VideoINSTA: Zero-shot Long Video Understanding via Informative Spatial-Temporal Reasoning with LLMs

Ruotong Liao[1,2,*], **Max Erler**[1,*], **Huiyu Wang**[3], **Guangyao Zhai**[2,3], **Gengyuan Zhang**[1,2], **Yunpu Ma**[1,2,4,†], **Volker Tresp**[1,2]

[1]LMU Munich [2]Munich Center for Machine Learning (MCML)
[3]Technical University of Munich [4]Siemens AG

ruotong.liao@outlook.com, cognitive.yunpu@gmail.com
volker.tresp@lmu.de

Abstract

In the video-language domain, recent works in leveraging zero-shot Large Language Model-based reasoning for video understanding have become competitive challengers to previous end-to-end models. However, long video understanding presents unique challenges due to the complexity of reasoning over extended timespans, even for zero-shot LLM-based approaches. The challenge of information redundancy in long videos prompts the question of what specific information is essential for large language models (LLMs) and how to leverage them for complex spatial-temporal reasoning in long-form video analysis. We propose a framework **VideoINSTA** , i.e. **IN**formative **S**patial-**T**empor**A**l Reasoning for zero-shot long-form video understanding. **VideoINSTA** contributes (1) a zero-shot framework for long video understanding using LLMs; (2) an event-based temporal reasoning and content-based spatial reasoning approach for LLMs to reason over spatial-temporal information in videos; (3) a self-reflective information reasoning scheme based on information sufficiency and prediction confidence while balancing temporal factors. Our model significantly improves the state-of-the-art on three long video question-answering benchmarks: EgoSchema, NextQA, and IntentQA, and the open question answering dataset ActivityNetQA. Code is released here.

1 Introduction

Large language models (LLMs) have demonstrated remarkable reasoning abilities, even in long-context situations (Chen et al., 2024; Mao et al., 2023; Kojima et al., 2022). These advancements have spurred interest in video reasoning. Previous works bridging video and text modalities depend on meticulously designed models suffering large-scale pretraining. This challenge is pronounced with videos, a data format characterized by a vast volume of information scaling with length. Consequently, these models exhibit limited generalizability across datasets and struggle to scale to long video within a single model (Sun et al., 2019; Yang et al., 2022). More recent models have gradually integrated LLMs' reasoning abilities by introducing lightly tuned adaptation layers (Yang et al., 2022; Zhang et al., 2023b; Lin et al., 2023a). However, they still struggle with the length of the videos. Recently, to avoid expensive training costs, early attempts have proposed a zero-shot solution by reasoning over semantic representations of video content using LLMs (Zhang et al., 2023a; Wang et al., 2024a; Choudhury et al., 2023). These approaches have become strong competitors to earlier end-to-end models. Nonetheless, long-form video understanding, which demands advanced reasoning over extended timespans, remains challenging even for LLM-based methods.

Even in light of these tryouts, many challenges remain unsolved: (1) *Information Quality*. Videos contain vast information even with some redundancy due to minor visual changes. Identifying the most crucial piece of information and extracting it effectively is essential to enhance the quality of data within the context window manageable by LLMs. How can we achieve this extraction? (2) *Neglect of Spatial and Temporal Characteristics*. Videos inherently exhibit temporal and spatial characteristics. How can we effectively preserve and convey this spatial-temporal information to support LLM reasoning? Especially, how do LLMs process temporal dynamics in videos? (3) *Complexity of Reasoning with Unbalanced Information over Temporal Span*. In long videos, the significance of information along the video temporal axis varies greatly. LLMs' implicit "intuition" to process all the information is insufficient. How do we develop an explicit reasoning algorithm for unbalanced information considering temporal factor?

[*] Equal contribution.
[‡] Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 6577–6602
November 12-16, 2024 ©2024 Association for Computational Linguistics

To address these challenges, we propose a framework **VideoINSTA**, i.e. **IN**formative **S**patial-**T**empor**A**l reasoning for zero-shot long-form video understanding, aiming to build a compound system extracting essential information from long-form videos – leveraging spatial-temporal reasoning and temporal-aware self-reflective reasoning to handle complex information with LLMs.

VideoINSTA is a zero-shot framework for reasoning with LLMs, augmented with visual-language tools. First, this framework emphasizes *event-based temporal reasoning* by proposing an automatic temporal segmentation method C-DPCKNN, which segments long videos into multiple events. Besides, it derives the global temporal information with the help of a unified temporal representation tool UniVTG (Qinghong Lin et al., 2023) and utilizes a temporal grounding scheme allowing the event to inherit the local temporal information. Second, this framework emphasizes *content-based spatial reasoning* by improving video captions with various visual-language captioning tools to extract richer spatial information. Specifically, event captioning is compensated by object detection and action caption as spatial information. A follow-up summarization serves as implicit spatial reasoning in a chain-of-thought manner. Third, this framework proposes *Iterative Information Reasoning* with LLMs, which iteratively merges the temporal and spatial information derived in the previous stages based on the self-evaluation of LLMs on the information sufficiency and prediction confidence.

Experiments have showcased remarkable improvements in existing long-form video question-answering tasks compared to end-to-end video-language models as well as other zero-shot LLM-based video understanding compound systems. Besides, VideoINSTA handles long videos with an average length of 3 minutes and is easily extensible for longer videos in a zero-shot manner. This framework also shows excellent results both on multi-choice and open-question answering tasks. The main contributions are summarized as follows:

- **VideoINSTA: A zero-shot framework for long-form video understanding with state-of-the-art performance.** We propose a new zero-shot and extensible framework based on LLMs augmented with visual-language tools.
- **Spatial-temporal reasoning on videos with LLMs.** We propose event-based temporal reasoning and content-based spatial reasoning with LLMs utilizing extracted spatial-temporal information for understanding long-form videos.
- **Self-reflective information reasoning with LLMs considering temporal factors.** Our framework contributes to an iterative reasoning scheme for LLMs to merge and reason on the spatial-temporal information in a self-reflective manner while considering the temporal factors.

2 Related Works

Video Question Answering with LLMs Long video question answering involves predicting the correct answer given videos and queries, and optional multi-choice options. With advancements in LLMs and their long-context reasoning abilities, video understanding using LLMs has been explored in various works (Xu et al., 2023; Maaz et al., 2023; Jin et al., 2024a; Yu et al., 2024; Lin et al., 2023b; Zhang et al., 2023c; Huang et al., 2024; Wang et al., 2023a). However, even with lightly tuned adaptation layers, scaling training costs increase significantly with video length. Recently, zero-shot methods like (Wang et al., 2022b) use image descriptors for video understanding tasks. Besides, LLoVi (Zhang et al., 2023a) and VideoAgent (Wang et al., 2024a), which use extensive captioning and iterative keyframe selection respectively, have aimed to achieve training-free video understanding. Additionally, works such as ProViQ (Choudhury et al., 2023) and MoReVQA (Min et al., 2024) investigate zero-shot understanding using neuro-symbolic programming. LangRepo (Kahatapitiya et al., 2024a) has a structured language repository to maintain textual video representations. TraverLER (Shang et al., 2024) iteratively gathers relevant information from keyframes with multiple LLMs and VideoTree (Wang et al., 2024b) is an extension of LLoVi with tree-based information searching scheme. Unlike these approaches, we allows LLMs to directly reason on extracted spatial-temporal information without neuro-symbolic programming.

Spatial-Temporal Reasoning on Video Spatial-temporal reasoning in video has been a topic of continuous discussion (Hussein et al., 2019; Wang et al., 2021; Xiao et al., 2023; Wu et al., 2021; Zhu et al., 2022; Jin et al., 2024a; Li et al., 2022; Xiao et al., 2022, 2024; Zhai et al., 2020) due to the dual characteristics of video data. Most previous approaches compress information and perform reasoning within the embedding space. Additionally, recent works have highlighted LLMs' capabilities in temporal (Tan et al., 2023; Han et al., 2023; Yuan et al., 2024; Liao et al., 2024; Ding et al., 2024; Xiong et al., 2024) and spatial reasoning (Ranasinghe et al., 2024b; Wu et al., 2024b; Ko et al., 2023; Sharma et al., 2024; Yamada et al., 2023; Wu et al., 2024a). However, applying LLMs' spatial-temporal reasoning abilities to video remains underexplored. Our work innovatively harnesses these abilities, augmenting them with spatial-temporal reasoning methods as tools, to effectively analyze long-form videos both spatially and temporally.

3 VideoINSTA: Informative Spatial-Temporal Reasoning with Large Language Models

In this section, we explain our **VideoINSTA** framework shown in Figure 1 following its three-phase methodology: event-based temporal reasoning, content-based spatial reasoning, and self-reflective information reasoning with LLMs.

3.1 Event-based Temporal Reasoning

The event-based temporal reasoning, as shown in Figure 2, consists of two sequential sub-steps differentiated by whether the query Q is a known, specifically, query-agnostic temporal segmentation and query-aware temporal grounding.

3.1.1 Query-agnostic Temporal Segmentation

KNN (Guo et al., 2003) Clustering has been a widely used algorithm for temporal segmentation for separating event clips in video. For example, (Zhou et al., 2024) utilizes KNN and ChatUniVi (Jin et al., 2024a) utilizes DPCKNN (Du et al., 2016), a density-based clustering algorithm to merge frames belonging to the same events. However, these methods are designed specifically for embedding-based reasoning. They share a common fallback that frames or even tokens belonging to the same cluster scatter across the video span, causing blended boundaries between events, and frames from different events are interleaved thus not fulfilling the temporal order. Therefore, we propose a *consecutive* clustering algorithm **C-DPCKNN** for automatic event parsing on videos with clear boundaries.

