# Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition

**Hao Fei** [1] **Shengqiong Wu** [1] **Wei Ji** [1] **Hanwang Zhang** [2] **Mong-Li Lee** [1] **Wynne Hsu** [1]

## Abstract

Existing research of video understanding still struggles to achieve in-depth comprehension and reasoning in complex videos, primarily due to the under-exploration of two key bottlenecks: *fine-grained spatial-temporal perceptive understanding* and *cognitive-level video scene comprehension*. This paper bridges the gap by presenting a novel solution. We first introduce a novel video Multimodal Large Language Model (MLLM), **MotionEpic**, which achieves fine-grained pixel-level spatial-temporal video grounding by integrating video spatial-temporal scene graph (STSG) representation. Building upon MotionEpic, we then develop a Video-of-Thought (**VoT**) reasoning framework. VoT inherits the Chain-of-Thought (CoT) core, breaking down a complex task into simpler and manageable subproblems, and addressing them step-by-step from a low-level pixel perception to high-level cognitive interpretation. Extensive experiments across various complex video QA benchmarks demonstrate that our overall framework strikingly boosts existing state-of-the-art. To our knowledge, this is the first attempt at successfully implementing the CoT technique for achieving human-level video reasoning, where we show great potential in extending it to a wider range of video understanding scenarios. Systems and codes will be open later.

## 1. Introduction

Enabling learning models to accurately interpret video data is one of the most paramount goals in the relevant community. In the current research, while there has been extensive exploration into building models for video action and dynamics recognition (Lei et al., 2018; Bertasius et al., 2021), mostly they fall prey to the type of straightforward perceptual-level understanding, i.e., for simple videos (Zolfaghari et al., 2018; Lin et al., 2019). And there remains a significant gap in research concerning comprehending and reasoning about complex videos in depth, an imperative capability urgently needed in real-world applications. Compared to shallow video perception, reasoning about complex videos poses greater challenges: it demands not only an intricate understanding of the video's spatiotemporal characteristics (Caballero et al., 2017), but also a profound grasp of the underlying implications behind pixels.

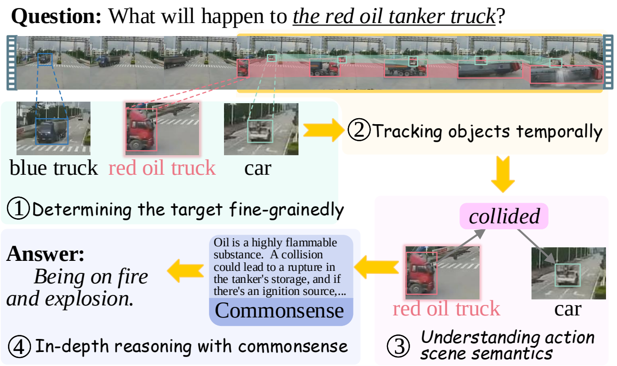

Question: What will happen to the red oil tanker truck?
blue truck
red oil truck
car
②Tracking objects temporally
①Determining the target fine-grainedly
collided
Answer:
Being on fire and explosion.
Oil is a highly flammable substance. A collision could lead to a rupture in the tanker's storage, and if there's an ignition source,...
Commonsense
red oil truck
car
④In-depth reasoning with commonsense
③ Understanding action scene semantics

Figure 1: Human-like video reasoning intuitively follows a multi-step procedure, from lower-level perceptive fine-grained pixel grounding and tracking, to higher-level cognitive action scene semantics understanding.

Drawing from human cognition patterns, we mark that reasoning about videos, especially for the complex ones, requires superior mastery in two points: perceptual capability of pixel understanding and cognitive ability for semantic understanding. **Firstly**, to achieve precise content perception, a fine-grained perceptive pixel understanding of the video movement is necessary. Most existing video understanding approaches focus on instance or patch-level analysis (Yuan et al., 2021; Neimark et al., 2021), lacking the precision for detailed granular control and accurate object-level recognition or tracking, let alone in-depth video comprehension. **Secondly**, profound reasoning demands cognitive capabilities allowing reasonable explanation and even causal imagination, i.e., with a reservoir of commonsense knowledge to link video pixels to the factual world. For example, understanding that jumping from a height can cause fractures, or that colliding with a tanker truck can cause an explosion. **Most importantly**, for humans, video reasoning is not an instantaneous process but follows a multi-hop procedure from lower level to higher level. This often involves first

[1]National University of Singapore [2]Nanyang Technological University. Correspondence to: Hanwang Zhang <hanwangzhang@ntu.edu.sg>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

identifying specific targets, like a "red oil truck" (cf. Fig. 1) in the video frames, then tracking and analyzing its temporal behaviors and interactions with the environment to deduce the scene semantics, and finally, integrating factual commonsense to formulate a cognitively coherent response.

Recently, the community of MLLMs has seen rapid advancement, exhibiting formidable data understanding and reasoning capabilities, among which video MLLMs have been extensively developed, such as Video-LLaMA (Zhang et al., 2023a), Video-ChatGPT (Maaz et al., 2023), and Video-LLaVA (Lin et al., 2023). Simultaneously, there is a growing interest in integrating CoT prompting technique (Wei et al., 2022) to augment the reasoning capabilities of LLMs. CoT works by intuitively breaking down a complex problem into a chain of simpler and more manageable sub-problems, facilitating a human-like reasoning process. While this technique has flourished in language understanding tasks extensively (Wang et al., 2022a), unfortunately, a CoT-based reasoning framework specifically tailored for video input with video MLLMs is yet under-explored.

To this end, this paper is dedicated to devising a solution that enables human-like complex video reasoning. We first propose the integration of a STSG representation (Ji et al., 2020), modeling both the input video and its STSG representation, where fine-grained spatial-temporal features are carefully integrated and modeled. To implement this, we introduce a novel video LLM, named **MotionEpic** (cf. Fig. 2), which, based on a similar architecture as existing video MLLMs, supports not only video input but also the encoding, understanding and generation of STSGs. To enable MotionEpic with fine-grained pixel-level spatial-temporal grounding between videos and STSGs, we also investigate various distinct video-STSG training objects. STSG annotations are used during the grounding-aware tuning phase, while in the subsequent stage, the system is learned to autonomously parse STSG, and thus supports STSG-free inference and reasoning for downstream tasks.

Building upon MotionEpic, we next design a novel reasoning framework, named Video-of-Thought (**VoT**), cf. Fig. 4. Inheriting the key spirit of CoT, VoT breaks down the raw intricate video reasoning problem into a chain of simpler sub-problems, and solves them one by one sequentially. These sub-questions follow a progression from lower to higher level, i.e., starting with pixel grounding for a precise understanding of target content, and then accurately interpreting corresponding semantic signals. ❶ Given an input video and a question, VoT identifies the possible target(s) involved in the question to observe. ❷ The system then grounds the temporal tracklet(s), which serves as supporting evidence/rationale for content perception in subsequent analysis. ❸ Combined with factual commonsense, VoT next interprets the target object's trajectory and its interactions with neighboring scenes to thoroughly understand the action dynamics and semantics. ❹ With in-depth understanding of the target actions in the video, we then carefully examine each optional answer with commonsense knowledge, where the final result is output after ranking those candidates. ❺ Finally, VoT performs verification for the answer from both pixel grounding perception and commonsense cognition perspectives, ensuring the most factually accurate result.

