CogStream: Context-guided Streaming Video Question Answering

Zicheng Zhao[1*], Kangyu Wang[1*], Shijie Li[1*], Rui Qian[2], Weiyao Lin[1], Huabin Liu[1†]

[1]Shanghai Jiao Tong University
[2]The Chinese University of Hong Kong
zzcsjtu7@sjtu.edu.cn, kangyuwang@sjtu.edu.cn, shijieli@sjtu.edu.cn, qr021@ie.cuhk.edu.hk, wylin@sjtu.edu.cn, huabinliu@sjtu.edu.cn

Abstract

Despite advancements in Video Large Language Models (Vid-LLMs) improving multimodal understanding, challenges persist in streaming video reasoning due to its reliance on contextual information. Existing paradigms feed all available historical contextual information into Vid-LLMs, resulting in a significant computational burden for visual data processing. Furthermore, the inclusion of irrelevant context distracts models from key details. This paper introduces a challenging task called **Co**ntext-**g**uided **Stream**ing Video Reasoning (**CogStream**), which simulates real-world streaming video scenarios, requiring models to identify the most relevant historical contextual information to deduce answers for questions about the current stream. To support CogStream, we present a densely annotated dataset featuring extensive and hierarchical question-answer pairs, generated by a semi-automatic pipeline. Additionally, we present *CogReasoner* as a baseline model. It effectively tackles this task by leveraging visual stream compression and historical dialogue retrieval. Extensive experiments prove the effectiveness of this method.

Code — https://github.com/LiamZhao326/CogStream
Datasets —
https://huggingface.co/datasets/SII-KYW/CogStream
Extended version — https://arxiv.org/abs/2506.10516

1 Introduction

Streaming video understanding has emerged as a crucial task. It is anticipated to conduct a dynamic and comprehensive interpretation of video stream. In a streaming context, ***streaming Video Question Answering*** (VQA) involves scenarios where users watch an ongoing video stream and continuously interact with the model. Users continuously ask questions about the latest video content, while the model provides answers based on the video content it has seen thus far.

However, current Vid-LLMs still face significant challenges in performing streaming VQA, stemming from: (1) ***Multi-turn contextual reasoning***, where dialogues are logically inter-connected, requiring Vid-LLMs to leverage historical dialogue information for accurately answering current questions. (2) ***Spatio-temporal information dynamics***, requiring the ability to update adaptive answers that evolve in sync with dynamic visual information over time. To this end, existing methods try to compress frames (Qian et al. 2024) to capture more comprehensive visual features or enhance memory mechanisms (Zhang et al. 2024a; Xiong et al. 2025) to retain more information from historical conversations.

However, the rapid growth of video data within streams poses great challenges for efficient visual information processing. Moreover, current methods typically rely on summarizing *all* historical textual information to understand the current stream. However, including irrelevant context easily distracts models, causing them to misinterpret insignificant details as important and thereby undermining the reasoning.

Based on the idea of streaming VQA, we introduce a novel and challenging task named ***Co***ntext-***g***uided ***Stream***ing Video Reasoning (***CogStream***). The core of this task is to identify the most relevant historical contextual information for streaming video reasoning. By focusing on pertinent cues derived from relevant historical context, Vid-LLMs can significantly enhance reasoning accuracy and efficiency, circumventing the need to process the entire historical stream. To support this task, we present a new dataset with distinctive features: (1) A semi-automatic pipeline for constructing an annotated dataset from unlabeled videos; (2) High-quality QA pairs where answers to questions about the current stream are supported and deduced by previous dialogue; and (3) Hierarchical reasoning tasks that offer various levels of streaming VQA complexity.

Furthermore, we propose a baseline method, ***CogReasoner***, which learns to (1) compress the accumulated video stream, (2) retrieve relevant historical QA pairs, and (3) reason over the integrated visual-textual information, enabling a more effective and streamlined solution for the CogStream task.

2 Related Work

Video Large Language Models The evolution of large language models (LLMs) (Chiang et al. 2023; Peng et al. 2023; Dubey et al. 2024) has significantly propelled the creation of video large language models (Vid-LLMs) (Fu et al. 2025; Zhang, Li, and Bing 2023; Wang et al. 2024b). Vid-LLMs augment LLMs with multimodal data, broadening their utility. Many of these, including VideoLLaMA (Zhang, Li, and Bing 2023), VideoLlaVA (Lin et al. 2023), and InternVL (Chen et al. 2024d), draw inspiration from BLIP-2 to conduct large-

*These authors contributed equally.
†Corresponding author.