Event Center Given a i^{th} frame in a video, we first use the vision encoder of CLIP (Radford et al., 2021) to provide its visual tokens $\boldsymbol{Z} = \{z_i\}_{i=1}^{L}$, where L is the number of visual tokens within each frame. Then we apply mean-pooling over all tokens to obtain the frame-level representation f_i. Specifically, we first compute the local density $\rho_m i$ as Eq. 1. Then we compute the distance index δ_i as Eq. 2 of each frame f_i. We set frames with the highest $\rho_i \times \delta_i, i \in [1, 2, \ldots, M]$ as cluster centers, where M is the total sampled frames in a video.

$$\rho_i = \exp\left(-\frac{1}{K} \sum_{z_k \in \mathrm{KNN}(z_i, \boldsymbol{Z})} \|z_k - z_i\|^2\right) \quad (1)$$

$$\delta_i = \begin{cases} \min_{j:\rho_j > \rho_i} \|z_j - z_i\|^2, & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i \\ \max_j \|z_j - z_i\|^2, & \text{otherwise.} \end{cases} \quad (2)$$

Event Clustering Given K cluster centers, we cluster consecutive frames in both, forward and backward directions. We deprecate setting other frames directly to their nearest cluster center based on Euclidean distances of the embeddings which causes interleaved event frames and blurred boundaries that are counterintuitive to how events are separated and sequenced in an untrimmed video. Instead, we set the event boundary according to the critical points with the $K-1$ minimum density values, i.e. minimum density peaks $\boldsymbol{\Delta} = \{\delta_i\}_{i=1}^{K-1}$, indicating drastic changes in the frame content and denote the set of indexes of the frames in the cluster as E. We treat each cluster as a critical event and parse the events consistent with the frame order.

Event Segmentation To set clear boundaries for each event, we store the indexes of boundary frames with $K-1$ minimum density peaks as $\boldsymbol{\mathcal{I}} = \{I_i\}_{i=1}^{K-1}$ to set the event set $\boldsymbol{\mathcal{E}} = \{E_i\}_{i=1}^{K}$ with respective starting and ending boundaries $\{(0, I_1), \ldots, (I_{K-1}, I_{EOV})\}_{i=1}^{K-1}$, I_{EOV} denotes the ending index of video. The video is then parsed into respective event clips.

3.1.2 Query-agnostic Temporal Grounding

Aside from automatic query-agnostic temporal segmentation, we introduce query-aware temporal grounding – providing semantic temporal representations to support richer informative reasoning.

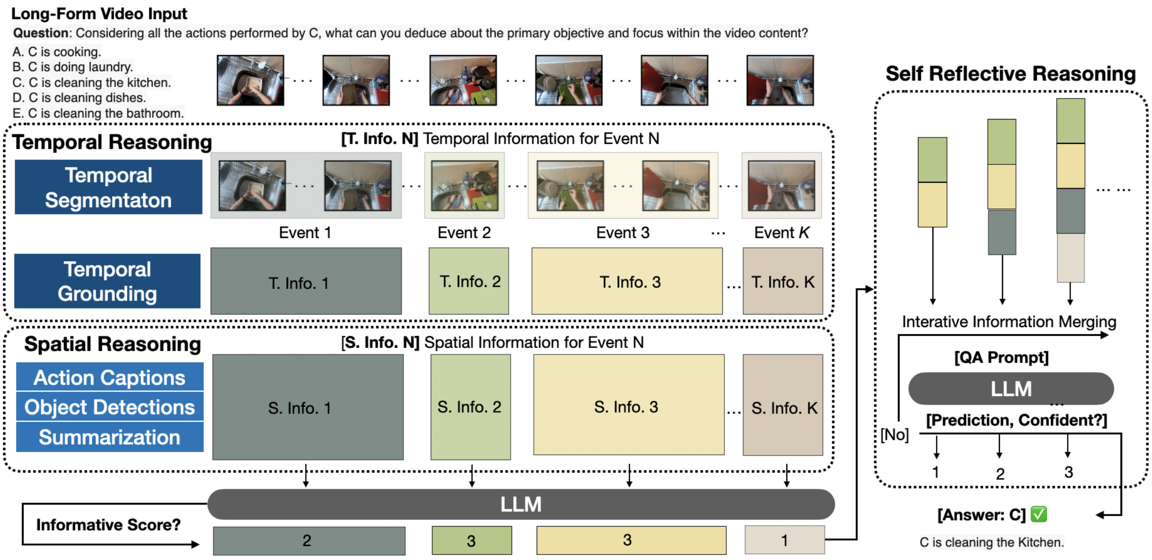

Figure 1: Framework of **VideoINSTA**. VideoINSTA consists of three phases. (1) Event-based Temporal Reasoning. Temporal Segmentation parses the video into events via proposed C-DPCKNN clustering, and Temporal Grounding derives semantic temporal information inherited from the global relevance of each event. (2) Content-based Spatial Reasoning. Action Captions are derived for each clip by video captioners as basic spatial information. Compensated with Object Detections, the spatial information is summarized to derive query-focused spatial information. (3) Self-reflective Information Reasoning. The previously derived spatial-temporal information is merged according to their information sufficiency in descending order and the LLM performs multi-round predictions after information merging until it comes to a confident self-evaluation.

Global Temporal Relevance Derivation We first derive the initial global temporal information, specifically, the relevance of the whole video given the query, with the help of the zero-shot unified video-language temporal grounding model UniVTG (Qinghong Lin et al., 2023). Given a video V and a question query Q, UniVTG divides the original V into fine-granular clips $V = \{v_i\}_{i=1}^{L_v}$ and evaluates each v_i with triple evaluators $(f_i, b_i, s_i)_{i=1}^{L_v}$, where L_v is the number of fine-grained clips. $s_i \in [0, 1]$ are continuous salience scores determining the relevance between the visual content of the video and the query Q spanning from totally irrelevant to highly correlated; f_i are the foreground indicators for query-based moment retrieval, and b_i are the boundary intervals for moment localization.

Local Temporal Relevance Inheritance As UniVTG derives global temporal relevance information for the whole video, we propose *Local Inheritance* which assigns query-aware global temporal relevance information to the automatically and query-agnostic parsed event clips $\mathcal{E} = \{E_i\}_{i=1}^{K}$ as local temporal relevance information. Specifically, a boundary-based inheritance scheme is performed. We rank fine-grained clips $\{v_i\}_{i=1}^{L_v}$ with predicted boundaries $\{b_i\}_{i=1}^{L_v}$ based on their $\{f_i\}_{i=1}^{L_v}$ probabilities and returns the Top-k clips as query-aware moment retrieval predictions and return their boundaries $\{b_i\}_{i=1}^{k}$ given a question V and a query Q. Then, we take boundary intersections between $\mathcal{I}$ and $\{b_i\}_{i=1}^{k}$ and calculated the percentage of $\{b_i\}_{i=1}^{k}$ allocated in each event E_i. The relevance percentage is translated into semantic representations for LLMs to reason. Hence, the temporal information is transformed as prompt $\mathcal{P}^t$.

3.2 Content-based Spatial Reasoning

The second phase of VideoINSTA contributes spatial reasoning with spatial information extraction. A common bottleneck from previous works on LLM-based video understanding is the redundant and inaccurate information in describing videos, especially overloading the LLMs' context window when processing long videos. It is necessary to address the importance of information density of the spatial information for LLMs to reason, especially for long-form videos. VideoINSTA shows that actions and objects occurring in the videos are the most crucial components. For each event clip in $\mathcal{E} = \{E_i\}_{i=1}^{K}$, we derive informative prompts with action captions $\mathcal{P}^a = \{P_i^a\}_{i=1}^{K}$ and object captions $\mathcal{P}^o = \{P_i^o\}_{i=1}^{K}$, detailed as follows.

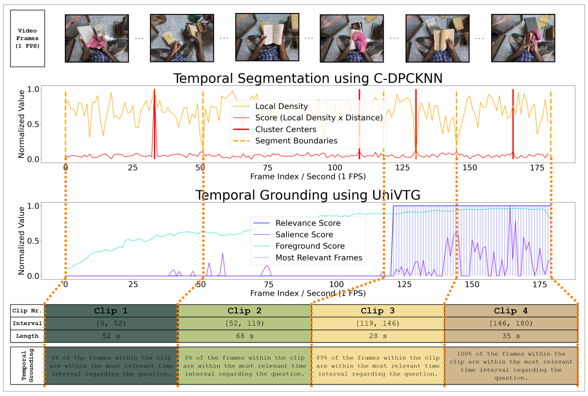

Figure 2: Illustration of Temporal Reasoning in **VideoINSTA**. In Temporal Segmentation, the proposed C-DPCKNN sets clear borders with minimum density peaks. In Temporal Grounding, each event inherits the global relevance information derived from UniVTG according to these borders. The inherited local temporal information is transformed into semantic prompts, empowering temporal reasoning in **VideoINSTA**.