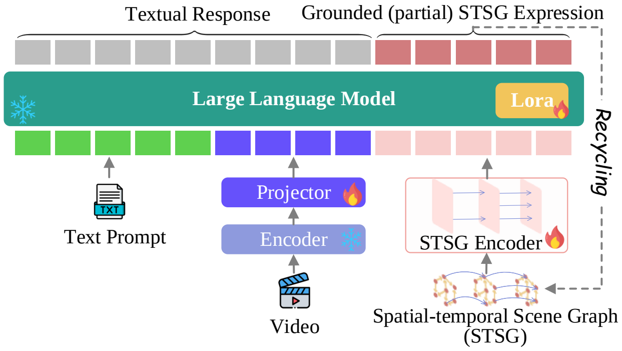


Figure 2: Overview of the MotionEpic video MLLM.

Our experiments mainly focus on video Question Answering (QA), a representative task reliant on in-depth video reasoning. We evaluate our system across 8 complex video QA benchmarks, where it strikingly boosts the current performances in both fine-tuning and zero-shot settings by max to 8.8% and 11.6% accuracy respectively, establishing a series of new states of the arts. We further conduct in-depth analyses of MotionEpic's capabilities in video grounding, and probe the video reasoning ability of VoT framework, providing insights into how the framework advances. To summarize, this work contributes in multiple aspects:

- proposing the first video Chain-of-Thought reasoning framework, VoT, which decomposes raw complex problems into a chain of sub-problems, and reasons through multiple steps from low to high levels, enabling not only pixel perceptive recognition but also semantic cognitive understanding of videos.
- contributing a novel video MLLM, MotionEpic, which supports fine-grained pixel-level spatial-temporal video grounding via STSG encoding and generation.
- empirically setting new state-of-the-art (SoTA) performances in a range of video QA benchmarks that require intricate reasoning capability.

## 2. Related Work

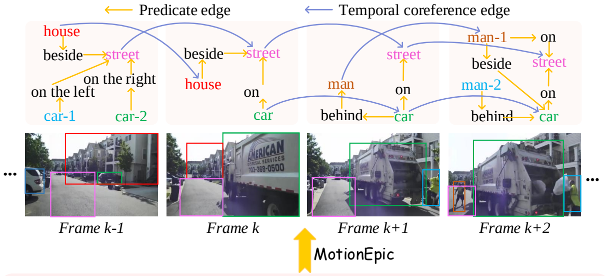
A key objective in the intelligence community is the understanding of various modalities of data. Currently, with the advent of LLMs such as ChatGPT (OpenAI, 2022a), we have attained unprecedented language reasoning capabilities, on par with the human level. This is largely due to the vast repository of commonsense knowledge and semantic understanding capabilities inherent in LLMs, enabling provide plausible causal explanations and even engage in imaginative reasoning. Particularly, the integration of the recent trending CoT technology, which deconstructs a problem into its constituent parts and provides rationale at each step, has made the reasoning process more reliable. As for image understanding, the rapid development of MLLMs, e.g., LLaVA (Liu et al., 2023), GPT-4V (OpenAI, 2022b), has also nearly achieved substantial comprehension ability. However, unlike language and images, video understanding or reasoning presents a dual challenge of static spatial and temporal dynamics.

Historically, earlier video understanding research efforts predominantly learn neural models over small-size in-domain training dataset (Zolfaghari et al., 2018; Lin et al., 2019). However, these 'small' models are limited to relatively superficial levels of perception, lacking the depth of human-level cognition. As a result, previous methods were mostly confined to the shallow understanding of simple videos, such as identifying contents and movements within a video. Unlike simple video comprehension, which relies mainly on perceptive abilities, understanding complex videos necessitates deeper cognitive reasoning, such as *explaining why certain actions occur in a video or hypothesizing potential outcomes*. Although MLLMs supporting video data have been developed (Li et al., 2023c; Zhang et al., 2023a), offering greater video understanding capabilities than smaller models, the research into penetrating beyond the perceptual surface of videos to deeply understand the implied semantic content and perform cognitive-level reasoning is still insufficiently explored. We observe that current video MLLMs either fail to achieve fine-grained spatial-temporal understanding of videos, or do not fully leverage the rich commonsense knowledge and causal reasoning inherent in LLMs for enhanced cognitive-level comprehension.

Meanwhile, we note a lack of research specifically focused on integrating CoT technology into video MLLM scenarios to establish a powerful video reasoning framework. To bridge this gap, this paper takes the initiative, and introduces the concept of Video-of-Thought. Unlike the original CoT approach that attempts to improve outputs with a simple "*Let's Think Step By Step*" prompt (Wei et al., 2022), we implement a more genuine thought chain. We encourage MLLM to first decompose the original problem into a series of more manageable sub-solutions before the model initiates reasoning, following the human-cognitive procedure from low-level pixel grounding and understanding to high-level cognitive semantic meaning inference, ultimately achieving human-level video understanding and reasoning capabilities.

## 3. MotionEpic: Fine-grained Spatial-temporal Grounded Video MLLM

In this section, we describe the MotionEpic video MLLM, and elaborate on how the STSGs are integrated, as well as the fine-grained spatial-temporal grounding-aware tuning.



```
Frame k-1: {
        Objects: [{"car-1": [0.0,13.4,7.0,8.1]}, {"car-2": [27.0,12.9,9.0, 4.0]},
                  {"house": [15.0, 0.0,34.5,17.0]}, {"street": [9.0,13.5,16.5,15.0]}],
        Triplets: [("car-1", "on the left", "street"), ("car-2", "on the right", "street"), ("house",
                  "beside", "street")]
}
...
Frame k+2: {
        Objects: [{"car": [9.0,2.4,33.7,25.6]}, {"man-1": [1.8,16.4,4.2,10.0]},
                  {"man-2": [43.0,14.4, 6.2,12.0]}, {"street": [17.9,0.0,10.2,10.0]}],
        Triplets: [("man-1", "beside", "car"), ("car", "on", "street"), ("man-2", "behind", "car"),...]
}
...
Temporal Coreference Edge: [
        {Frame k-1→ k: [("house", "house"), ("street", "street")]},
        ...
]
```

Grounded (partial) STSG Expression

Figure 3: The STSG expression generated by MotionEpic, with its corresponding structural STSG illustration.

### 3.1. Architecture Briefing

Fig. 2 presents a schematic overview of MotionEpic, where MotionEpic takes as input three sources: text prompt, video, and STSG representation of video. We follow the most common practice, and employ the Vicuna-7B (v1.5) (Chiang et al., 2023) as the backbone LLM. To perceive video input, we adopt the ViT-L/14 encoder (Dosovitskiy et al., 2020) and Q-Former projector (Li et al., 2023a). We also design MotionEpic to support the STSG signal, where we retrofit the Graph Transformer (Dwivedi & Bresson, 2020) with recurrent propagation to encode the multi-frame STSG information. Appendix §A.1 details STSG encoding.