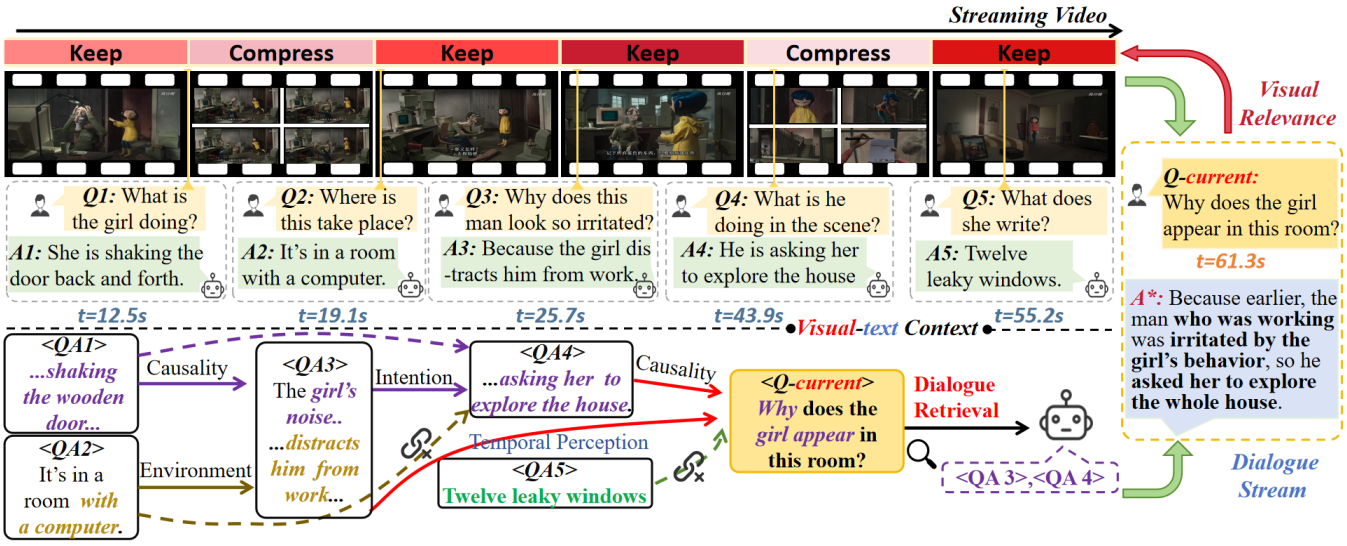

Figure 1: Illustration of CogStream. Given a streaming video, users continuously interact with models by asking questions. Both the video data and the history of QA dialogue grow with the stream. To answer the latest question, models must deduce the answer from relevant historical context, thereby forming the dialogue stream. Our CogReasoner addresses this task by compressing streaming video based on current questions and accurately retrieving relevant historical QAs to deduce the answer.

scale video-text pre-training, enabling them to comprehend and analyze video content. Nevertheless, handling long and streaming video content continues to present hurdles due to high computational demands, memory constraints, and the complexity of modeling long-range temporal relationships.

Streaming Video Understanding Existing streaming video research primarily focuses on specific visual tasks, such as real-time object tracking (Liu et al. 2022; Wang et al. 2020), action recognition (Luvizon, Picard, and Tabia 2020; Zhang et al. 2016), and instantaneous video content description (Chen et al. 2024a). While these methods excel in their individual domains, they often fall short in complex multi-task scenarios and lack the ability for in-depth understanding across varied time segments. Similarly, current benchmarks like SVBENCH (Yang et al. 2025) predominantly evaluate relationships between adjacent video segments, neglecting the deep reasoning required for *longer temporal contexts*. This limitation inherently constrains performance on tasks that necessitate integrating information across multiple segments.

Video Question-Answering VideoQA is crucial for evaluating Vid-LLMs' capabilities, particularly their grasp of long-term context. Existing VQA datasets fall into two categories: static (e.g., REXTIME (Chen et al. 2024b), NextQA (Xiao et al. 2021), Video-MME (Fu et al. 2025)) and streaming (e.g., VStream-QA (Zhang et al. 2024b), STREAMBENCH (Wu et al. 2024), SVBench (Yang et al. 2025)). While static benchmarks leverage global video input for reasoning, they lack streaming support and associations between QAs. Streaming datasets, conversely, target long-span temporal understanding but often underutilize historical QAs for context-driven reasoning. Differently, this paper introduces a novel dataset & approach focusing on complex contextual relationships within streaming videos. Our dataset features QA pairs spanning extended time periods and cross-segment associations, thereby compelling models to leverage historical QA information for dynamic reasoning.

Tab. 1 compares our work with current benchmarks.

3 CogStream Task and Dataset

Task Setup

As illustrated in Fig. 1, the CogStream task simulates a real-world scenario where users watch an ongoing video stream and continuously interact with a model, asking questions about the content presented so far. Formally, at current time step t, a new video segment v_t is presented, and the cumulative streaming video viewed up to this point is represented by $V_t = \{v_1, \ldots, v_t\}$. The model also maintains the historical dialogue with the user, consisting of previous question-answer (QA) pairs: $QA_{t-1} = \{qa_1, qa_2, \ldots, qa_{t-1}\}$. Concurrently, the user may pose a new question q_t concerning the video content. Answering such a question requires the model to access and integrate both the cumulative video content V_t and the historical textual information QA_{t-1}.

Dataset Overview

To support this task, we introduce a new dataset, which empowers and validates streaming video reasoning capabilities via the QA paradigm. Specifically, as shown in Fig. 2, we categorize all QA pairs into three distinct types based on the *temporal coverage of historical information required for answering*: ***Basic QA*, *Streaming QA*, and *Global QA*.**

Basic QA understands the current video segment v_t from four key aspects: *action* ("What's the man doing?"), *objects* ("What's the girl holding?"), *attributes* ("What style is this hat?"), and *co-reference* (refers back to a specific object

Datasets	Streaming VQA	QA Logical Assoc.	Context Retrieval	Temporal Span	Total QA
REXTIME (Chen et al. 2024b)	✗	✗	✗	-	12,759
NextQA (Xiao et al. 2021)	✗	✗	✗	-	52,044
Video-MME (Fu et al. 2025)	✗	✗	✗	-	2,700
VStream-QA (Zhang et al. 2024b)	✔	✗	✗	-	3,500
STREAMBENCH (Wu et al. 2024)	✔	✔	✔	Long-span	4,500
SVBench (Yang et al. 2025)	✔	✔	✗	Adjacent, Long-span	49,979
CogStream (Ours)	✔	✔	✔	Adjacent, Long-span	**59,032**

Table 1: Comparison of Video Question-Answering Datasets. **QA Logical Assoc.** indicates if a question's answer depends on prior QAs. **Temporal Span** refers to the time distance between associated QA pairs in streaming Video QA, which can be adjacent (consecutive segments) or long-span (distant segments).

mentioned earlier:"How is it used?"). Basic QA provides essential context for subsequent Streaming and Global QA.

Streaming QA is designed based on the nature of streaming video reasoning (Fig. 2), requiring the model to attend to continuously updated visual $V_t = \{v_1, \cdots, v_t\}$ and textual information $QA_{t-1} = \{qa_1, qa_2, \ldots, qa_{t-1}\}$. These questions are designed to assess five distinct capabilities of Vid-LLMs: (1) *Sequence Perception*, requiring the model to reconstruct the chronological evolution of events across segments, based on prior visual-textual context; (2) *Dialogue Recalling*, focusing on the model's capacity to retrieve specific content from historical dialogue; (3) *Dynamic Updating*, where answers must evolve based on the ongoing video stream; (4) *Object Tracking*, challenging the model to recognize and follow the same entity across multiple segments; and (5) *Causal Reasoning*, requiring inference over cumulative visual and textual information to analyze causes or predict outcomes.

