Ablation analysis on training strategy and data organization impact on model performance.

sions: **(1) Structure.** Table 4 demonstrates the necessity of each memory layer. We systematically evaluated the impact of removing each layer while maintaining computational parity by adjusting token allocations. We maintained computational equivalence by ensuring the total token count remained constant at 832 tokens across all configurations. The results reveal distinct patterns: spatial-biased memory configurations significantly enhance SP performance (comparing "w/o $m_t$" with "w/o $m_{main}$" and "w/o $m_s$", while temporally distributed memory structures improve TP and STP metrics. Our final architecture achieves optimal spatiotemporal balance through strategic memory allocation. **(2) Memory Updating Policy.** Table 5 presents a comparative analysis of various memory bank update strategies, including Token Merge, FIFO queuing, uniform temporal sampling, and uncompressed frame input. While FIFO inherently prioritizes recent temporal information, our proposed strategy demonstrates superior performance by facilitating cross-temporal information interaction between memory banks. This advantage stems from addressing a fundamental limitation of MLLMs: their tendency to encode frames independently, where token similarity reflects relationships between frames. Simple merging operations fail to preserve crucial inter-frame temporal dynamics.

**Impact of the Training Paradigm.** Table 6 presents a systematic analysis of different training strategies across three key dimensions: **(1) Effect of Online Data:** Training exclusively with offline data yields a baseline performance of 45.23%. Incorporating online data through progressive training elevates performance to 52.42%, representing an

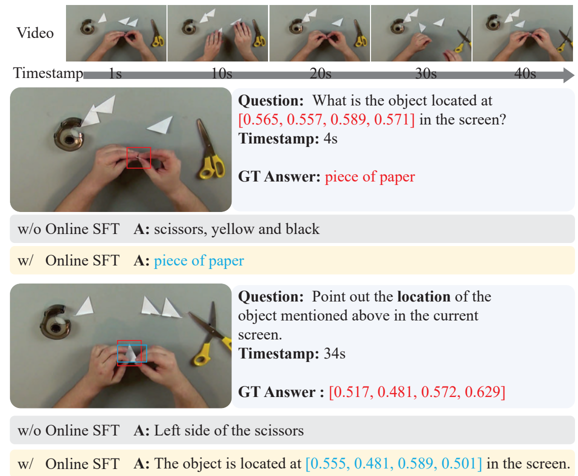


Figure 5. Qualitative comparison on online data training

8.66% improvement. Notably, this substantial enhancement is achieved with merely 96K online samples (6% of the total training data), demonstrating the critical role of online data in developing temporal context understanding capabilities. **(2) Impact of Interleaved Data Format:** The integration of interleaved data format with offline training improves performance from 45.23% to 51.84%, with a 6.61% increase. This enhancement suggests that interleaved data organization facilitates more effective learning of temporal relationships between question-answer pairs, particularly in dynamic online scenarios. **(3) Progressive Training Strategy:** The combination of progressive training with interleaved format achieves 53.89%, surpassing the joint training approach (51.84%) by 2.05%. This improvement indicates that transitioning from joint training to progressive online data introduction in the second epoch facilitates a better alignment progression from coarse to fine-grained spatiotemporal understanding.

**Qualitative comparison of Online SFT.** As shown in Figure 5, we verified it in a multi-round dialogue scenario and found that compared with the model based on offline training alone, the model after Online SFT accurately outputs the location and better builds the connection dialogue.

## 6. Conclusion

This work presents the following contributions to advance streaming video understanding: (1) OVBench, a comprehensive benchmark designed to evaluate real-time spatiotemporal understanding capabilities; (2) VideoChat-Online, an efficient streaming video model that effectively balances efficiency. It achieves state-of-the-art performance while maintaining deployment flexibility. These advances provide a solid foundation for future research in streaming video understanding and real-world applications.

**Acknowledgment**: This work is supported by the National Key R&D Program of China (No. 2022ZD0160900), Jiangsu Frontier Technology Research and Development Program (No. BF2024076), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and Nanjing University-China Mobile Communications Group Co., Ltd. Joint Institute.

## References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 6

[2] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*, 2023. 7

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 6

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 6

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18407–18418, 2024. 1, 2, 3, 7

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 6

[7] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 7

[8] 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. 2, 6, 7

[9] Wei-Lin Chiang et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[10] Yanwei Li et al. Llama-vid: An image is worth 2 tokens in llms. In *ECCV*, 2024. 2, 7

[11] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[12] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024. 7

[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, 7

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

[15] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. *CVPR*, 2017. 6

[16] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 2, 7

[17] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 2, 7

[18] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer, 2025. 7

[19] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020. 6

[20] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021. 6

[21] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: toward spatio-temporal reasoning

in visual question answering. In *CVPR*, pages 1359–1367, 2017. 2

[22] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. 2

[23] 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. 2, 7

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

[25] 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, 6, 7

[26] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *ArXiv*, abs/2305.06355, 2023. 1, 2

[27] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *CoRR*, abs/2311.17005, 2023. 1, 2, 6, 7

[28] 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. *arXiv preprint arXiv:2311.17005*, 2023. 2, 6

[29] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv:2311.10122*, 2023. 7

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

[31] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *ArXiv*, abs/2308.09126, 2023. 1, 7

[32] Meta. Build the future of ai with meta llama 3. https://llama.meta.com/llama3/, 2024. 2

[33] Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2265–2269, 2021. 6

[34] OpenAI. Chatgpt. https://openai.com/blog/chatgpt/, 2023. 2

[35] OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023. 1

[36] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 1

[37] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan, Banarse, Mateusz Malinowski, Yezhou Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine, Miech, Skanda Koppula, Alexander Fréchette, Hanna Klimczak, R. Koster, Junlin Zhang, Stephanie, Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test : A diagnostic benchmark for multimodal models. In *NeurIPS*, 2023. 1, 6

[38] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *Advances in Neural Information Processing Systems*, pages 119336–119360. Curran Associates, Inc., 2024. 3

[39] David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024. 2

[40] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, 2024. 2, 5, 7

[41] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv:2307.16449*, 2023. 2, 3, 5, 7, 8

[42] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024. 2, 5

[43] 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 *CVPR*, pages 1207–1216, 2019. 6

[44] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez,

Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. 2

[45] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 7

[46] 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. 2

[47] Bo Wu, Shoubin Yu, and Tenenbaum Chen, Zhenfang. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 6

[48] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. 7

[49] Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation, 2024. 2

[50] 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 *ICME*, 2017. 2

[51] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1

[52] 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 *AAAI*, 2019. 2

[53] 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 (CVPR)*, pages 23056–23065, 2023. 6

[54] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv:2306.02858*, 2023. 1, 7

[55] Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *CoRR*, abs/2312.04817, 2023. 2

[56] 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. 1, 2, 3, 4, 5, 6, 7

[57] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 7

[58] 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. 7

[59] 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, 7

[60] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, 2017. 6