### 3.2. Integrating STSG Representation

By definition (Ji et al., 2020), an STSG consists of a sequence of single SGs corresponding to all video frames, with each SG comprising triplets in the video frame, i.e., '*subject*'-'*predicate*'-'*object*', where '*subject*' and '*object*' refer to two visual proposals (RoIs) that are connected with the '*predicate*' relationship. STSG intuitively depicts the underlying core semantics representations of videos while filtering the less-informative background information, aiding the perceptive understanding of videos (Zhao et al., 2023). Also, such fine-grained structural feature helps effectively model the compositional spatiotemporal semantics.

In our practice, we slightly retrofit the vanilla STSG definition to meet the demand in our reasoning framework. Since a video has redundant temporal contents across frames, we first evenly sample the frames (with a proper sampling rate), which can effectively reduce computation costs. We denote each single SG at $k$-th frame as $G_k$=$(V_k; E_k)$, where $V_k$ is a list of nodes, i.e., object proposal, and $E_k$ is a list of predicate edges. For each object proposal $v_{k,i}$ we record the category label $c_i$, the proposal's neural representation $f_i$, and the bounding box (bbox) annotation $b_i=(x, y, w, h)$), i.e., the 2D coordinate in the image. Thus, each $v_{k,i}=(c_i, f_i, b_i)_k$. All nodes (i.e., $v_{k,i}$ and $v_{k,j}$) are connected with edges $e_{k,i,j}$. To enhance the connectivity of STSG, we further create a type of *temporal coreference edges* across each single-frame SG, where the same objects are linked together with time-persistent edges, $e_{k-1\rightarrow k}$, mimicking the 'tracking' process.

MotionEpic achieves fine-grained spatial-temporal video grounding by simultaneously understanding and generating STSGs. After full tuning (cf. §3.3), MotionEpic can directly output (partial) STSG based on input video (with text prompts), essentially grounding the specific portions of the video content as indicated in the input prompts. In Fig. 3 we illustrate how the generated STSG expression corresponds to the structural STSG. Further, the output STSG serving as the rationale will be recycled in the system, i.e., repurposed as the input for the subsequent round.

### 3.3. Fine-grained Video-Scene Grounding-aware Tuning

Intuitively, we expect our system to perform video reasoning for downstream tasks without relying on any external STSG annotations, i.e., STSG-free inference. This requires an accurate spatial-temporal grounding between videos and STSGs. To this end, we carry out tuning for MotionEpic such that it is learned to autonomously parse STSG according to input instructions. The grounding-aware tuning is performed based on video-STSG pairs. We design various training objects, which can be further divided into coarse-grained and fine-grained levels:

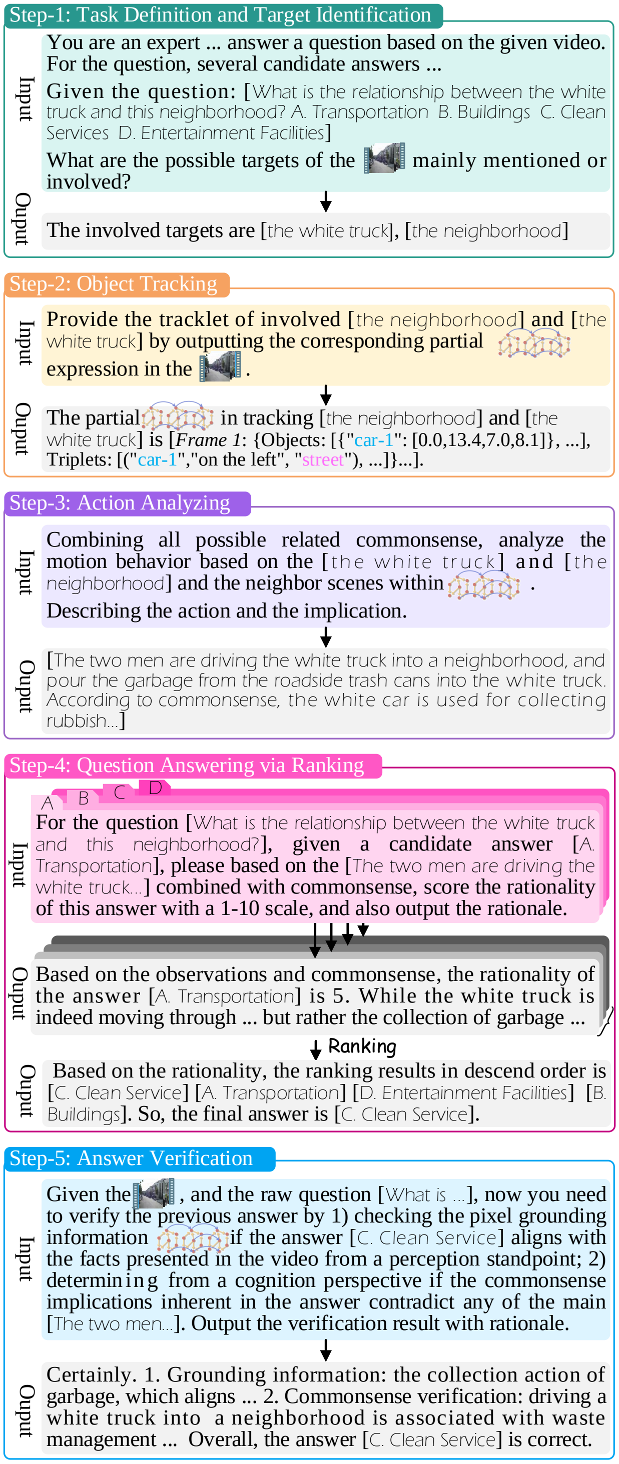
**1) Enhancing coarse-grained correspondence**:

- $\mathcal{L}_1$: predicting if the overall input video and STSG are paired.
- $\mathcal{L}_2$: given a video, generating the whole STSG (expression) of the video.

**2) Enhancing fine-grained correspondence**:

- $\mathcal{L}_3$: given a video and action description(s), outputting the corresponding object tracklet(s), i.e., a partial STSG.
- $\mathcal{L}_4$: given a video and key object(s), describing the corresponding temporal action(s) in textual response, and outputting the corresponding object tracklet(s).
- $\mathcal{L}_5$: given a video and a bbox of a certain frame's object, outputting the object label, as well as the corresponding tracklet.

For each learning objective, we wrap up the inputs with instruction-tuning (Liu et al., 2023) style question-answer pairs, being consistent with the following downstream inference. Appendix §B.2 extends more details about the grounding-aware tuning. Overall, except for the STSG encoder and video projector, the video encoder and the backbone LLM are kept frozen throughout all the learning stages. To tune the LLM, we leverage LoRA (Hu et al., 2022) to enable a small subset of parameters to be updated.



You are an expert ... answer a question based on the given video. For the question, several candidate answers ...

Given the question: [What is the relationship between the white truck and this neighborhood? A. Transportation B. Buildings C. Clean Services D. Entertainment Facilities]

What are the possible targets of the mainly mentioned or involved?

The involved targets are [the white truck], [the neighborhood]

Provide the tracklet of involved [the neighborhood] and [the white truck] by outputting the corresponding partial expression in the .

The partial in tracking [the neighborhood] and [the white truck] is [*Frame 1*: {Objects: [{"car-1": [0.0,13.4,7.0,8.1]}, ...], Triplets: [("car-1","on the left", "street"), ...]}...].

Combining all possible related commonsense, analyze the motion behavior based on the [the white truck] and [the neighborhood] and the neighbor scenes within .