Global QA Once the entire video $V_n = \{v_1, \ldots, v_t, \ldots, v_n\}$ has been processed (i.e., stream ends), the model is tasked with reviewing the entire video in conjunction with its associated QA pairs. This review aims to achieve a comprehensive understanding and enable higher-level reasoning. Global QA addresses two tasks: (1) *Global Analysis*: detailed examination of complex topics, events, or underlying meanings within the video, requiring the model to interpret abstract concepts and recognize intricate relationships; (2) *Overall Summary*: synthesizes information from all segments into a coherent summary of the overarching narrative or theme.

Dataset Construction Pipeline

We propose a *semi-automatic* pipeline (Fig. 3) to construct our dataset from unlabeled videos. It comprises four steps: (1) *Video Segmentation:* dividing raw videos into event-based segments; (2) *QA Pairs Generation:* generating various types of QA pairs for each segment; (3) *Relevance QA set Construction:* Identify relevant QA set; (4) *Dialogue Stream Generation:* generating the dialogue stream based on QAs.

Video segmentation To simulate continuous interactions in CogStream, we first divide each video into a series of non-overlapping, event-based segments using the SceneTiling method (Wang et al. 2024b). The timestamp t at the end of each segment v_t serves as an interaction point. To ensure high segmentation quality, we also perform a manual review and refinement process. Details are provided in the Extended version.

QA pairs generation We utilize a Multimodal Large Language Model (MLLM), such as GPT-4o, to generate QA pairs for each video segment based on its visual content. To ensure logical relevance of QA pairs across segments, we introduce a *semantic propagation strategy*. This strategy generates a *title* and *summary* for each segment v_t, serving as contextual priors for the next segment v_{t+1}. Specifically, the MLLM generates QA pairs and a title L_t (representing the segment's theme), refines these QA pairs to eliminate answer-revealing hints for question integrity and challenge, and produces a summary s_t of the refined QA pairs as detailed context. These titles and summaries, collected into sets L_t and S_t up to t, serve as contextual priors for subsequent segments. Formally, QA pairs for v_t are generated as: $qa_t^{Bas.} = \texttt{MLLM}(v_t)$, $qa_t^{Str./Glo.} = \texttt{MLLM}(v_t, L_{t-1}, S_{t-1})$, where $qa_t^{Bas.}$ denotes Basic QA pairs depending only on v_t, $qa_t^{Str./Glo.}$ denotes Streaming/Global QA pairs incorporating prior titles L_{t-1} and summaries S_{t-1} for context. The process iterates over all segments to yield candidate QA set QA for the input video. Further details are in the Extended version.

Relevance QA set construction Next, we establish a *relevance scoring* mechanism to quantify logical and contextual dependency between QA pairs across video segments. Specifically, for each current QA pair $qa_c \in QA_t$, we instruct an MLLM to estimate a relevance score $\mathrm{RS}_{c,p}$ with each previous QA pair $qa_p \in QA_{t-1}$: $\mathrm{RS}_{c,p} = \texttt{MLLM}(qa_c, qa_p)$. This is based on two criteria: (1) *content relevance*, assessing shared content (e.g., objects, events) between qa_c and qa_p, and (2) *logical supportiveness*, evaluating if qa_c extends or builds upon qa_p. The MLLM assigns $\mathrm{RS} \in (0, 7)$ per pair, with higher scores indicating stronger relevance. Only pairs with $\mathrm{RS} > 4$ are appended to the relevant QA set of the current QA. Further details are in the Extended version.

Dialogue stream generation We simulate streaming user interactions by building coherent dialogue streams with strong QA contextual dependencies, using a two-step chronological selection. Initially, two basic QA pairs are randomly added per video segment v_t for foundational understanding. Next, when adding complex QAs (Streaming and Global), we prioritize interdependence with prior QAs. This is achieved via a *Composite Score* that combines relevance to previous QAs with the size of their relevant QA sets. Probabilistic selection based on normalized scores ensures strong logical coherence. Multiple randomized iterations are performed to diversify