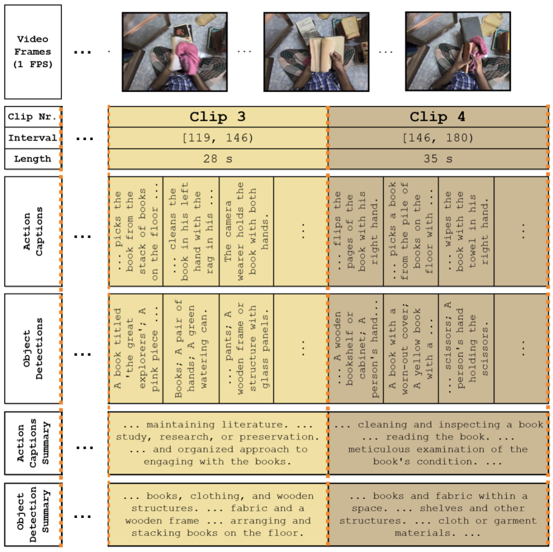

Figure 3: Spatial Reasoning in **VideoINSTA**.

3.2.1 Action Captioning

We leverage generative visual-language models (VLMs) to convert the video context to language descriptions. To ensure zero-shot quality of the extracted spatial information and as a fair comparison to other approaches, we utilize LaViLa (Zhao et al., 2023a) – pre-trained on Ego4D dataset (Grauman et al., 2022), following (Zhang et al., 2023a) – on ego-centric videos, to create automatic video narrations. The auto-generated narrations densely cover long videos while reserving temporal synchronization of the visual information and descriptions of the video actions within the event clip. For exocentric videos, we follow (Wang et al., 2024a) utilizing CogAgent (Hong et al., 2024) to provide descriptions of the sequential video frames with a special focus on events and actions, denoted as $\mathcal{P}^a = \{P_i^a\}_{i=1}^K$, as in Appendix D.2.

3.2.2 Object Detections

Spatial awareness enhances reasoning by incorporating structural and contextual object descriptions of an image (Chen et al., 2023; Ranasinghe et al., 2024b). We leverage the high-fidelity VLM CogAgent (Hong et al., 2024) to extract objects from video frames as interactive subjects, aiding LLMs' spatial understanding. The VLM identifies a fixed number of prominent objects per frame. To maintain temporal consistency within an event clip, objects are sequentially stored as semantic representations (Fig. 3) for LLM reasoning, denoted as $\mathcal{P}^o = \{P_i^o\}_{i=1}^K$, as in Appendix D.2.

3.2.3 Query-dependent Summarization

Given a query, we prompt the LLMs to get a query-based summarization of the spatial information. The query-based summarization serves as an implicit Chain-of-Thought (Wei et al., 2023) for LLMs to reason over the spatial information, focusing on the query about long clips. The summarization step $\mathcal{P}_s = \{P_i^s\}_{i=1}^K = \{(sum_{\mathcal{LLM}}(P_i^a, Q), sum_{\mathcal{LLM}}(P_i^o, Q))\}_{i=1}^K$ contains action summarizations focusing on event information and object summarizations focusing on environment information, as in Appendix D.2.

3.3 Informative Reasoning with Self-Reflection

Inspired by Reflexion (Shinn et al., 2023), the third phase of VidoeINSTA proposes a self-reflective information reasoning scheme – with LLMs to reason on spatial-temporal information collected in the previous stages. Particularly, we balance between information sufficiency and the temporal order. Two evaluation scores are defined as intermediate metrics in our algorithm.

Informative Score. The LLM is required to generate an Informative Score $S_I = \{S_i^I\}_{i=1}^K \in [1, 2, 3]$ for each clip indicating $[\text{not sufficient}, \text{marginal sufficient}, \text{sufficient}]$, which is an initial evaluation of the information sufficiency of the prompts derived in previous stages.

Confidence Score. The LLM is required to generate a Confidence Score $S_C = \{S_i^C\}_{i=1}^K \in [1, 2, 3]$ for each question-answering round indicating

Algorithm 1: VideoINSTA

```
Input     : Video $V$, Question $Q$, Options $\{o_0, o_1, o_2, o_3, o_4\}$
Parameter : Number of segments $K \in \mathbb{N}^+$
Output    : Final Prediction $answer \in \{o_0, o_1, o_2, o_3, o_4\}$
$V' \leftarrow \emptyset$;        // for clip descriptions and informative scores
$\mathcal{E} \leftarrow$ temporal_segmentation($V, K$);
$T \leftarrow$ temporal_grounding($V, Q$);
$A \leftarrow$ action_captions($V$);
$O \leftarrow$ object_detections($V$);
for $E_i \in \mathcal{E}$ do
    $P_i^a \leftarrow$ inherit($A, E_i$);
    $P_i^o \leftarrow$ inherit($O, E_i$);
    $P_i^t \leftarrow$ inherit($T, E_i$);
    $P_i^{sa} \leftarrow$ summarize($P_i^a, Q$);
    $P_i^{so} \leftarrow$ summarize($P_i^o, Q$);
    $P_i \leftarrow (P_i^a, P_i^o, P_i^t, P_i^{sa}, P_i^{so})$;
    $S_i^I \leftarrow$ informative_eval($P_i, Q, (o_0, o_1, o_2, o_3, o_4)$);
    $V'$.insert(($P_i, S_i^I$));    // $i$-th clip description and info
      score
end
$V'' \leftarrow$ sort_descending($V'$, key $= V'.S_I$);    // by info scores
$L \leftarrow \emptyset$;      // for merged clip descriptions without info scores
for $E_i \in V''$ do
    $P_i, S_i^I \leftarrow E_i$;
    $L$.insert($P_i$);
    if $i \neq |V''| - 1$ and $S_{(i+1)}^I = 3$ then
        continue;
    end
    else
        $L' \leftarrow$ sort_temporally($L$);
        $P_{L'} \leftarrow$ concatenate($L'$);
        $answer, prompt, completion \leftarrow$
          QA($P_{L'}, Q, (o_0, o_1, o_2, o_3, o_4)$);
        $S_i^C \leftarrow$ self_reflect($prompt, completion$);
        if $S_i^C = 3$ then
            break;
        end
    end
end
return $answer$;
```

[not confident, marginal confident, very confident], which is a self-evaluation of the answer prediction.

Self-reflective reasoning. The algorithm shown in Alg. 1 starts with an initial evaluation step for the LLM to derive an informative score for each clip. Then, the informative states are sorted in descending order according to their informative scores and maintained in a list. Within the same informative level, the prompts are ordered temporally. Then, the algorithm performs a multi-round self-reflective scheme, specifically merging informative clips and evaluating the question-answering confidence. In the first round, sufficient informative states are merged and prompted to the LLM for question-answering. Then, the LLM is required to derive a confidence score. If the LLM is not confident enough about its prediction, a further clip with a lower informative score is merged into the state which gets temporally re-ordered. The alternating merge-and-evaluate scheme ends until all clips are merged or the prediction confidence reaches the top value. The VideoINSTA is detailed in Alg. 1 and on the right of Figure. 1.

4 Extensibility of the Framework

Extensible API tools VideoINSTA is a general framework for informative spatial-temporal reasoning on videos and maintains the extensibility to improve both, the temporal reasoning and spatial reasoning phases by acquiring informative prompts from different expert tools through APIs. For example, expert temporal segmentation models can be utilized for better event parsing in the temporal reasoning phase in VideoINSTA. Expert spatial models like high-fidelity captioning models and object detectors can provide more accurate informative prompts for the spatial reasoning phase.

Open Question Answering Apart from single-choice question answering, VideoINSTA can also be easily adapted to open question answering. We tested VideoINSTA on AcitivityNet-QA (Yu et al., 2019), which is a dataset for open-ended question answering over complex web videos. Following (Maaz et al., 2024), we also conduct evaluation in a zero-shot manner, employing LLM-assisted evaluation to assess the predictions' accuracy of VideoINSTA.

5 Experimental Setup

In this section, we describe the experimental setup of the VideoINSTA framework. We present quantitative results and a qualitative analysis on the EgoSchema (Mangalam et al., 2024), Next-QA (Xiao et al., 2021), and Intent-QA (Li et al., 2023a) benchmarks.

EgoSchema EgoSchema is a benchmark for long-form video understanding, featuring 5,000 single-choice questions derived from egocentric videos. A distinctive feature of this dataset is the length of its videos, each lasting 180 seconds. EgoSchema comprises only a test set, with a subset of 500 questions having available labels.

NextQA The NExT-QA dataset includes 5,440 natural videos that feature object interactions in daily life, accompanied by 48,000 single-choice questions. The average length of the video is 44 seconds. In line with standard practices, our zero-shot evaluation is focused on the validation set.

IntentQA IntentQA focuses on intent reasoning. It contains 4,303 videos and 16K single-choice question-answer pairs focused on reasoning about people's intent in the video. The videos are more than 44 seconds in average length. We perform a zero-shot evaluation on the test set.

Evaluation Metrics Since each dataset features single-choice questions and VideoINSTA generates option predictions directly, we utilized accuracy as the evaluation metric.

Baselines The baselines include recent representative LLM-based zero-shot video understanding methods – including LLoVi, VideoAgent, ProViQ and MoReVQA – and other baselines include supervised end-to-end models, see Table 1.

Experiment Design To comprehensively analyze VideoINSTA, there are two research questions. **RQ1**: How is the performance of the proposed VideoINSTA framework compared to the existing end-to-end models and LLM-based compound systems? **RQ2**: How do the components of the VideoINSTA affect its effectiveness?

Implementation Details Following LLoVi and VideoAgent, we utilize the LaViLa model retrained on Ego4D, filtering out videos that overlap with EgoSchema to ensure zero-shot evaluation.