Describing the action and the implication.

[The two men are driving the white truck into a neighborhood, and pour the garbage from the roadside trash cans into the white truck. According to commonsense, the white car is used for collecting rubbish...]

For the question [What is the relationship between the white truck and this neighborhood?], given a candidate answer [A. Transportation], please based on the [The two men are driving the white truck...] combined with commonsense, score the rationality of this answer with a 1-10 scale, and also output the rationale.

Based on the observations and commonsense, the rationality of the answer [A. Transportation] is 5. While the white truck is indeed moving through ... but rather the collection of garbage ...

Based on the rationality, the ranking results in descend order is [C. Clean Service] [A. Transportation] [D. Entertainment Facilities] [B. Buildings]. So, the final answer is [C. Clean Service].

Given the , and the raw question [What is ...], now you need to verify the previous answer by 1) checking the pixel grounding information if the answer [C. Clean Service] aligns with the facts presented in the video from a perception standpoint; 2) determining from a cognition perspective if the commonsense implications inherent in the answer contradict any of the main [The two men...]. Output the verification result with rationale.

Certainly. 1. Grounding information: the collection action of garbage, which aligns ... 2. Commonsense verification: driving a white truck into a neighborhood is associated with waste management ... Overall, the answer [C. Clean Service] is correct.

Figure 4: An illustrative view of VoT framework. The complete I/O and prompts are detailed in Appendix §A.2.

## 4. Video-of-Thought Reasoning Framework

Based on MotionEpic, we now perform video reasoning with VoT. Different from the vanilla CoT with one straightforward prompt, i.e., "*Let's think step by step*", VoT divides the raw problem into much smaller and finer-grained sub-problems. We consider an exact paradigm of task decomposition, which encompasses five chained steps, following a process from low-level perceptive pixel grounding to high-level cognitive semantic comprehension. In Fig. 4 we illustrate the overall VoT framework.

### ► Step-1: Task Definition and Target Identification

First, MotionEpic is fed with the raw video along with the text prompt of the task definition, format, and raw question, all of which serve as the background foundation information of the reasoning. As the initial phase, we expect to identify the target within the video that requires analysis, which is a crucial prerequisite for determining the subsequent in-depth reasoning. It is noteworthy that sometimes the question may explicitly include targets visible in the video, or implicitly involve related targets. Therefore, we proceed to prompt the model, to infer from the question what the target object(s) involved or related to in the video might be:

> Given the question [`Question`], what are the possible targets of the mainly mentioned or involved?

After this step, all the possible [`Target`] involved in the question will be confirmed.

### ► Step-2: Object Tracking

In the second step, we aim to further ground the object's full spatial-temporal characteristics, i.e., to track the target's trajectory. We note that grounding the targets' temporal tracking is pivotal for pursuing fine-grained video understanding, as only accurately perceiving the behaviors in the video can ensure that the subsequent cognitive-level understanding is meaningful. In this work, we leverage the STSG for the temporal grounding, rather than directly tracking the original video frames. Such semantic representation carried by STSG is highly concise, ensuring that the tracking of the video's target is more accurate and reliable. Also notably, object tracking and pixel grounding based on the STSG can effectively mitigate the hallucination issues (Zhang et al., 2023b) inherent in existing MLLMs.

Having performed grounding-aware tuning, MotionEpic develops the full capability to ground from object to (partial) STSG. Therefore, we directly prompt the model with:

> Provide the tracklet of involved [`Target`] by outputting the corresponding partial expression.

The yielded grounded [`Target Tracklet`] of STSG will serve as low-level evidence (i.e., supporting rationale) for the next step of behavior analysis.

### ► Step-3: Action Analyzing

In this step, VoT further analyze the corresponding actions and behaviors by integrating the target tracking in STSG. For an accurate understanding of the target object's motion, merely observing the target itself is insufficient. This process should also reference the higher-order neighbor nodes within the STSG representation, interacting targets with their neighboring scenes. On the other hand, directly inferring actions from video pixels alone is still inadequate, as interpretations based solely on pixel information often remain superficial. Therefore, we further prompt the model to consider more potentially relevant commonsense knowledge, allowing the model to connect video pixel observations with the factual world, achieving a more in-depth understanding of the video. Given that MLLMs possess the necessary repository of commonsense knowledge via extensive pre-training, all that is required is to properly prompt the model:

> Combining all possible related commonsense, analyze the motion behavior based on the [`Target Tracklet`] and the neighbor scenes within . Describing the action observations and implications.

This step yields the target action's [`Observation and Implication`].

### ► Step-4: Question Answering via Ranking

Having established an in-depth understanding of the target actions in the video, we can now consider answering the original question. We contemplate a multiple-choice QA format, where multiple candidate answers are provided.[1] Inspired by the human pattern of answering multi-choice questions, we also consider a ranking mechanism to determine the final answer. Specifically, for each candidate answer, we prompt the model to score its likelihood (from 1 to 10) in conjunction with commonsense knowledge, and provide a corresponding rationale:

> For the question [`Question`], given a candidate answer [`Answer`], please based on the action's [`Observation and Implication`] combined with commonsense, score the rationality of this answer with a 1-10 scale, and also output the rationale.

We then rank the scores of all options and select the most optimal answer [`Answer`].

### ► Step-5: Answer Verification

Given that complex video task often involves intricate questions and answers, and the entire reasoning process encompasses lengthy chained steps, it is essential to verify the answer provided in the previous step. Our basic idea to verification is that, assuming that answer A is correct, we retrospectively evaluate whether the answer results in contradictions with the input question and video in two aspects: 1) First, check the pixel grounding information if it aligns with the facts presented in the video from a perception standpoint. 2) On the other hand, prompt the model again from a cognition perspective to determine if the commonsense implications inherent in the answer contradict any of the main observations inferred in the 3-$rd$ reasoning step.

[1] Note that for open-ended QA, we consider prompting the model to output multiple distinct optional answers, such that we unify different types of QA formats into a multi-choice format. For the open-ended QA format, we detail processing and prompt methods in Appendix §A.2.

> Given the 🎞, and the raw question [Question], now you need to verify the previous answer by
> 1) checking the pixel grounding information if the answer [Answer] aligns with the facts presented in the video from a perception standpoint;
> 2) determining from a cognition perspective if the commonsense implications inherent in the answer contradict any of the main [Observations] inferred in the 3-$rd$ reasoning step.
> Output the verification result with rationale.

If any inconsistencies are found in perception and cognition perspectives, we record the corresponding rationale, and re-execute the 4-$th$ step to reselect the answer. This approach ensures that the final outcome is the most factually accurate.

## 5. Experiments

### 5.1. Settings

**Task and Data.** While in theory all video understanding tasks could benefit from our reasoning framework, we mainly focus on the most representative task, video QA. For fine-tuning setting, we adopt 6 benchmarks characterizing complex video QA where advanced video abilities, e.g., explanation, causality, foresight and imagination are required: VLEP (Lei et al., 2020), STAR (Wu et al., 2021), IntentQA (Li et al., 2023b), Social-IQ (Zadeh et al., 2019), Causal-VidQA (Li et al., 2022a) and NExT-QA (Xiao et al., 2021). For zero-shot setting, we further MSR-VTT (Xu et al., 2016) and ActivityNet (Heilbron et al., 2015) datasets.