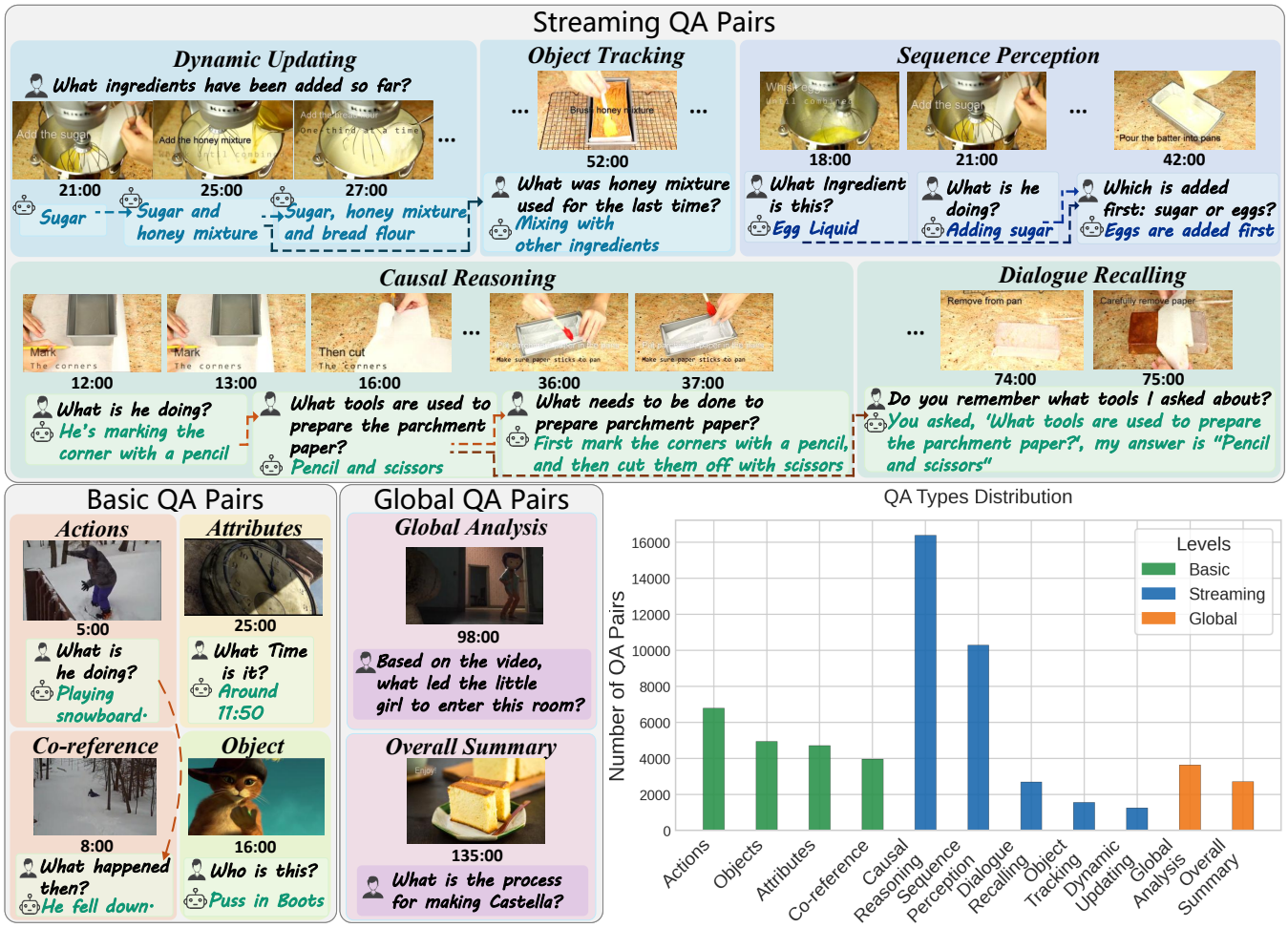

Figure 2: Illustration of different QA settings and type distribution in the dataset. Top: Streaming QA. Bottom-left: Basic QA (left) and Global QA (right). Bottom-right: Distribution of QA types.

Review result	Value
Answer Acc.(%)	86.72
Relevant QA set Acc.(%)	96.35
Avg. T/V (Pipeline)	0.25 h
Avg. T/V (Manual Annotation)	1.45 h

Table 2: Human quality assessment for generated dataset. Relevant QA Set Acc. evaluates QA set relevance by excluding irrelevant sets and including all strongly related ones. Avg. T/V (Pipeline) and Avg. T/V (Manual Annotation) denotes the average time per video for automated QA pair generation and human manual annotation, respectively.

sequence difficulty. See Extended version for full details.

Dataset Analysis

Video source and dataset scale To build our dataset, we collected 6,361 unannotated videos from six public sources: MovieChat (Song et al. 2024) (accounting for 40.2%), MECD (Chen et al. 2024c) (16.8%), QVhighlights (Lei, Berg, and Bansal 2021) (9.8%), VideoMME (Fu et al. 2025) (6.5%), COIN (Tang et al. 2019) (18.0%), and YouCook2 (Zhou, Xu, and Corso 2018) (8.6%). After selecting videos with high-quality annotations, our final dataset comprises **1,088** videos, yielding a total of **59,032** question-answer (QA) pairs. The videos range from 1 to 7 minutes, with some over **10** minutes, and are segmented into **5.02** segments on average via manual annotation. These pairs were derived by sampling and reorganizing an initial set of **58,030** QA pairs, each associated with distinct *relevant QA set* labels. Fig. 2 (right-bottom) presents the QA type distribution of our dataset: Basic (34.6%), Streaming (54.6%), and Global (10.8%). We then allocated **236** of these videos to the testing set and the remaining **852** videos to the training set. Analysis details are in the Extended version.

Manual review To validate the quality of our generated dataset, we conducted human evaluations on a random sample from the dataset. We assessed the generated data in terms of answer accuracy, information completeness, and contextual logical consistency, and provided manual annotations for comparison. As summarized in Tab. 2, the results show a high overlap between the generated and human-annotated

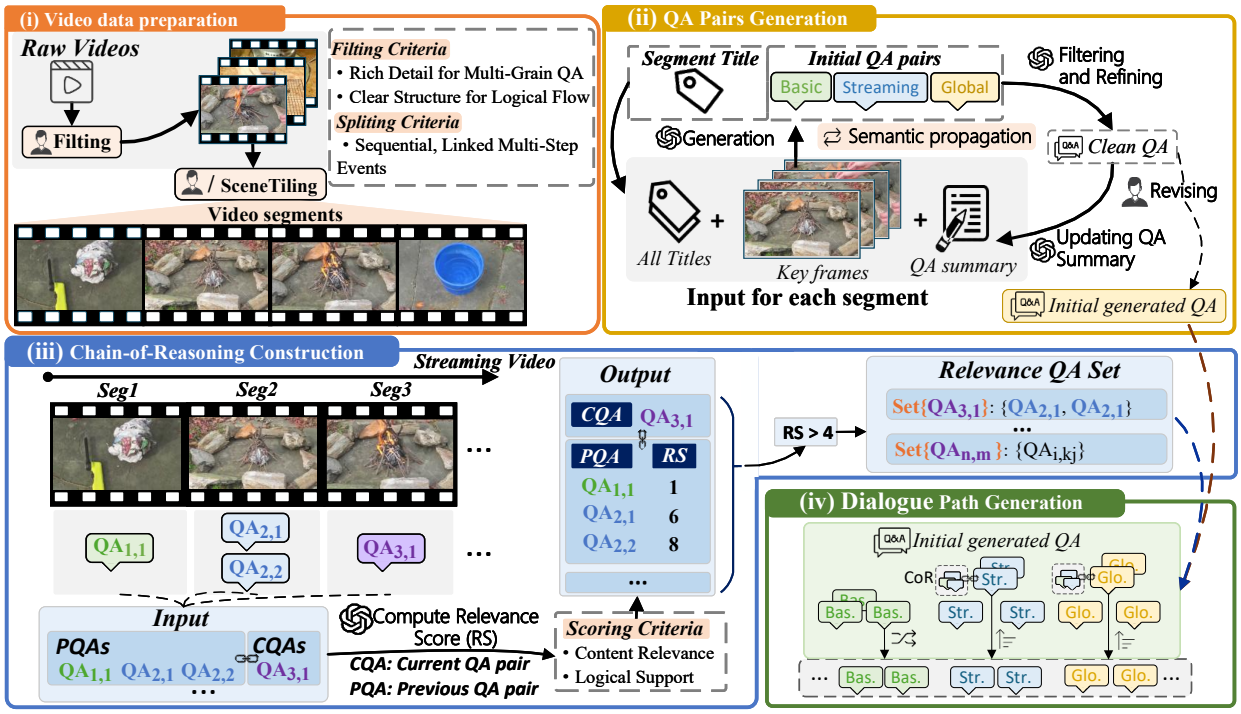

Figure 3: The generation pipeline of CogStream dataset.