6 Experimental Results

6.1 Main Results

Comparison with State-of-the-arts To answer the RQ1, our average results over multiple run from Table 1 achieve state-of-the-art performance, surpassing all types of existing end-to-end models, proprietary models, and zero-shot compound systems across three datasets.

Noticeably, **VideoINSTA with ChatGPT3.5 surpasses the other zero-shot LLM-based baselines LLoVi and VideoAgent with ChatGPT-4**. Our method demonstrates spatial-temporal informative reasoning to serve as the foundational framework for zero-shot video reasoning, opening a new state-of-the-art in the video question-answering domain.

Open Question Answering We measure the accuracy by utilizing an LLM to evaluate the generated prediction by comparing it to the ground truth answer and assigning a true or false value accordingly. Table 2 shows the results with Llama-3. VideoINSTA achieves more than double the performance compared to the baseline LLoVi with **151.3%** relative improvement.

6.1.1 Ablation on Main Stage

We undertake ablation studies on EgoSchema to evaluate the contribution of each phase in VideoINSTA with three distinct variations: VideoINSTA w/o TA (without event-based temporal reasoning), VideoINSTA w/o S (without content-based spatial reasoning), and VideoINSTA w/o IN (without self-reflective information reasoning). We further investigate event-based temporal reasoning and the contribution of the query-unaware temporal segmentation (VideoINSTA w/o TA-Seg.) and the query-aware temporal inheritance (VideoINSTA w/o TA-Inhr.). Figure 5 concludes that all phases in the VideoINSTA framework contribute to distinct performance improvements including the two sub-steps in the temporal reasoning. The whole pipeline enables VideoINSTA to outperform existing methods.

6.2 Ablation on Temporal Reasoning

Clustering in Temporal Segmentation To evidently prove the effectiveness of our proposed C-DPCKNN, we conduct experiments on variants VideoINSTA w. TA-Seg. (Uniform), w. TA-Seg. (KNN), w. TA-Seg. (DPCKNN) and w. TA-Seg. (C-DPCKNN) on both EgoSchema and NExT-QA. The quantitative results of this comparison are illustrated in Figure 6. The results validate that our proposed C-DPCKNN method for query-unaware temporal segmentation is superior to the other approaches. Additionally, the worse performance of Uniform, KNN, and DPCKNN highlights that improper segmentation can severely impact subsequent reasoning steps. We conclude that they have the same drawback of improper segmentation, further validating the effectiveness of C-DPCKNN.

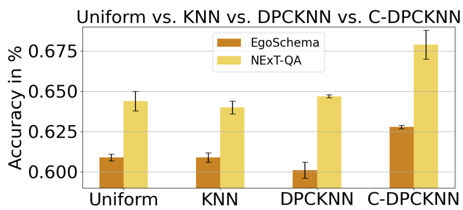

Figure 4: Ablation on different temporal segmentation of VideoINSTA methods.

Number of Events in Temporal Segmentation To further explore the impact of C-DPCKNN in temporal segmentation within our temporal reasoning framework, we conducted a series of experiments on the EgoSchema dataset. We varied the number of event clips K from the set $\{2, 4, 8\}$. For each configuration, we kept the implementation of other components in VideoINSTA consistent. Empirical results reveal an optimal critical value

LLM (Method \ Dataset)	System	EgoSchema	NExT-QA	IntentQA
Random Chance		20.0	20.0	20.0
Supervised State-of-the-Art				
LongViViT (Papalampidi et al., 2023)		56.8	-	-
MC-ViT-L (Balažević et al., 2024)		62.6	65.0	-
Training-Free State-of-the-Art				
LLM	System			
PaLM-2 (Anil et al., 2023)	MoReVQA (Ranasinghe et al., 2024a)	51.7†	69.2	-
FlanT5-3B (Raffel et al., 2020)	SeViLA (Yu et al., 2024)	25.7	63.6	60.9
Mistral-7B (Jiang et al., 2023)	LangRepo (Kahatapitiya et al., 2024b)	60.8	54.6	53.8
	MVU (Ranasinghe et al., 2024a)	60.3	55.2	-
Llama2-7B (Touvron et al., 2023)	LLoVi (Zhang et al., 2023a)	34.0	-	-
Llama2-13B (Touvron et al., 2023)	LLoVi (Zhang et al., 2023a)	40.4	-	-
Llama2-70B (Touvron et al., 2023)	LLoVi (Zhang et al., 2023a)	50.6	-	-
	VideoAgent (Wang et al., 2024a)	45.4	-	-
GPT-3 (Brown et al., 2020)	ViperGPT (Surís et al., 2023)	-	60.0	-
GPT-4V (OpenAI, 2024a)	IG-VLM (Kim et al., 2024)	59.8	68.6	64.2
	GPT-4V (Balažević et al., 2024)	63.5	-	-
Llama3-8B (Dubey et al., 2024)	LLoVi (Zhang et al., 2023a) (ours)	47.6	46.6	48.9
	VideoINSTA	52.6	58.3	53.0
ChatGPT-4 (OpenAI, 2024a)	LLoVi (Zhang et al., 2023a)	61.2	67.7	64.0
	AssistGPT (Gao et al., 2023)	-	58.4	-
	VideoAgent (Wang et al., 2024a)	60.2	71.3	-
	VideoAgent (Fan et al., 2024)	62.8	70.8	-
	TraveLER (Shang et al., 2024)	-	68.2	-
	VideoTree (Wang et al., 2024b)	**66.2**	**73.5**	<u>66.9</u>
	VideoINSTA	<u>65.0</u>	<u>72.3</u>	**72.8**
ChatGPT-3.5 (OpenAI, 2024a)	LLoVi (Zhang et al., 2023a)	58.8	-	-
	ProViQ (Choudhury et al., 2023)	57.1	63.8‡	-
	VideoAgent (Wang et al., 2024a)	-	48.8	-
	VideoTree (Wang et al., 2024b)	57.6	-	-
	VideoINSTA	62.8	67.9	64.4

Table 1: Video Reasoning Results. The best accuracy (%) is highlighted in orange and the second best in yellow for each training-free (zero-shot or few-shot) method respectively. Note that we are strictly zero-shot without using in-context examples in our prompts. The best result among all methods is **bold** and the second best is <u>underlined</u>.

LLM	Model	Accuracy (%)
Llama-3-8B-Instruct (AI@Meta, 2024)	LLoVi	14.75
	VideoINSTA	**37.06 (151.3% ↑)**

Table 2: Accuracy performance of VideoINSTA on open question answering dataset ActivityNet-QA.

for the number of events K, as shown in Figure 5(b). EgoSchema videos are characterized by their uniform length of 3 minutes, with a high temporal certificate - a metric indicating the proportion of necessary informative segments to the total video duration. The empirical findings suggest that K intuitively corresponds to the actual number of events observed in the videos.

6.3 Ablation on Spatial Stage

Spatial Captioners We provide an ablation study over captioners comparing CogAgent vs. LLaVA-1.5 (Liu et al., 2023) on NExT-QA, indicating that a better captioner leads to better information quality as CogAgent is a captioner with higher fidelity since it was especially designed for Graphical User Interface understanding and navigation, which requires fine-granular perception. Therefore, CogAgent facilitates better informativeness in tasks involving visual and linguistics.

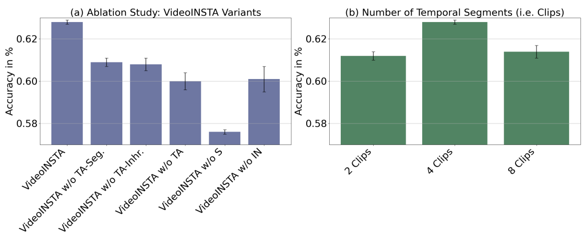

Figure 5: Ablation Studies on EgoSchema. (a) All three phases contribute to **VideoINSTA**. (b) $K = 4$ is the best empirical clustering number for EgoSchema.

LLM	Object Captioner	Accuracy
ChatGPT-3.5 (OpenAI, 2024a)	CogAgent	0.679
	LLaVA-1.5	0.628

Table 3: Performance metrics for different captioners using ChatGPT3.5 on the NExT-QA dataset.

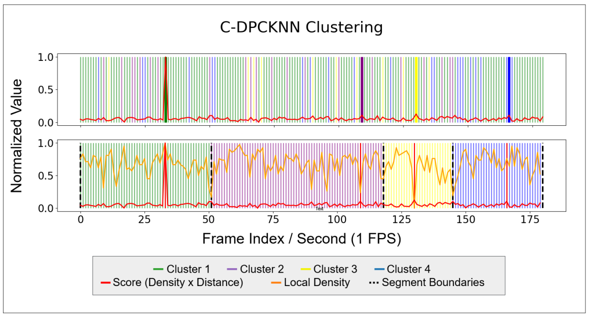

Figure 6: The Upper figure illustrates the intermediate results of DPCKNN clustering with blended boundaries among clusters. The bottom figure illustrates clearer event boundaries with the proposed C-DPCKNN.

6.4 Qualitative Analysis

Event Segmentation with Clear Borders We visualize the temporal segmentation performance on EgoSchema. As seen in Figure 6, the upper figure illustrates the intermediate clustering results with the original DPCKNN. According to the results, frames clustered to the same event are scattered across the video, and the event boundaries are blended, which is counter-intuitive to how untrimmed videos present their content. The bottom figure illustrates the results of how our proposed C-DPCKNN utilizes density peaks as sharp boundaries. This qualitative visualization shows that events are parsed correctly around clustering centers and the respective borders align to the regions with high fluctuations among frame features.