**Grounding-aware Tuning Corpus.** To construct the video-STSG pairs, we leverage the Action Genome data (Ji et al., 2020), which contains 10K high-quality manual annotated STSGs of videos. To enrich the data amount, we also use part of WebVid-10M videos (Bain et al., 2021), where we select 350K videos with clear actions, and parse the STSGs via SoTA parser (Li et al., 2022b).

**Baselines and Implementations.** The evaluations are compared with recent SoTA baselines of these complex video QA datasets, including InternVideo (Wang et al., 2022b), LLaMA-VQA (Ko et al., 2023), VLAP (Wang et al., 2023), SeViLA (Yu et al., 2023), TranSTR (Li et al., 2023f) and HiTeA (Ye et al., 2023). The results are faithfully copied from their papers. Also we reimplement current video MLLMs, including VideoChat2 (Li et al., 2023e), Video-LLaMA (Zhang et al., 2023a), Video-ChatGPT (Maaz et al., 2023), VideoChat (Li et al., 2023d) and Video-LLaVA (Lin et al., 2023). For fairness, we compare MotionEpic with these video MLLMs in vanilla CoT setting. Further, we also implement the Video-LLaVA by integrating the STSG features. In MotionEpic, the recurrent Graph Transformer has 6 layers with 768-d hidden sizes. The object neural representation $f_i$ is obtained via CLIP, which will be used as node embedding initiation. We adopt accuracy as the main metric of task performance. The results from our implementation are averaged over 5 random runs.

Table 1: Results on four VideoQA datasets. STAR data includes four subsets: Interaction (Int.), Sequence (Seq.), Prediction (Pre.), Feasibility (Fea.). The best scores of baselines are underlined, and the new best results are **bold**, the improvements between two of which are marked.

| Model | VLEP | STAR | | | | IntentQA | Social-IQ | |
|---|---|---|---|---|---|---|---|---|
| | | Int. | Seq. | Pre. | Fea. | | 2-Way | 4-Way |
| **• SoTA baselines** | | | | | | | | |
| InternVideo | 63.9 | 62.7 | 65.6 | 54.9 | 51.9 | - | - | - |
| LLaMA-VQA | 71.0 | 66.2 | 67.9 | 57.2 | 52.7 | - | - | - |
| VLAP | 69.6 | 70.0 | 70.4 | 65.9 | 62.2 | - | - | - |
| SeViLA | 68.9 | 63.7 | 70.4 | 63.1 | 62.4 | - | - | - |
| VideoChat | 62.0 | 63.2 | 66.8 | 54.1 | 49.6 | 59.3 | 67.7 | 37.8 |
| Video-LLaVA | 65.8 | 64.3 | 67.0 | 56.5 | 50.1 | 62.5 | 68.9 | 39.2 |
| **• CoT** | | | | | | | | |
| Video-LLaVA | 65.7 | 65.0 | 67.7 | 57.8 | 52.0 | 65.2 | 70.5 | 42.4 |
| Video-LLaVA+STSG | 68.0 | 65.9 | 68.9 | 60.0 | 55.7 | 67.9 | 71.8 | 43.7 |
| MotionEpic | 70.2 | 67.8 | 69.6 | 63.6 | 60.4 | 70.1 | 73.2 | 45.0 |
| **• VoT** | | | | | | | | |
| MotionEpic | **76.4** | **74.5** | **74.6** | **71.0** | **67.7** | **76.8** | **78.8** | **53.0** |
| | ↑5.4 | ↑4.5 | ↑4.2 | ↑5.1 | ↑5.5 | ↑6.7 | ↑5.6 | ↑8.0 |

Table 2: Results on Causal-VidQA data. D: Description, E: Explanation, P: Prediction, C: Counterfactual.

| Model | Acc@D | Acc@E | Acc@P | | | Acc@C | | |
|---|---|---|---|---|---|---|---|---|
| | | | A | R | AR | A | R | AR |
| **• SoTA baselines** | | | | | | | | |
| TranSTR | 73.6 | 75.8 | 65.1 | 65.0 | 48.9 | 68.6 | 65.3 | 50.3 |
| Video-LLaMA | 69.2 | 71.0 | 63.6 | 62.4 | 44.4 | 65.4 | 60.1 | 45.0 |
| VideoChat | 73.9 | 74.9 | 66.2 | 64.1 | 46.9 | 67.0 | 63.7 | 47.8 |
| Video-ChatGPT | 77.1 | 78.1 | 68.0 | 66.9 | 49.0 | 70.8 | 65.6 | 50.9 |
| Video-LLaVA | 78.7 | 78.4 | 68.6 | 67.4 | 52.7 | 71.0 | 67.9 | 52.5 |
| **• CoT** | | | | | | | | |
| Video-LLaVA | 79.8 | 79.0 | 69.4 | 67.7 | 53.3 | 72.3 | 66.7 | 52.9 |
| Video-LLaVA+STSG | 81.0 | 80.9 | 70.9 | 69.2 | 55.0 | 74.7 | 68.2 | 54.6 |
| MotionEpic | 83.5 | 82.2 | 72.8 | 72.8 | 56.4 | 76.2 | 69.7 | 55.8 |
| **• VoT** | | | | | | | | |
| MotionEpic | **89.2** | **91.0** | **79.3** | **80.7** | **61.7** | **81.5** | **76.8** | **60.6** |
| | ↑5.7 | ↑8.8 | ↑6.5 | ↑7.9 | ↑5.3 | ↑5.3 | ↑7.1 | ↑4.8 |

More implementation and specification details are presented in Appendix § B.

### 5.2. Main Performance on Video QA Reasoning

In Table 1, 2 and 3 we present the main results of different systems. Overall, our MotionEpic under the VoT reasoning framework has boosted all the SoTA baselines by very large margins consistently. Notably, ours improves on Social-IQ 4-way by 8.0% accuracy, and further improves on Explanation by 8.8% accuracy. Beyond enhanced performance, we further gain some key observations. First, by observing Video-LLaVA without/with CoT prompting, we see that the improvement from CoT for video reasoning could be quite limited. Further, by comparing Video-LLaVA without/with STSG integration, we notice that the structural fine-grained STSG features play a positive role in understanding videos. Third, by comparing Video-LLaVA+STSG with our MotionEpic under the same CoT, it is clear that the implicit integration of the scene graph features is quite superior to the explicit integration. Also, even our MotionEpic with vanilla

Table 3: Results on NExT-QA data.

| **Model** | **Acc@All** | **Acc@C** | **Acc@T** | **Acc@D** |
|---|---|---|---|---|
| **• SoTA baselines** | | | | |
| InternVideo | 63.2 | 62.5 | 58.5 | 75.8 |
| HiTeA | 63.1 | 62.4 | 58.3 | 75.6 |
| LLaMA-VQA | 72.0 | 72.7 | 69.2 | 75.8 |
| SeViLA | 73.8 | 73.8 | 67.0 | 81.8 |
| VLAP | 75.5 | 74.9 | 72.3 | 82.1 |
| Video-LLaMA | 60.6 | 59.2 | 57.4 | 72.3 |
| VideoChat | 61.8 | 63.5 | 61.5 | 74.6 |
| Video-ChatGPT | 64.4 | 66.9 | 64.1 | 77.7 |
| Video-LLaVA | 71.3 | 70.7 | 66.8 | 78.9 |
| **• CoT** | | | | |
| Video-LLaVA | 72.7 | 72.0 | 68.9 | 80.5 |
| Video-LLaVA+STSG | 74.0 | 74.6 | 69.6 | 80.9 |
| MotionEpic | 75.2 | 75.4 | 71.1 | 81.7 |
| **• VoT** | | | | |
| MotionEpic | **81.0** ↑5.5 | **82.8** ↑7.4 | **78.6** ↑6.3 | **85.8** ↑3.7 |