data, demonstrating both the efficiency of our semi-automatic pipeline and its ability to maintain high data quality.

4 CogReasoner

In the CogStream task, answering questions necessitates leveraging historical visual and dialogue content. However, as video data accumulates over time, the expanding volume of historical data introduces redundancy and computational burden, leading to inefficient processing and challenges in isolating critical content. To this end, we propose CogReasoner, a baseline designed to effectively (1) ***compress*** accumulated video stream, (2) ***retrieve*** relevant historical QA pairs, and (3) ***reason*** over integrated visual-textual information.

Visual Stream Compression

Temporal-semantic clustering Given sampled frames, they are first encoded and projected into frame-wise features $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ ($\mathbf{x}_i \in \mathbb{R}^{P \times D}$), paired with timestamps $\mathbf{T} = \{t_1, \ldots, t_N\}$. Building upon vanilla K-means, we then devise a temporal-aware clustering algorithm. It groups the features into coherent event representations by employing a composite distance metric, D, designed to jointly model semantic and temporal similarities between frames. Specifically, for each frame feature $\mathbf{x}_i$ at t_i, we compute two distances: the feature distance $d_f(\mathbf{x}_i, \mathbf{c}_j)$ to the j-th cluster centroid $\mathbf{c}_j$, and the temporal distance $d_t(t_i, \tau_j)$ to j-th cluster centroid time τ_j:

$$d_f(\mathbf{x}_i, \mathbf{c}_j) = \|\mathbf{x}_i - \mathbf{c}_j\|, \quad d_t(t_i, \tau_j) = |t_i - \tau_j|. \quad (1)$$

The composite distance D is derived by integrating both semantic and temporal components:

$$D = \sqrt{\mathcal{F}_{nf}(d_f(\mathbf{x}_i, \mathbf{c}_j))^2 + \alpha \mathcal{F}_{nt}(d_t(t_i, \tau_j))^2}, \quad (2)$$

where $\mathcal{F}_{nf}(\cdot)$ and $\mathcal{F}_{nt}(\cdot)$ denote min-max normalization applied along the feature and time dimensions, α controls the relative weight of the temporal component. Finally, based on the composite distance, we perform K-means clustering to produce multiple events. The number of K is set proportionally to the video length. In this way, by jointly considering semantic similarity and temporal coherence, we effectively group original frames into coherent events, providing a structured perception of video for subsequent streaming compression. Ablation studies about K, more details, detailed schematic diagram are presented in the Extended version.

Question-aware streaming compression Treating all historical visual events equally not only increases computational overhead but also impedes reasoning. Our question-aware compression retains highly relevant events while aggressively compressing those of lower relevance. Specifically, for event j covering frame set $X_j = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, we first concatenate all frame-wise features in X_j and then input to the LLM of CogReasoner alongside a carefully designed prompt instructing the model to summarize information of the event. The event embedding $\mathbf{h}_j$ is produced as:

$$\mathbf{h}_j = \texttt{MeanPool}(f_{\text{LLM}}(\texttt{Cat}(\mathbf{x}_1, \ldots, \mathbf{x}_k), \text{Prompt})), \quad (3)$$

$f_{\text{LLM}}(\cdot)$ denotes the output from the final hidden layer of the LLM, $\texttt{Cat}(\cdot)$ is concatenation. Then, we can estimate the relevance s_j between event $\mathbf{h}_j$ and current encoded question $\mathbf{q} \in \mathbb{R}^D$ by cosine similarity. Based on the relevance score, frames belonging to events deemed highly relevant (i.e., $s_j \geq \theta$, where θ is a predefined threshold) are preserved in their original form. Conversely, for events of low relevance, each constituent frame is compressed into a single