Clear Segmentation for Correct Grounding We further investigate how the two variants of VideoINSTA w/. TA-Seg(KNN) and TA-Seg(C-DPCKNN) affects the grounding descriptions. We can find that the density-based clustering in C-DPCKNN successfully captures the scene transitions indicating the borders are set to where the content changes drastically, when the man starts to catch fish in a fishbowl in the bathroom as underlined in Figure 7. The consequent actions of

[†]Obtained on the hidden test split of EgoSchema (5,000 tasks) instead of the public test split (500 tasks) as all the other results.

[‡]Not obtained on the validation split of NExT-QA as the other results, but on the test split.

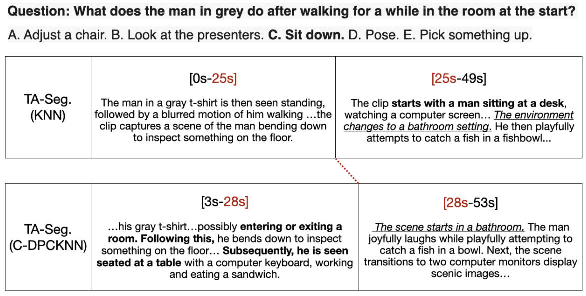

Figure 7: Performance of C-DPCKNN leads to clearer boundaries over KNN that contributes to exact semantic representations for videos segments.

the man in gray before he went to the bathroom are fully tracked in the same clip, leading to the correct answer "C) sit down". However, the KNN method falsely sets the border causing important information loss leading to the false answer "E) pickup something".

Spatially Informative Captions VLMs share a tendency to focus on describing the actions and events happning in the video clips or frames. However, the environment in videos and the interactions between human and objects provide more trivial but essential information for accurate reasoning in a fine-grained level, to which the spatially informative reasoning with object detections contributes. An example in IntentQA has the answer *"Seat belt"* to the question *"How did the people make sure that the babies will not fall off the swing easily when playing on them?"*. Basic video narrations will lead to captions like *"Some people are standing around the babies and playing swings."*, leading to a false prediction of *"Standing Around"*, while neglecting the crucial factor for safety, which actually is the object *seat belt*.

7 Conclusion

This work focuses on understanding long-form videos with LLMs – particularly emphasizing information quality, spatial-temporal reasoning, and explicit complex reasoning across unbalanced distributed information. The proposed training-free framework **VideoINSTA** for long-form video understanding showcases exceeding performance over state-of-the-art end-to-end and zero-shot LLM-based methods. It further reveals the potential on open question answering and the extensibility of various visual-language tool-augmented spatial-temporal reasoning approaches.

Limitation

The limitation of **VideoINSTA** lies in its nature as a compound system, centered around a large language model (LLM) and incorporating various visual-language tools to process spatial-temporal information. If the number of tools or the rounds of reasoning increase to some level, there is a heightened risk of inconsistency and randomness of generated intermediate thoughts, potentially introducing additional noise into the reasoning process.

Ethics Statement

VideoINSTA is tailored as a compound system utilizing various visual-language tools for spatial-temporal information extraction. This framework might help with developing visual understanding systems for assisting daily life since it has exceeding results on the first-view dataset EgoSchema. The risk of VideoINSTA might be inherited from open-source LLMs, such as bias and hallucinations. Besides, We only use AI assistants (e.g., ChatGPT) to conduct experiments in this research.

Liscences

The datasets used in this research work are open-sourced and can be seen in references. We use the datasets from the original version within the intended use term. The licenses of the models used in this paper are listed.

- LLoVi
- EgoSchema
- NExT-QA
- Intent-QA
- UniVTG
- Chat-UniVi
- CogAgent
- LLama3

Acknowledgements

This work was funded by the Munich Center for Machine Learning and supported by the Federal Ministry of Education and Research and the State of Bavaria.

References

AI@Meta. 2024. Llama 3 model card.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J. Hénaff. 2024. Memory consolidation enables long-context video understanding.

Peijun Bao, Zihao Shao, Wenhan Yang, Boon Poh Ng, and Alex C Kot. 2024. E3m: Zero-shot spatio-temporal video grounding with expectation-maximization multimodal modulation. ECCV.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. Longlora: Efficient fine-tuning of long-context large language models.

Minkyu Choi, Harsh Goel, Mohammad Omama, Yunhao Yang, Sahil Shah, and Sandeep Chinchali. 2024. Neuro-symbolic video search. *arXiv preprint arXiv:2403.11021*.

Rohan Choudhury, Koichiro Niinuma, Kris M Kitani, and László A Jeni. 2023. Zero-shot video question answering with procedural programs. *arXiv preprint arXiv:2312.00937*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

Zifeng Ding, Heling Cai, Jingpei Wu, Yunpu Ma, Ruotong Liao, Bo Xiong, and Volker Tresp. 2024. zrLLM: Zero-shot relational learning on temporal knowledge graphs with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1877–1895, Mexico City, Mexico. Association for Computational Linguistics.

Mingjing Du, Shifei Ding, and Hongjie Jia. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding.

Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. 2023. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn.

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

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.

Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer.

Zhen Han, Ruotong Liao, Jindong Gu, Yao Zhang, Zifeng Ding, Yujia Gu, Heinz Koeppl, Hinrich Schütze, and Volker Tresp. 2023. ECOLA: Enhancing temporal knowledge embeddings with contextualized language representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5433–5447, Toronto, Canada. Association for Computational Linguistics.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290.

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

HuggingFace. 2024. Hugging face website. Accessed: 2024-06-13.

Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. 2019. Videograph: Recognizing minutes-long human activities in videos. *arXiv preprint arXiv:1905.05143*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024a. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024b. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13700–13710.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. 2024a. Language repository for long video understanding. *arXiv preprint arXiv:2403.14622*.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S. Ryoo. 2024b. Language repository for long video understanding.

Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. 2024. An image grid can be worth a video: Zero-shot video question answering using a vlm.

Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. 2023. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023a. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11963–11974.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Rongchang Li, Zhenhua Feng, Tianyang Xu, Linze Li, Xiao-Jun Wu, Muhammad Awais, Sara Atito, and Josef Kittler. 2024. C2c: Component-to-composition learning for zero-shot compositional action recognition. *arXiv preprint arXiv:2407.06113*.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937.

Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. 2024. GenTKG: Generative forecasting on temporal knowledge graph with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4303–4317, Mexico City, Mexico. Association for Computational Linguistics.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023a. Video-llava: Learning united visual representation by alignment before projection.

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

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023c. Univtg: Towards unified video-language temporal grounding.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

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

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

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2024. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36.

Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. Large language models know your contextual search intent: A prompting framework for conversational search.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. 2024. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245.

OpenAI. 2024a. Chatgpt: Gpt-3.5 and gpt-4 and gpt-4v(ision). `https://www.openai.com/chatgpt`. Accessed: 2024-06-13.

OpenAI. 2024b. Chatgpt model documentation. Accessed: 2024-06-13.

Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. 2023. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. Univtg: Towards unified video-language temporal grounding. *arXiv e-prints*, pages arXiv–2307.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S. Ryoo. 2024a. Understanding long videos in one multimodal language model pass.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. 2024b. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987.

Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. 2024. Traveler: A multi-lmm agent framework for video question-answering.

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. 2024. A vision check-up for language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14410–14419.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Showlab. 2024. Univtg model documentation. Accessed: 2024-06-13.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

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

Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. 2023a. Vamos: Versatile action models for video understanding. *arXiv preprint arXiv:2311.13627*.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024a. Videoagent: Long-form video understanding with large language model as agent.

Yang Wang, Gedas Bertasius, Tae-Hyun Oh, Abhinav Gupta, Minh Hoai, and Lorenzo Torresani. 2021. Supervoxel attention graphs for long-range video modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 155–166.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023b. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022a. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022b. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2024b. Videotree: Adaptive tree-based video representation for llm reasoning on long videos.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. 2024. Longvlm: Efficient long video understanding via large language models. *arXiv preprint arXiv:2404.03384*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024a. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint arxiv:2404.03622*.

Xiaoqian Wu, Yong-Lu Li, Jianhua Sun, and Cewu Lu. 2024b. Symbol-llm: Leverage language models for symbolic system in visual human activity reasoning. *Advances in Neural Information Processing Systems*, 36.

Xinxiao Wu, Ruiqi Wang, Jingyi Hou, Hanxi Lin, and Jiebo Luo. 2021. Spatial–temporal relation reasoning for action prediction in videos. *International Journal of Computer Vision*, 129(5):1484–1505.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 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*, pages 9777–9786.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2804–2812.

Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. 2023. Contrastive video question answering via video graph transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. 2023. Retrieval-based video language model for efficient long video question answering. *arXiv preprint arXiv:2312.04931*.

Yutaro Yamada, Yihan Bao, Andrew K Lampinen, Jungo Kasai, and Ilker Yildirim. 2023. Evaluating spatial understanding of large language models. *arXiv preprint arXiv:2310.14540*.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 124–141. Curran Associates, Inc.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2024. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 1963–1974.

Guangyao Zhai, Liang Liu, Linjian Zhang, Yong Liu, and Yunliang Jiang. 2020. Poseconvgru: A monocular approach for visual ego-motion estimation by learning. *Pattern Recognition*, 102:107187.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2023a. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*.

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

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

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023a. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023b. Learning video representations from large language models. In *CVPR*.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. 2024. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.

Wencheng Zhu, Yucheng Han, Jiwen Lu, and Jie Zhou. 2022. Relational reasoning over spatial-temporal graphs for video summarization. *IEEE Transactions on Image Processing*, 31:3017–3031.