Table 4: Zero-shot Video QA results. Verify-G/C: verification in terms of Grounding and Commonsense perspectives.

| **Model** | **MSR-VTT** | **ActivityNet** | **NExT-QA** | **STAR** | ***AVG.*** |
|---|---|---|---|---|---|
| **• Zero-shot SoTA baselines** | | | | | |
| InternVideo | - | - | 49.1 | 41.6 | - |
| Video-LLaMA | 49.6 | 21.4 | 43.5 | 36.4 | 37.7 |
| VideoChat | 52.0 | 26.5 | 52.8 | 45.0 | 44.1 |
| Video-ChatGPT | 54.3 | 35.2 | 53.0 | 48.7 | 47.8 |
| Video-LLaVA | 59.2 | 45.3 | 57.3 | 50.6 | 53.1 |
| VideoChat2 | 54.1 | 49.1 | 61.7 | 59.0 | 56.0 |
| **• CoT** | | | | | |
| Video-LLaVA | 60.1 | 48.9 | 60.5 | 54.0 | 55.9 |
| Video-LLaVA+STSG | 63.5 | 51.4 | 63.0 | 56.7 | 58.7 |
| MotionEpic | 64.4 | 52.0 | 63.9 | 58.5 | 59.7 |
| **• VoT** | | | | | |
| MotionEpic | **69.8**<br>↑5.4 | **59.6**<br>↑7.6 | **75.5**<br>↑11.6 | **68.4**<br>↑9.4 | **68.3**<br>↑8.6 |
| w/o Verify-G | 67.1<br>↓2.7 | 56.4<br>↓3.2 | 71.0<br>↓4.5 | 64.6<br>↓3.8 | 64.8<br>↓3.5 |
| w/o Verify-C | 69.1<br>↓0.7 | 58.4<br>↓1.2 | 71.8<br>↓3.7 | 64.4<br>↓4.0 | 65.9<br>↓2.4 |

CoT we beat the SoTA methods on certain datasets. Lastly, observing the MotionEpic under CoT and our proposed VoT, we see there are huge performance gaps in between consistently on all reasoning scenarios and tasks, indicating the great potential of our proposed video reasoning framework.

### 5.3. Zero-shot Performance

We then examine the performance in zero-shot setting. Table 4 presents the comparisons. In general, we can notice that CoT exhibits stronger improvements than direct prompting methods under the zero-shot scenario, compared with the scenario of the above fine-tuning. Notedly, the improvements by VoT over the CoT become larger and clearer under the zero-shot setting. Specifically, compared with the fine-tuned results on NExT-QA and STAR, the improvements by ours climb to 11.6% and 9.4%, respectively. The enhancements on these two complex video QA tasks are clearer than those on the comparatively simpler tasks, i.e., MSR-VTT and ActivityNet. This is largely because the latter datasets more tend to perceptive understanding (e.g., describing *what's in video*), rather than cognitive understanding (e.g., explanation, foresight or imagination). Further we cancel the verification mechanism (at 6-$th$) of either the pixel grounding perspective or the commonsense perspective. We see that on MSR-VTT and ActivityNet, the perceptive-level pixel grounding verification is more crucial than the commonsense cognitive verification. For those complex videos, both types of verifications are pivotal.

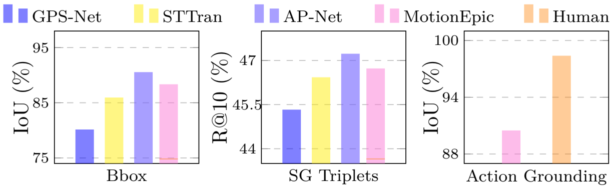


Figure 5: MotionEpic performance on object grounding, scene graph triplet classification, and action grounding.

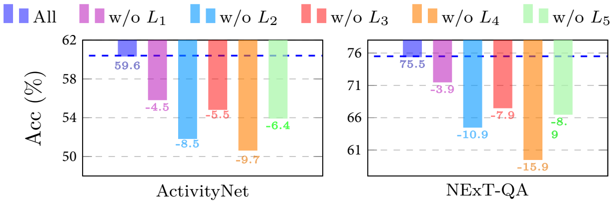
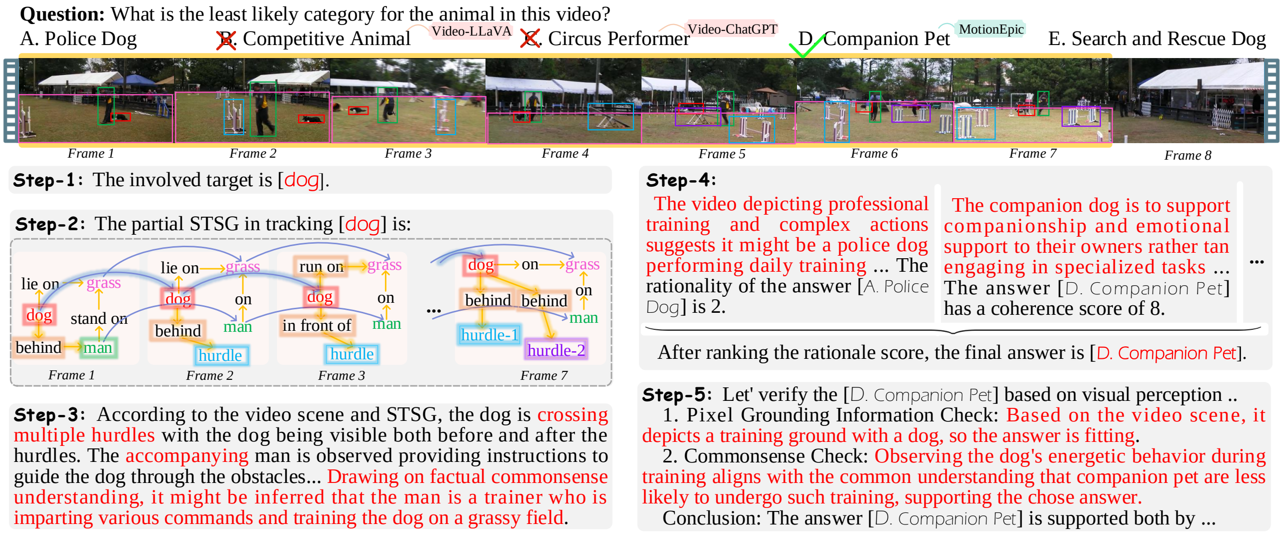


Figure 6: Performance drop (zero-shot) of MotionEpic after ablating different grounding-aware tuning item.