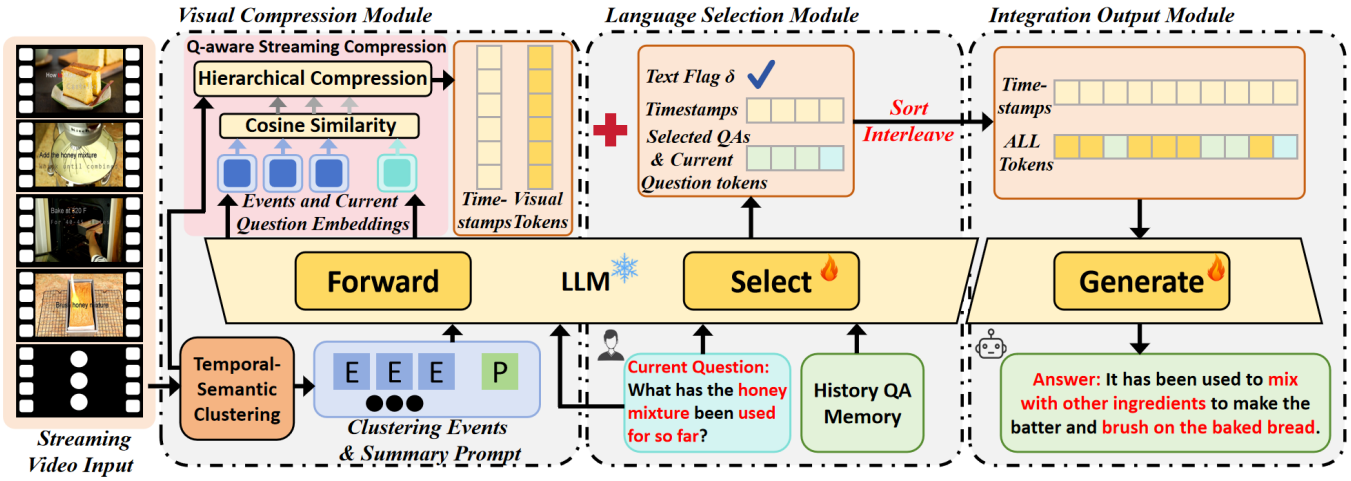

Figure 4: The overview of CogReasoner. It comprises three modules: the **Visual Stream Compression** uses Temporal-Semantic Clustering and Question-aware Streaming Compression to process video streams into relevant events; the **Historic Dialogue Retrieval** employs an LLM to select relevant historical QA pairs and assess visual input necessity; the **Video-text Interleave Reasoning** interleaves visual and textual tokens time-sequentially for answer generation.

token representation using average pooling. This strategy enables the model to focus on highly pertinent events and significantly reduces computational overhead. Meanwhile, preserving summary representations for these low-relevance events maintains the video's overall temporal coherence. Further ablation studies about θ are presented in the Extended version.

Historic Dialogue Retrieval

As current LLM-based reasoning heavily relies on text, often prioritizing it over visual context, irrelevant historical text will impair response accuracy. Moreover, in streaming dialogues, some purely linguistic questions can be resolved solely using historical QA pairs (e.g., '*Dialogue Recalling*' task in our dataset). Forcing visual processing in these cases introduces unnecessary interference and computational costs. Therefore, we design a ***Historic Dialogue Retrieval*** mechanism to (1) identify and select relevant historical QAs and (2) determine if the question can be resolved using only textual information, thus avoiding unnecessary visual processing. Specifically, we employ an LLM to select relevant historical QA pairs and determine whether the current question is purely textual in nature. Formally, given a current question q_t and a set of historical QA pairs $QA_{t-1} = \{qa_1, qa_2, \ldots, qa_{t-1}\}$, both are fed into the LLM: $QA_{retrieved}, \delta = \texttt{LLM}(QA_{t-1}, q_t)$, where indicator $\delta \in \{0, 1\}$ indicates whether q_t is a purely linguistic question that can be answered solely based on the historical QA pairs QA_{t-1}; $QA_{retrieved}$ denotes the subset of historical QA pairs that are relevant to the current question.

Video-text Interleave Reasoning

Given the compressed visual stream and retrieved textual information, ***Video-text Interleave Reasoning*** deduces the final answers. For a current question, we interleave and concatenate the visual $V_{\text{compressed}}$ and textual $QA_{\text{retrieved}}$ inputs in a temporally ordered manner to form the input sequence. For questions that only require textual information $QA_{\text{retrieved}}$, the corresponding visual stream is omitted. This process is formulated as: $a_t = \texttt{LLM}(V_{\text{compressed}}, QA_{\text{retrieved}}, \delta, q_t)$, where indicator δ is computed during the dialogue retrieval, a_t denotes the model's output answer to q_t.

Overall Training

CogReasoner is fine-tuned in two stages, centered around a shared LLM. First, to enhance Historic Dialogue Retrieval, this LLM is fine-tuned with a task-specific LoRA adapter on 100k+ text-only historical QA and current question combinations for precise QA selection. Second, for Video-text Interleave Reasoning, the visual encoder is frozen, while the projection layer and the same LLM are fine-tuned, using a distinct LoRA adapter for end-to-end reasoning. This stage uses 48k+ QA pairs across 800+ videos, incorporating ground-truth preceding QA pairs and their selection status. Further details about training and inference are in the Extended version.

5 Experiments

Experimental Setup

Baselines Recent methods for streaming video understanding are selected for comparison, including Flash-VStream (Zhang et al. 2025b) and ReKV (Di et al. 2025). Besides, state-of-the-art (SOTA) Vid-LLMs are selected, including: *Open-source models*: LongVA (Zhang et al. 2024c), InternVL2 (Chen et al. 2024d), VideoLLaMA2 (Cheng et al. 2024), VideoLLaMA3 (Zhang et al. 2025a), MiniCPM-V 2.6 (Yao et al. 2024), and MiniCPM-o 2.6 (Team 2025). *Proprietary models*: Gemini-1.5 Pro (Team et al. 2024), Qwen2-VL-Max (Wang et al. 2024a), and GPT-4o (Achiam et al. 2023). All experiments are conducted on CogStream test set. To align with the QA generation pipeline (sampling 20 frames per segment), all

Method	Prm.	Frm.	Basic				Streaming					Global		Avg.↑
			Att.	Obj.	Co-ref.	Act.	Rea.	Seq.	Dial.	Dyn.	Obj.	Over.	Glob.	
Open-Source Models														
InternVL2	7B	12/seg	52.3	59.0	36.6	36.3	52.6	41.9	39.2	39.1	49.3	52.4	59.8	48.66
LongVA	7B	12/seg	63.6	55.0	42.0	33.6	53.1	40.9	55.4	25.3	40.1	42.4	53.3	48.76
VideoLLaMA2	7B	20/seg	60.0	61.7	47.8	46.4	47.5	47.4	54.1	30.2	62.3	54.3	54.8	50.72
MiniCPM-o 2.6	8B	20/seg	77.3	76.4	63.6	60.6	65.9	61.0	47.1	50.9	44.5	57.4	62.8	64.08
VideoLLaMA3	7B	1fps	75.7	71.8	62.6	64.6	67.7	61.5	56.9	52.4	69.1	66.0	72.3	66.52
MiniCPM-V-2.6	8B	20/seg	78.6	73.6	70.7	59.6	70.5	59.7	50.0	49.2	60.6	64.2	69.4	66.84
Flash-VStream	7B	1fps	53.1	41.2	41.8	43.3	46.9	37.2	23.1	8.9	49.9	19.1	26.5	40.58
ReKV	7B	1fps	51.1	54.2	40.1	43.6	46.3	37.9	51.0	13.9	39.3	26.7	33.8	43.18
CogReasoner	7B	1fps	74.8	71.5	62.9	64.6	66.3	68.0	66.3	50.8	71.0	70.9	64.6	67.32
VideoLLaMA3†	7B	1fps	**79.9**	75.0	68.4	65.0	**70.8**	68.4	63.1	66.5	**82.8**	74.0	72.1	70.70
CogReasoner†	7B	1fps	77.3	**78.9**	**74.6**	**70.0**	69.7	**68.8**	**83.4**	**70.5**	74.8	**75.4**	**76.0**	**72.26**
Proprietary Models														
Gemini 1.5 Pro	-	20/seg	75.5	73.4	66.4	62.5	66.2	61.1	64.1	42.0	35.1	69.4	74.4	66.04
Qwen2-VL-Max	-	50(max)	77.2	**76.7**	**70.4**	**69.2**	76.7	66.5	62.3	**53.7**	52.0	76.2	76.6	72.58
GPT-4o	-	20/seg	**78.4**	73.9	68.2	66.1	**77.5**	**72.1**	**73.0**	52.4	**74.0**	**77.0**	**79.6**	**73.90**