A Case Studies

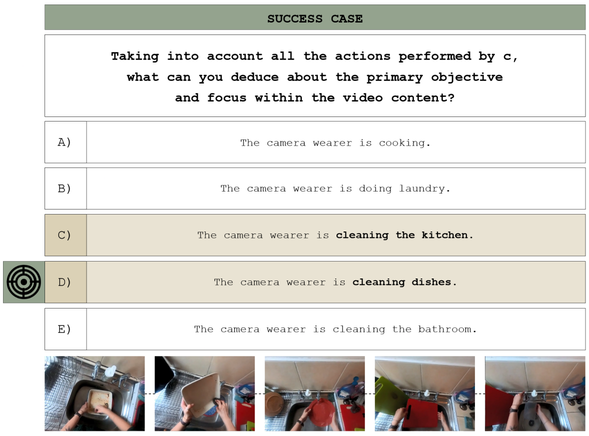

Figure 8: Sucess case. The ambiguity between the answer options C) and D) is highlighted in bold. The ground truth answer option is marked with a bullseye symbol and the prediction of the VideoINSTA framework is indicated with a crosshair symbol. In this case, they are overlapped.

Success Case As shown in Figure 8, the VideoINSTA framework effectively addresses the ambiguity between the actions "cleaning dishes" and "cleaning the kitchen." While "cleaning the kitchen" appears broader and potentially applicable, "cleaning dishes" is more specific to the actual video content. A human viewer, after watching the video and reviewing the answer options, would likely determine that the individual is focused solely on cleaning dishes, rather than wiping kitchen surfaces or completing other tasks. Thus, "cleaning dishes" is the more accurate selection.

Failure Case Figure 9 shows the failure case. The task is to determine whether the importance of precision stems from the need to cut the wood "evenly and consistently" (option B) or to the "correct size" (option D). A brief review of the video might suggest that both options are plausible. However, watching the full video reveals that only a single piece of wood is involved throughout, making "cutting to the correct size" the more accurate answer. The option of "cutting evenly and consistently" would imply the presence of multiple pieces, which is not the case, even when the wood temporarily leaves the camera's view. Unlike a human, who intuitively recognizes that the reappearing wood is still the same and that no other pieces exist, VideoINSTA struggles to track it consistently due to its lack of an environmental conciousness and the inability to track object identity. This shortcoming prevents VideoINSTA from recognizing that "cutting evenly and consistently" is irrelevant in this scenario, leading to the selection of an incorrect answer instead of the ground-truth response.

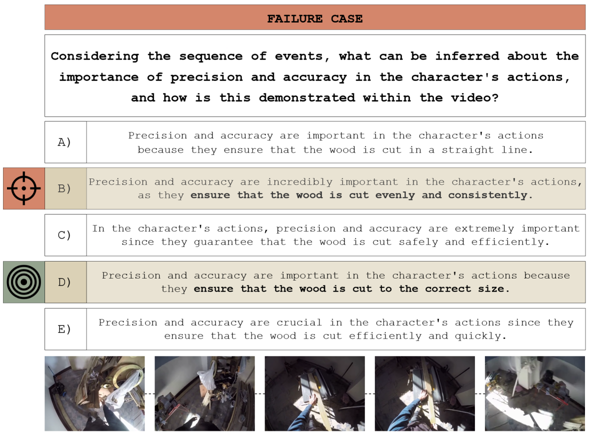

Figure 9: Failure case. The ambiguity between the answer options B) and D) is highlighted in bold. The ground truth answer option is marked with a bullseye symbol and the prediction of the VideoINSTA framework is indicated with a crosshair symbol. In this case, our algorithm fails to predict the ground truth option D and aims for B) instead.

B More Related Works

Video Language Models With the in-depth investigation into Multi-modal Large Language Models(MLLMs) (Gu et al., 2023; Wu et al., 2023; Cui et al., 2024), there has been growing attention to bridging video modality to generative large language models such as Video-llama (Zhang et al., 2023c), Video-LLaVA (Lin et al., 2023b), LanguageBind (Zhu et al., 2023), VideoChat (Li et al., 2023b), ChatUniV (Jin et al., 2024b) InternVideo (Wang et al., 2022a), etc., which are dependent on meticulously designed model structures, or adaptation layers. They suffer from large- scale pertaining, or requiring proper datasets for instruction tuning such as InternVid (Wang et al., 2023b). Therefore, a line of work utilizing LLM as a compound system center or agent-based reasoning for video understanding has been introduced, which we discussed extensively in our baselines in Sec. 2 and experiments 1. Another line of work, focusing on low-resource and even zero-shot understanding of videos emerges, such as LLaVA-Next (Liu et al., 2024), E3M (Bao et al., 2024), LongVLM (Weng et al., 2024), C2C (Li et al., 2024), (Choi et al., 2024), etc, where they enlight the task more lightly.

C Supplementary Statistics

Dataset Statistics We report the split that we use for our experiments in Table 4, the number of tasks in those splits – i.e. the number of question-answer-pairs – as well as the number of videos within those splits. Furthermore, we report the average, minimum and maximum video length in seconds of the videos in the corresponding split – these numbers can vary from the ones for the whole datasets.

Datasets	Split	#Tasks	#Videos	Avg. Length	Min. Length	Max. Length
EgoSchema	Public Test	500	500	180.0	180.0	180.0
NExT-QA	Validation	4,996	570	42.2	10.0	180.0
IntentQA	Test	2,134	576	44.9	6.0	180.0
ActivityNet-QA	Test	8,000	800	112.1	3.0	285.7

Table 4: Dataset statistics.

Pre-trained model versions and statistics As shown in Table 5, we abbreviate Large Language Model with LLM, Vision Language Model with VLM, Visual Temporal Grounding Model with VTGM, and Vision Encoder with VE. Please refer to the implementation details for the exact hyper-parameters that we use, since they vary for some different experiments and use cases.

Models	Version	Type	#Params	Context
ChatGPT 3.5	gpt-3.5-turbo-1106	LLM	N/A	16k
ChatGPT 4	gpt-4-1106-preview	LLM	N/A	128k
Llama3	meta-llama/Meta-Llama-3-8B-Instruct	LLM	8B	8k
UniVTG	CLIP-B/32 Pretraining (Finetuned)	VTGM	N/A	N/A
LaViLa	Fair Checkpoint (Zhang et al., 2023a)	VLM	N/A	0
CogAgent	THUDM/cogagent-vqa-hf, lmsys/vicuna-7b-v1.5	VLM	18B	N/A

Table 5: Pre-trained model versions and statistics.

Method	EgoSchema	NExT-QA
w. TA-Seg. (Uniform)	0.600 (±0.004)	0.644 (±0.006)
w. TA-Seg. (KNN)	0.609 (±0.003)	0.640 (±0.004)
w. TA-Seg (DPCKNN)	0.601(±0.001)	0.647 (±0.001
w. TA-Seg. (C-DPCKNN)	**0.628 (±0.001)**	**0.679 (±0.009)**

Table 6: Ablation on different temporal segmentation of VideoINSTA methods on EgoSchema and NExT-QA datasets.

D Implementation Details

D.1 Experiment Setup

We split a dataset into equal-sized chunks and run a sub-experiment on each of them for parallelization purposes. We collect and aggregate the results of all sub-experiments afterward to obtain the final experiment result. We use the types of GPU servers: NVIDIA RTX A6000 GPU, NVIDIA A100-PCIE-40GB, Quadro RTX 8000, NVIDIA RTX 3090.

D.1.1 Details of Llama3

When we refer to Llama3, we use the instruction-tuned version *meta-llama/Meta-Llama-3-8B-Instruct* (AI@Meta, 2024), which is available on HuggingFace (HuggingFace, 2024). We use greedy sampling – comparable with a temperature of 0.0 – throughout all our experiments.

D.1.2 Details of ChatGPT

When we refer to ChatGPT 3.5, we use the instruction-tuned version *gpt-3.5-turbo-1106*, and when we refer to ChatGPT 4, we use the instruction-tuned version *gpt-4-1106-preview* OpenAI, 2024a,b. Following (Zhang et al., 2023a), we use a temperature of 1.0 for the summarization.

D.1.3 Details of LaViLa

For our experiments on EgoSchema, we use LaViLa (Zhao et al., 2023b) as the action captioner. Following (Zhang et al., 2023a), we use their re-trained model checkpoint to avoid data leakage and ensure a fair comparison. We uniformly sample 4 frames from each consecutive $1s$-interval of the video to obtain a caption.

D.1.4 Details of CogAgent

Following (Wang et al., 2024a), we leverage the VLM CogAgent (Hong et al., 2024) as the action captioner for our experiments on NExT-QA, IntentQA and ActivityNetQA. Moreover, we use it as the label-free object detector for our experiments on all datasets. Specifically, we use the model *THUDM/cogagent-vqa-hf* together with the tokenizer *lmsys/vicuna-7b-v1.5*, which are available on HuggingFace (HuggingFace, 2024).

D.1.5 Details of UniVTG

We leverage UniVTG (Lin et al., 2023c) to get the temporal grounding of a video and finally retrieve the most important interval regarding the question of a task. We use *ViT-B/32* as the CLIP vision encoder model version (Radford et al., 2021) together with their best-fine-tuned model checkpoint (Showlab, 2024).

D.1.6 Details of C-DPCKNN

We use the CLIP vision encoder *openai/clip-vit-large-patch14* (Radford et al., 2021), which is available on HuggingFace (HuggingFace, 2024).