### 5.4. Analyses on MotionEpic Video MLLM

**Probing Video Grounding Ability.** To evaluate how well our MotionEpic is capable of video grounding, we here perform the probing test. Specifically, we evaluate the performance of MotionEpic on STSG parsing on the Action Genome test set, by comparing with SoTA DSG parsers: GPS-Net (Lin et al., 2020), STTran (Cong et al., 2021) and AP-Net (Li et al., 2022b). We measure three aspects: 1) the object grounding (bbox detection), 2) SG triplet classification (categories of entities, and relation predicates between entities), and 3) temporal action grounding (the start and end times of actions). Fig. 5 illustrates the results, where we see that MotionEpic achieves very competitive performance on par with SoTA parser, even with human-level performance. This reveals that MotionEpic shows reliable capability in providing video grounding information to support the subsequent in-depth video reasoning.

**Influence of Various Grounding-aware Tuning Strategies.** We further study the impacts/contributions of different grounding-aware tuning objectives introduced in §3.3. We design five groups of ablations where each tuning goes without one item, and the resulting model performs zero-shot end task. The results are shown in Fig. 6, where different items come with varied impacts, indicating the importance of video-STSG grounding fine-tuning. Notably, the lack of $\mathcal{L}_2$ and $\mathcal{L}_4$ result in the greatest degradation. This is intuitive, as these two objectives directly associate with the subsequent reasoning process, i.e., understanding STSG from video, and generating (partial) STSG given objects.



**Question:** What is the least likely category for the animal in this video?

A. Police Dog B. Competitive Animal C. Circus Performer D. Companion Pet E. Search and Rescue Dog

**Step-1:** The involved target is [dog].

**Step-2:** The partial STSG in tracking [dog] is:

**Step-3:** According to the video scene and STSG, the dog is crossing multiple hurdles with the dog being visible both before and after the hurdles. The accompanying man is observed providing instructions to guide the dog through the obstacles... Drawing on factual commonsense understanding, it might be inferred that the man is a trainer who is imparting various commands and training the dog on a grassy field.

**Step-4:**

The video depicting professional training and complex actions suggests it might be a police dog performing daily training ... The rationality of the answer [A. Police Dog] is 2.

The companion dog is to support companionship and emotional support to their owners rather tan engaging in specialized tasks ... The answer [D. Companion Pet] has a coherence score of 8.

...

After ranking the rationale score, the final answer is [D. Companion Pet].

**Step-5:** Let' verify the [D. Companion Pet] based on visual perception ..

1. Pixel Grounding Information Check: Based on the video scene, it depicts a training ground with a dog, so the answer is fitting.
2. Commonsense Check: Observing the dog's energetic behavior during training aligns with the common understanding that companion pet are less likely to undergo such training, supporting the chose answer.

Conclusion: The answer [D. Companion Pet] is supported both by ...

Figure 7: Visualization of qualitative example showcasing how our VoT framework achieves successful video reasoning.

| Data | CoT | | VoT | Human |
|---|---|---|---|---|
| | Video-LLaVA | MotionEpic | MotionEpic | |
| Causal-VidQA | 32.4 | 56.8 | 74.3 | 80.6 |
| Social-IQ | 22.3 | 40.1 | 61.4 | 72.7 |

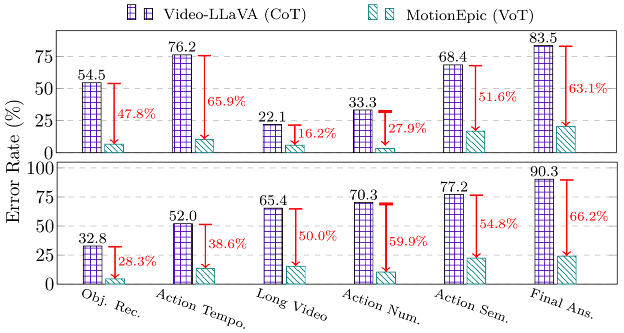


Figure 8: Above: human evaluation of video QA. Below: error rate under various specific categories.

### 5.5. Analyses on VoT Video Reasoning Framework

**Reasoning Ability Breaking-down.** Previously, we validated the overall stronger performance of the VoT reasoning system through extensive experimentation. Here, we aim to provide a more in-depth analysis of VoT. First, we select 200 hard instances each from the Causal-VidQA and Social-IQ test sets, and then compare the performance of Video-LLaVA and MotionEpic under CoT and VoT frameworks, respectively. Also we conduct human evaluation on this subset to gauge its difficulty level. The results, in the above table of Fig. 8, show that MotionEpic with VoT reasoning framework achieves quite exceptional results, comparable even to human performance. We further summarize the error cases and analyze differences in the 6 most frequent categories of errors. As seen in the below figure, MotionEpic (with VoT) significantly reduces the error rate of Video-LLaVA (with CoT), especially in terms of action semantics and commonsense understanding.

**Video Reasoning Visualization.** Finally, we present a case study to intuitively understand the superiority of our system. As shown in Fig. 7, the video displays a complex scene, and the given question is abstract and complex, not directly answerable through mere perception of the video itself. However, our MotionEpic provides the correct answer, while the other two baselines err. At the content perception level, VoT ensures accurate and robust understanding through STSG-based video grounding, preventing hallucination, i.e., it correctly interprets that the animal is a dog, then infers from commonsense that the scene involves a trainer training the dog. Then, at the cognitive level, it analyzes each option to determine the best answer. Through further verification, the result aligns with both the video content and factual commonsense understanding. Overall, the entire reasoning greatly improves the accuracy at each step through problem decomposition, while ensuring an explainable process decision rationale.

## 6. Conclusion

In this work, we present an advanced solution for complex video reasoning, including a novel video MLLM, MotionEpic, and an innovative VoT framework. MotionEpic achieves fine-grained pixel-level spatial-temporal video grounding by adeptly integrating video STSG representation. With MotionEpic, the VoT framework resolves the intricate video task by skillfully dissecting it into manageable subproblems, tackling them sequentially from low-level pixel perception to advanced cognitive interpretation. Our experiments across various complex video QA benchmarks have not only proven the efficacy of our approach but have also boosted the existing state-of-the-art standards. Overall, this work marks a substantial contribution to the video modeling community, paving the way for more nuanced, human-level analysis in video reasoning research.

## Statement of Potential Broader Impact

This paper aims to construct a robust, human-level video understanding and reasoning framework. The system must be built upon existing LLM to realize its full potential. Potential implications include substantial energy consumption during LLM system training, leading to environmental degradation, and the necessity for more extensive data corpora for training. Moreover, due to the powerful video reasoning and comprehension capabilities, there exists the potential for malicious actors to exploit this framework for nefarious intents, posing a societal threat. Consequently, the release of this framework necessitates the establishment of specific licensing mechanisms to ensure responsible deployment and mitigate potential misuse.

## References

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *ICML*, volume 2, pp. 4, 2021.

Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., and Shi, W. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4778–4787, 2017.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 902023.

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., and Yang, M. Y. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16372–16382, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Dwivedi, V. P. and Bresson, X. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the CVPR*, pp. 961–970, 2015.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *Proceedings of the ICLR*, 2022.

Ji, J., Krishna, R., Fei-Fei, L., and Niebles, J. C. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10236–10247, 2020.

Ko, D., Lee, J., Kang, W.-Y., Roh, B., and Kim, H. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4300–4316, 2023.

Lei, J., Yu, L., Bansal, M., and Berg, T. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379, 2018.