Table 3: Performance metrics of different models in 11 *CogStream* capabilities. Prm. denotes the number of model parameters, Frm. denotes the number of sampled frames. Models denoted by † were fine-tuned on our training set; all other results are zero-shot.

the baselines process video segments using this standardized frame rate, with two exceptions: (1) CogReasoner and VideoLLaMA3 employ their efficient compression to handle sparser 1fps inputs, and (2) Qwen2-VL-Max operates up to 50 frames due to its access constraints. All open-source models are deployed with 16-bit precision.

Evaluation Inspired by SVbench (Yang et al. 2025), we enhance the LLM-based VQA metric (GPT4-score (Muhammad Maaz and Salman Khan 2023)) for evaluation. We introduce the following metrics: Information Accuracy (IA), Detail Completeness (DC), Context Awareness (CA), Temporal Precision (TP), and Logical Consistency (LC). Each is scored between 0 and 100, and we report their average. See our Extended version for details of these metrics.

Implementation VideoLLaMA3 (Zhang et al. 2025a) served as our baseline, comprising VL3-SigLIP-NaViT (vision encoder), an MLP projection layer, and Qwen2.5 (language model). To ensure fairness, we evaluate models under two settings: **Zero-shot** and **Fine-tuned** (denoted with an †). In the zero-shot setting, all baselines are evaluated directly. Our CogReasoner is also evaluated zero-shot, with the necessary exception that its dialogue retrieval module is pre-trained to ensure meaningful performance (see Tab. 7). For the fine-tuned setting, both CogReasoner† and the VideoLLaMA3† baseline are trained on our dataset for a direct comparison. All models use carefully designed prompts to meet task requirements. Further details are in the Extended version.

Main Results

Comparative results on streaming VQA Tab. 3 compares our framework against SOTA methods. It demonstrates that our framework achieves strong results across all three tasks: Basic (**Bas.**), Streaming (**Str.**), and Global (**Glo.**). Thanks to the design of our Visual Stream Compression, we observe notable improvements in the Basic task. More importantly, the model shows substantial gains on the context-dependent Streaming and Global tasks, where accurate contextual understanding is essential for effective reasoning. This effectiveness benefits from the precise selection of textual context by our Historic Dialogue Retrieval, synergizing with the Visual Stream Compression that provides refined visual inputs and the Video-text Interleave Reasoning that ensures their effective fusion. Further visualization results are in the Extended version.

Visual	Bas.	Str.	Glo.	Avg.↑
Baseline	73.80	68.50	73.90	70.64
Only Selection	**75.40**	67.50	73.10	70.52
No Clustering	73.80	69.80	75.10	71.48
Ours	75.30	**70.20**	**75.70**	**72.26**

Table 4: Ablation study on visual stream compression.

Con.	Bas.	Str.	Glo.	Avg.↑
No context	75.60	62.30	61.80	66.04
All context	75.20	67.90	72.50	70.48
Retrieved (***Ours***)	75.30	70.20	75.70	72.26
Ground-truth	**75.90**	**77.30**	**82.30**	**77.40**

Table 5: Different historical context strategies.

Ablation Study

Effectiveness of visual stream compression Our ablation study (Tab. 4) validates the key components of our visual compression module. We find that removing temporal-semantic clustering (*No Clustering*) significantly degrades perfor-

Method	Con.	Rat.	Bas.	Str.	Glo.	Avg.↑
VideoLLaMA3	All Context	30%	65.90	62.40	68.40	65.02
		15%	68.30	63.50	69.00	65.48
		0%	69.70	64.50	69.10	66.52
Ours	All Context	30%	73.40	68.20	68.30	69.74
		15%	71.00	68.70	73.50	69.82
		0%	75.20	67.90	72.50	70.48
Ours	Retrieved	30%	74.60	70.40	74.10	71.94
		0%	75.30	70.20	**75.70**	72.26
		15%	**75.70**	**70.50**	74.50	**72.40**

Table 6: Comparison between context strategies under different interference QA ratios ('Rat.')

Method	HDR	Bas.	Str.	Glo.	Avg.↑
VideoLLaMA2	+Itself	54.80	49.00	47.90	50.52
	-	55.00	48.00	54.50	50.72
	+Ours	**57.70**	**53.30**	**58.50**	**55.16**
MiniCPM-V-2.6	+Itself	70.60	63.80	61.40	65.46
	-	70.90	65.20	**66.80**	66.84
	+Ours	**72.30**	**66.20**	65.90	**67.78**

Table 7: Performance comparison of different Vid-LLMs with our proposed historical dialogue retrieval (HDR).

mance by losing temporal and semantic coherence. Furthermore, summarizing less-relevant visual information (Ours) proves superior to completely discarding it (*Only Selection*), which underscores the importance of maintaining the video's overall continuity for robust reasoning.

Comparison of historical context handling strategies We evaluate our dialogue retrieval strategy (*Retrieved*) in Tab. 5, comparing it against baselines that use no context (*No Context*) or all available context (*All Context*), and an upper-bound with ground-truth context (*GT*). Our method significantly outperforms both baselines, confirming that while historical context is crucial, indiscriminately including all of it degrades performance. This result validates the importance of selective context retrieval. The performance gap to the *GT* upper-bound suggests that further improvements to the retrieval module remain a focus for future work.