D.2 Prompts

```
You are given some language descriptions of a first-person view video. The video is {length} seconds long. Each sentence describes a 1.0s clip. The descriptions are sequential and non-overlapping which cover the whole video exactly. Here are the descriptions: {interval_text}.\n Please give me a {words} words summary. When doing summarization, remember that your summary will be used to answer this multiple choice question: {question}
```

Table 7: Action Captions Summarization Prompt Template for ChatGPT. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}.

You are given some language descriptions of a first person view video. The video is {length} seconds long. Each sentence describes a 1.0s clip. The descriptions are sequential and non-overlapping which cover the whole video exactly. Here are the descriptions: {interval_text}.\n **Please give me a summary of these action captions. Please write an easy-to-read continuous text. You can use paragraphs, but do not use special formatting such as bulleted or numbered lists. Please use {words} words for your summary.** When doing summarization, remember that your summary will be used to answer this multiple choice question: {question}

Table 8: Action Captions Summarization Prompt Template for Llama3. The difference to the prompt template for ChatGPT is highlighted in **bold**. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}.

```
You are given a list of the most eye-catching objects that were detected in each frame of a video clip using a visual large language model. The list appears in the temporal order of the frames. The video is {length} seconds long. Each sentence describes the objects of a 1.0s clip. The object detections are sequential and non-overlapping which cover the whole video exactly. Here are the object detections:\n\n{interval_text}.\n\nPlease give me a {words} words summary of these object detections. When doing summarization, remember that your summary will be used to answer this multiple choice question: {question}
```

Table 9: Object Detections Summarization Prompt Template for ChatGPT. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}.

You are given a list of the most eye-catching objects that were detected in each frame of a video clip using a visual large language model. The list appears in the temporal order of the frames. The video is {length} seconds long. Each sentence describes the objects of a 1.0s clip. The object detections are sequential and non-overlapping which cover the whole video exactly. Here are the object detections:\n\n{interval_text}.\n\n**Please give me a summary of these object detections. Please write an easy-to-read continuous text. You can use paragraphs, but do not use special formatting such as bulleted or numbered lists. Please use {words} words for your summary.** When doing summarization, remember that your summary will be used to answer this multiple choice question: {question}

Table 10: Object Detections Summarization Prompt Template for Llama3. The difference to the prompt template for ChatGPT is highlighted in **bold**. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}.

```
# Video Question Answering
\n\nHi there! Now that you have studied the topic of video question answering for years, you
find yourself in the final exam of your studies. Please take your time to solve this task.
You can do it! You know everything that is required to master it. Good luck!

\n\n## What is Video Question Answering?
\n\nVideo Question Answering is a task that requires reasoning about the content of a video
to answer a question about it. In this exam, you will be given purely textual information
about a single clip of the video that has been extracted beforehand. Your task is to read
the information about the clip carefully and evaluate whether the given clip is needed to
answer the question about the video or not.

\n\n## Here is your task
\n\nPlease think step by step to evaluate the answerability of the given question and options
based on the given clip. The question is a single choice question with five answer options,
such that there is exactly one best answer option.  Is the information in the given clip
sufficient to answer the given question with one of the given options? Please make sure to
include all relevant information in your evaluation.

\n\nPlease use the following criteria for evaluation:
\n    1. Irrelevant information {{'answerability': 1}}: If information of this clip is not
even relevant to the question.
\n    2. Insufficient information {{'answerability': 2}}: If information of this clip is
potentially useful to answer the question, but more clips are needed to confidently answer
the question.
\n    3. Sufficient information {{'answerability': 3}}: If the information of this clip is
sufficient to answer the question and no other clip is needed.

\n\nPlease write your answerability X in JSON format {{'answerability':  X}}, where X
is in {{1, 2, 3}}.

\n\n## Here is the information about the video clip
\n\n### Information about one of four clips of the video
\n{lexical_node_state_representation}

\n\n### Question
\n\n{question}

\n\n### Five answer options
\n\n    A) {option_0}
\n    B) {option_1}
\n    C) {option_2}
\n    D) {option_3}
\n    E) {option_4}

\n\n## Now it is your turn
\n\nPlease think step by step to provide your evaluation and provide the answerability X in
JSON format {{'answerability': X}}, where X is in {{1, 2, 3}}:
\n\n
```

Table 11: Answerability Rating Prompt Template for ChatGPT. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}. JSON-formatting is indicated by {{double curly brackets}}, as one level of brackets will be removed when the prompt prompt template gets filled.

```
# Video Question Answering
\n\nHi there! Now that you have studied the topic of video question answering for years, you find yourself in the final exam of your studies. Please take your time to solve this task. You can do it! You know everything that is required to master it. Good luck!

\n\n## What is Video Question Answering?
\n\nVideo Question Answering is a task that requires reasoning about the content of a video to answer a question about it. In this exam, you will be given purely textual information about a single clip of the video that has been extracted beforehand. Your task is to read the information about the clip carefully and evaluate whether the given clip is needed to answer the question about the video or not.

\n\n## Here is your task
\n\nPlease think step by step to evaluate the answerability of the given question and options based on the given clip. The question is a single choice question with five answer options, such that there is exactly one best answer option. Is the information in the given clip sufficient to answer the given question with one of the given options? Please make sure to include all relevant information in your evaluation. Moreover, make sure that you always provide an answerability, even if it seems ambiguous or unsolvable.

\n\nPlease use the following criteria for evaluation:
\n    1. Irrelevant information {{'answerability': 1}}: If information of this clip is not even relevant to the question.
\n    2. Insufficient information {{'answerability': 2}}: If information of this clip is potentially useful to answer the question, but more clips are needed to confidently answer the question.
\n    3. Sufficient information {{'answerability': 3}}: If the information of this clip is sufficient to answer the question and no other clip is needed.

\n\nPlease write your answerability X in JSON format {{'answerability': X}}, where X is in {{1, 2, 3}}.

\n\n## Here is the information about the video clip
\n\n### Information about one of four clips of the video
\n{lexical_node_state_representation}

\n\n### Question
\n\n{question}

\n\n### Five answer options
\n\n    A) {option_0}
\n    B) {option_1}
\n    C) {option_2}
\n    D) {option_3}
\n    E) {option_4}

\n\n## Now it is your turn
\n\nPlease think step by step to provide your evaluation and provide the answerability X in JSON format {{'answerability': X}}, where X is in {{1, 2, 3}}:
\n\n
```

Table 12: Answerability Rating Prompt Template for Llama3. The difference to the prompt template for ChatGPT is highlighted in **bold**. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}. JSON-formatting is indicated by {{double curly brackets}}, as one level of brackets will be removed when the prompt template gets filled.

```
# Video Question Answering
\n\nHi there! Now that you have studied the topic of video question answering for years, you find yourself in the final exam of your studies. Please take your time to solve this task. You can do it! You know everything that is required to master it. Good luck!

\n\n## What is Video Question Answering?
\n\nVideo Question Answering is a task that requires reasoning about the content of a video to answer a question about it. In this exam, you will be given purely textual information about one or more clips of a video that has been extracted beforehand. So your task is to read the information about the video carefully and answer the question about it.

\n\n## Here is your task
\n\nBased on the given information about the most relevant clips of the video regarding the question, please select exactly one of the given options as your best answer to the given question. This is a single choice setting, such that there is exactly one best answer option. Think step by step to find the best candidate from the given answer options regarding the given question. Please write the letter of the best answer X in JSON format {{'best_answer': 'X'}}, where X is in {{'A', 'B', 'C', 'D', 'E'}}.

\n\n## Here is the information about the video
\n\n### Information about the most relevant clips of the video regarding the question
\n{whole_video_state}

\n\n### Question
\n\n{question}

\n\n### Five answer options (please select exactly one)
\n\n    A) {option_0}
\n    B) {option_1}
\n    C) {option_2}
\n    D) {option_3}
\n    E) {option_4}

\n\n## Now it is your turn
\n\nPlease choose the best option now. Think step by step and provide the best answer (friendly reminder: in the requested JSON format {{'best_answer': 'X'}}, where X is in {{'A', 'B', 'C', 'D', 'E'}}):
\n\n
```

Table 13: Question Answering Prompt Template for ChatGPT. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}. JSON-formatting is indicated by {{double curly brackets}}, as one level of brackets will be removed when the prompt template gets filled.

```
# Video Question Answering
\n\nHi there! Now that you have studied the topic of video question answering for years, you
find yourself in the final exam of your studies. Please take your time to solve this task.
You can do it! You know everything that is required to master it. Good luck!
\n\n## What is Video Question Answering?

\n\nVideo Question Answering is a task that requires reasoning about the content of
a video to answer a question about it. In this exam, you will be given purely textual
information about one or more clips of a video that has been extracted beforehand. So
your task is to read the information about the video carefully and answer the question about it.

\n\n## Here is your task
\n\nBased on the given information about the most relevant clips of the video regarding the
question, please select exactly one of the given options as your best answer to the given
question. This is a single choice setting, such that there is exactly one best answer option.
Think step by step to find the best candidate from the given answer options regarding the
given question. Please write the letter of the best answer X in JSON format {{'best_answer':
'X'}}, where X is in {{'A', 'B', 'C', 'D', 'E'}}. Make sure that you always select the best
answer option, even if it seems ambiguous or unsolvable.

\n\n## Here is the information about the video
\n\n### Information about the most relevant clips of the video regarding the question
\n{whole_video_state}

\n\n### Question
\n\n{question}

\n\n### Five answer options (please select exactly one)
\n\n    A) {option_0}
\n    B) {option_1}
\n    C) {option_2}
\n    D) {option_3}
\n    E) {option_4}

\n\n## Now it is your turn
\n\nPlease choose the best option now. Think step by step and provide the best answer
(friendly reminder: in the requested JSON format {{'best_answer': 'X'}}, where X is in {{'A',
'B', 'C', 'D', 'E'}}):
\n\n
```