Generalizability of the historic dialogue retrieval To test the generalizability of our Historic Dialogue Retrieval (HDR), we integrated it with VideoLLaMA2 and MiniCPM-V-2.6 (Tab. 7). The results show that when these baseline models use their own LLM for retrieval (*+Itself*), their performance degrades below their original configuration. In stark contrast, applying our trained HDR module (*+Ours*) yields consistent performance improvements. This demonstrates that our HDR is an effective and generalizable module that enhances video reasoning across different Vid-LLM architectures.

Robustness evaluation We assess our model's robustness to dialogue noise by introducing irrelevant QA pairs at varying interference ratios, with results in Tab. 6. Our retrieval-enabled CogReasoner demonstrates strong resilience, with performance remaining largely stable as noise increases. In stark contrast, models processing all dialogue history (*All Context*) exhibit a significant performance degradation. This result underscores the critical role of our retrieval mechanism in filtering irrelevant context to maintain reasoning quality.

6 Conclusion

This paper proposes a challenging task, CogStream, which requires high-quality context for streaming video reasoning. To achieve this, we constructed a dataset based on a semi-automatic generation pipeline. Extensive experiments demonstrate that state-of-the-art Vid-LLMs struggle to efficiently acquire relevant context and achieve strong performance in this task. Therefore, we devised a baseline for CogStream, which achieves highly competitive performance.

Acknowledgments

The paper is supported in part by the National Natural Science Foundation of China (No. U21B2013 , 62325109), and in part by the Shanghai 'The Belt and Road' Young Scholar Exchange Grant (24510742000).

We thank the reviewers for their insightful comments and valuable suggestions.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, J.; Lv, Z.; Wu, S.; Lin, K. Q.; Song, C.; Gao, D.; Liu, J.-W.; Gao, Z.; Mao, D.; and Shou, M. Z. 2024a. VideoLLM-online: Online Video Large Language Model for Streaming Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18407–18418.

Chen, J.-J.; Liao, Y.-C.; Lin, H.-C.; Yu, Y.-C.; Chen, Y.-C.; and Wang, Y.-C. F. 2024b. Rextime: A benchmark suite for reasoning-across-time in videos. *arXiv preprint arXiv:2406.19392*.

Chen, T.; Liu, H.; He, T.; Chen, Y.; Gan, C.; Ma, X.; Zhong, C.; Zhang, Y.; Wang, Y.; Lin, H.; et al. 2024c. MECD: Unlocking Multi-Event Causal Discovery in Video Reasoning. *arXiv preprint arXiv:2409.17647*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; and Bing, L. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Di, S.; Yu, Z.; Zhang, G.; Li, H.; Cheng, H.; Li, B.; He, W.; Shu, F.; Jiang, H.; et al. 2025. Streaming Video Question-Answering with In-context Video KV-Cache Retrieval. In *ICLR*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24108–24118.

Lei, J.; Berg, T.; and Bansal, M. 2021. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *URL https://arxiv.org/abs/2107.09609*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, Y.; Yu, R.; Yin, F.; Zhao, X.; Zhao, W.; Xia, W.; and Yang, Y. 2022. Learning quality-aware dynamic memory for video object segmentation. In *European Conference on Computer Vision*, 468–486. Springer.

Luvizon, D. C.; Picard, D.; and Tabia, H. 2020. Multi-task deep learning for real-time 3D human pose estimation and action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2752–2764.

Muhammad Maaz, H. R.; and Salman Khan, F. S. K. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424*.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Qian, R.; Dong, X.; Zhang, P.; Zang, Y.; Ding, S.; Lin, D.; and Wang, J. 2024. Streaming long video understanding with large language models. *arXiv preprint arXiv:2405.16009*.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Tang, Y.; Ding, D.; Rao, Y.; Zheng, Y.; Zhang, D.; Zhao, L.; Lu, J.; and Zhou, J. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1207–1216.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Team, O. M.-o. 2025. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; Xie, C.; Liu, Y.; and Zheng, Z. 2024b. Videollamb: Long-context video understanding with recurrent memory bridges. *arXiv preprint arXiv:2409.01071*.

Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; and Wang, S. 2020. Towards real-time multi-object tracking. In *European conference on computer vision*, 107–122. Springer.

Wu, C.-K.; Tam, Z. R.; Lin, C.-Y.; Chen, Y.-N. V.; and Lee, H.-y. 2024. Streambench: Towards benchmarking continuous improvement of language agents. *Advances in Neural Information Processing Systems*, 37: 107039–107063.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xiong, H.; Yang, Z.; Yu, J.; Zhuge, Y.; Zhang, L.; Zhu, J.; and Lu, H. 2025. Streaming Video Understanding and Multi-round Interaction with Memory-enhanced Knowledge. *arXiv preprint arXiv:2501.13468*.

Yang, Z.; Hu, Y.; Du, Z.; Xue, D.; Qian, S.; Wu, J.; Yang, F.; Dong, W.; and Xu, C. 2025. SVBench: A Benchmark with Temporal Multi-Turn Dialogues for Streaming Video Understanding. *arXiv preprint arXiv:2502.10810*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025a. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; and Wang, H. 2016. Real-time action recognition with enhanced motion vector CNNs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2718–2726.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, H.; Wang, Y.; Tang, Y.; Liu, Y.; Feng, J.; Dai, J.; and Jin, X. 2024a. Flash-VStream: Memory-Based Real-Time Understanding for Long Video Streams. *arXiv preprint arXiv:2406.08085*.

Zhang, H.; Wang, Y.; Tang, Y.; Liu, Y.; Feng, J.; Dai, J.; and Jin, X. 2024b. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*.

Zhang, H.; Wang, Y.; Tang, Y.; Liu, Y.; Feng, J.; and Jin, X. 2025b. Flash-VStream: Efficient Real-Time Understanding for Long Video Streams. *arXiv preprint arXiv:2506.23825*.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024c. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Zhou, L.; Xu, C.; and Corso, J. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.