Table 14: Question Answering Prompt Template for Llama3. The difference to the prompt template for ChatGPT is highlighted in **bold**. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}. JSON-formatting is indicated by {{double curly brackets}}, as one level of brackets will be removed when the prompt template gets filled.

```
# Assessment of Decision-Making
\n\nHi there! You are given an exam task and a students answer to the task.
\nYou are asked to assess the confidence level of the decision-making process in your students answer based on the information provided in the exam task. Imagine you are the teacher of the student and you want to know if you have provided enough information in the task to make a well-informed decision. At the same time, you want to know if the student has made a well-informed decision based on the information provided in the task.

\n\n## Here is the exam
\n\n{reasoning_history}

\n\n## Criteria for Evaluation
\n\n    1. Insufficient Information {{'confidence': 1}}: If information is too lacking for a reasonable conclusion.
\n    2. Partial Information {{'confidence': 2}}: If information partially supports an informed guess.
\n    3. Sufficient Information {{'confidence': 3}}: If information fully supports a well-informed decision.

\n\n## Assessment Focus
\nPlease evaluate based on the relevance, completeness, and clarity of the provided information in the task in relation to the decision-making context of the students answer.\nPlease provide the confidence in JSON format {{'confidence': X}} where X is in {{1, 2, 3}}.\n\n
```

Table 15: Self-Reflection Prompt Template for ChatGPT. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}. JSON-formatting is indicated by {{double curly brackets}}, as one level of brackets will be removed when the prompt template gets filled.

```
# Assessment of Decision-Making
\n\nHi there! You are given an exam task and a students answer to the task.
\nYou are asked to assess the confidence level of the decision-making process in your students answer based on the information provided in the exam task. Imagine you are the teacher of the student and you want to know if you have provided enough information in the task to make a well-informed decision. At the same time, you want to know if the student has made a well-informed decision based on the information provided in the task.

\n\n## Here is the exam
\n\n{reasoning_history}

\n\n## Criteria for Evaluation
\n\n    1. Insufficient Information {{'confidence': 1}}: If information is too lacking for a reasonable conclusion.
\n    2. Partial Information {{'confidence': 2}}: If information partially supports an informed guess.
\n    3. Sufficient Information {{'confidence': 3}}: If information fully supports a well-informed decision.

\n\n## Assessment Focus
\nPlease evaluate based on the relevance, completeness, and clarity of the provided information in the task in relation to the decision-making context of the students answer.\nPlease make sure that you always provide a confidence, even if it seems ambiguous or unsolvable. Please provide the confidence in JSON format {{'confidence': X}} where X is in {{1, 2, 3}}.\n\n
```

Table 16: Self-Reflection Prompt Template for Llama3. The difference to the prompt template for ChatGPT is highlighted in **bold**. Note that only linebreaks explicitly indicated with "\n" are true linebreaks at runtime – the linebreaks of this document are just for more readability. Parameters being filled at runtime are indicated with {coloured single curly brackets}. JSON-formatting is indicated by {{double curly brackets}}, as one level of brackets will be removed when the prompt template gets filled.

```
You are given some language descriptions of a first person view video. The video is 63
seconds long. Each sentence describes a 1.0s clip. The descriptions are sequential and
non-overlapping which cover the whole video exactly. Here are the descriptions: The camera
wearer pours the water in the. The camera wearer picks a. The camera wearer washes the plate.
The camera wearer washes the. The camera wearer washes the. The camera wearer scrapes the
container. The camera wearer washes the plate. The camera wearer washes the. The camera
wearer washes the. The camera wearer washes the tray with the sponge. The camera wearer
washes the. The camera wearer washes the. The camera wearer washes the. The camera wearer
washes the spoon. The camera wearer picks a. The camera wearer picks the bowl. The camera
wearer washes the tray. The camera wearer washes the. The camera wearer washes the. The
camera wearer washes the bowl. The camera wearer washes the. The camera wearer washes the.
The camera wearer washes the. The camera wearer washes the. The camera wearer washes the
tray. The camera wearer rinses the. The camera wearer pours water in the. The camera wearer
rinses the. The camera wearer washes the tray. The camera wearer washes the. The camera
wearer rinses the tray. The camera wearer closes the. The camera wearer lifts the basin.
The camera wearer holds the tray with both. The camera wearer washes the. The camera wearer
opens the. The camera wearer washes the tray with the sponge. The camera wearer washes the
tray with the. The camera wearer closes the. The camera wearer holds the tray. The camera
wearer opens the container. The camera wearer scrubs the. The camera wearer scrubs the
sink. The camera wearer scrubs the sponge with a sponge scrub. The camera wearer scrubs the.
The camera wearer scrubs the. The camera wearer scrubs the tray with a. The camera wearer
wipes the board with a sponge. The camera wearer scrubs the board with a. The camera wearer
squeezes the sponge. The camera wearer washes the chopping board. The camera wearer scrubs
the chopping board with a. The camera wearer washes the chopping board. The camera wearer
washes the chopping board with the. The camera wearer washes the. The camera wearer rinses
chopping board. The camera wearer washes the chopping board. The camera wearer rinses the
chopping. The camera wearer washes the chopping board. The camera wearer rinses the sponge.
The camera wearer washes the chopping board with the. The camera wearer opens the sink. The
camera wearer closes the dish.
Please give me a 180 words summary. When doing summarization, remember that your summary
will be used to answer this multiple choice question: Taking into account all the actions
performed by the camera wearer, what can you deduce about the primary objective and focus
within the video content?
```

Table 17: Action Caption Summarization Prompt Example for ChatGPT.

```
You are given a list of the most eye-catching objects that were detected in each frame of a video clip using a visual large language model. The list appears in the temporal order of the frames. The video is 63 seconds long. Each sentence describes the objects of a 1.0s clip. The object detections are sequential and non-overlapping which cover the whole video exactly. Here are the object detections:

Sink; Dish rack; Square dish. Sink; Dishwashing soap dispenser; Dish rack. Sink; Dish soap dispenser; Dish rack. Sink; Soap dispenser; Plastic bottle. Sink; Hand; Pan. Sink; Dish soap dispenser; Black pan. Sink; Dish soap dispenser; Plastic bottle. Sink; Dish soap dispenser; Plastic container. Sink; Hand; Dish soap. Sink; Dishwashing spray bottle; Dish rack. A sink; A dish rack; A person's hands. A sink; A faucet; A dish rack. Sink; Dishwashing soap dispenser; Dish rack. Sink; Dish rack; Soap dispenser. Sink; Plate with food remnants; Hand. Sink; Cutting board; Spray bottle. A sink; A hand washing dish soap dispenser; A red chopping board. A sink; A faucet; A spray bottle. A sink; A faucet; A bottle of dish soap. A sink; A black dish or container; A red cutting board. Sink; Dish soap dispenser; Plastic bottle. Sink; Hand; Plastic bottle. A sink; A faucet; A bottle of dish soap. Sink; Dish soap dispenser; Cutting board. Sink; Hands; Plastic bottle. Sink; Dishwashing soap dispenser; Plastic bottle. A black tray or dish; A white container or bowl; A bottle of liquid soap. Sink; Faucet; Dishwashing soap dispenser. Sink; Faucet; Dishwashing soap. A sink; A faucet; A dish rack. A black container; A white container; A faucet. A sink; A faucet; A black object (possibly a pan or a lid). A black plate; A silver dish rack; A silver sink with a faucet. A sink; A faucet; A dishwashing soap dispenser. A sink; A faucet; A dish rack. Sink; Plate; Cleaning spray bottle. Sink; Plate; Cleaning spray bottle. Sink; Plate; Dish soap. A sink; A white plate; A bottle of liquid. A white plate; A sink; A bottle. A green lid or cover; A red cutting board; A black container or pot. A white plate; A red cutting board; A bottle of cleaning solution. Plate; Sink; Dish rack. Sink; Plate; Dish rack. Sink; Dish rack; Plastic container. A white plate or dish; A metal dish rack; A sink. Sink; Dishwashing detergent bottle; Cutting board. A sink; A plate or tray; A bottle of dish soap. Sink; Plate; Cleaning bottle. A plate; A sink; A bottle of dish soap. A sink; A faucet; A bottle of dish soap. A sink; A dish rack; A bottle of dish soap. A sink; A dish rack; A bottle of dish soap. A sink; A dish rack; A bottle of dish soap. Sink; Plate; Cutting board. Sink; Plate; Soap dispenser. Sink; Plate; Dish soap dispenser. Sink; Plate; Dish soap. Sink; Plate; Soap dispenser. A sink; A dish rack; A bottle of dish soap. Sink; Plate; Dish soap. Sink; Dish soap dispenser; Red cutting board. A green container with a lid; A black frying pan or skillet; A metal dish rack.

Please give me a 180 words summary of these object detections. When doing summarization, remember that your summary will be used to answer this multiple choice question: Taking into account all the actions performed by the camera wearer, what can you deduce about the primary objective and focus within the video content?
```

Table 18: Action Caption Summarization Prompt Example for ChatGPT.