Lei, J., Yu, L., Berg, T., and Bansal, M. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8769–8784, 2020.

Li, J., Niu, L., and Zhang, L. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21273–21282, 2022a.

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ICML*, pp. 19730–19742, 2023a.

Li, J., Wei, P., Han, W., and Fan, L. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11963–11974, 2023b.

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023c.

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023d.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023e.

Li, Y., Yang, X., and Xu, C. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13874–13883, 2022b.

Li, Y., Xiao, J., Feng, C., Wang, X., and Chua, T.-S. Discovering spatio-temporal rationales for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13869–13878, 2023f.

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

Lin, J., Gan, C., and Han, S. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7083–7093, 2019.

Lin, X., Ding, C., Zeng, J., and Tao, D. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3746–3753, 2020.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023.

Maaz, M., Rasheed, H. A., Khan, S. H., and Khan, F. S. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424, 2023.

Neimark, D., Bar, O., Zohar, M., and Asselmann, D. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3163–3172, 2021.

OpenAI. Introducing chatgpt. 2022a.

OpenAI. Gpt-4 technical report. 2022b.

Schuster, M. and Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

Sun, Y., Xue, H., Song, R., Liu, B., Yang, H., and Fu, J. Long-form video-language pre-training with multimodal temporal contrastive learning. In *Proceedings of the NeurIPS*, 2022.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022a.

Wang, X., Liang, J., Wang, C.-K., Deng, K., Lou, Y., Lin, M., and Yang, S. Vlap: Efficient video-language alignment via frame prompting and distilling for video question answering. *arXiv preprint arXiv:2312.08367*, 2023.

Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022b.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

Wu, B., Yu, S., Chen, Z., Tenenbaum, J. B., and Gan, C. Star: A benchmark for situated reasoning in real-world videos. In *Annual Conference on Neural Information Processing Systems*, 2021.

Xiao, J., Shang, X., Yao, A., and Chua, T. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the CVPR*, pp. 9777–9786, 2021.

Xu, J., Mei, T., Yao, T., and Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the CVPR*, pp. 5288–5296, 2016.

Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., and Guo, B. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the CVPR*, pp. 5026–5035, 2022.

Ye, Q., Xu, G., Yan, M., Xu, H., Qian, Q., Zhang, J., and Huang, F. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15405–15416, 2023.

Yu, S., Cho, J., Yadav, P., and Bansal, M. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., and Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 558–567, 2021.

Zadeh, A., Chan, M., Liang, P. P., Tong, E., and Morency, L.-P. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019.

Zhang, H., Li, X., and Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *CoRR*, abs/2306.02858, 2023a.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023b.

Zhao, Y., Fei, H., Cao, Y., Li, B., Zhang, M., Wei, J., Zhang, M., and Chua, T.-S. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5281–5291, 2023.

Zolfaghari, M., Singh, K., and Brox, T. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 695–712, 2018.

# A. Extended Method Specifications

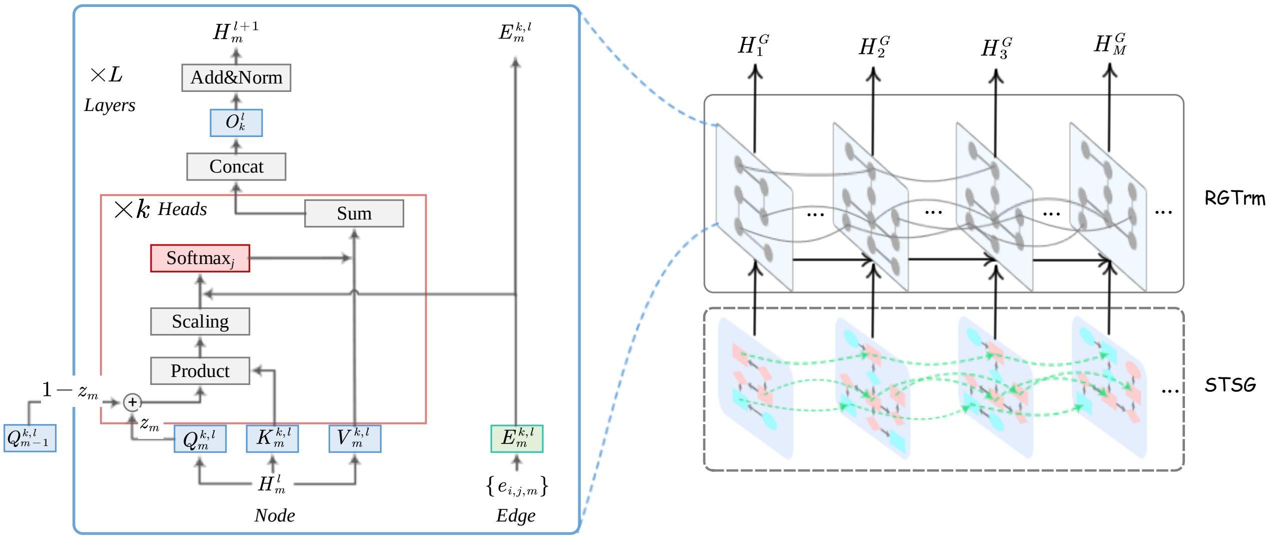


Figure 9: Illustration of the Recurrent Graph Transformer (RGTrm) for STSG encoding.

## A.1. STSG Encoder

STSG is characterized by temporal and spatial dynamics, compared with the single SG. Thus, we introduce a recurrent graph Transformer (RGTrm) to model STSG, for which we draw the main inspiration from the recurrent networks (Schuster & Paliwal, 1997). As demonstrated in Fig. 9, RGTrm is built upon the GTrm propagation Technically, GTrm has $L$ stacked layers, where the update of representation $h_i^l$ of node $i$ in $\kappa$-th head in $l$-th layer is written as:

$$h_i^{l+1} = O_\kappa^l \, ||_{\kappa=1}^{H} \, (\sum_{j \in \mathcal{N}_i} w_{i,j}^{\kappa,l} \, V^{\kappa,l}) \,, \quad \text{where} \quad w_{i,j}^{\kappa,l} = \text{Softmax}_j (\frac{Q^{\kappa,l} \cdot K^{\kappa,l}}{\sqrt{d_\kappa}}) \cdot E^{\kappa,l} \,, \tag{1}$$

where $\kappa$ denotes the attention head number. $O_\kappa^l \in \mathbb{R}^{d \times d}$ is the attention head representation. $K^{\kappa,l}$=$\boldsymbol{W}^K\{h_j^l\}$, $Q^{\kappa,l}$=$\boldsymbol{W}^Q\{h_j^l\}$, $V^{k,l}$=$\boldsymbol{W}^V\{h_j^l\}$ (all $\in \mathbb{R}^{d \times d}$) are the key, query and value representation derived from the last layer representation. $E^{\kappa,l} \in \mathbb{R}^{d \times d}$=$W_E\{e_{i,j}\}$ is the embedding of edge $e_{i,j}$ in an SG. Furthermore, the dynamics are additionally modeled in RGTrm through the *temporal coreference edges* of nodes within STSG (i.e., $e_{k-1 \rightarrow k}$). Specifically, we calculate each attention weight at $k$-th time-frame in STSG:

$$w_{i,j}^{\kappa,l,k} = \text{Softmax}_j (\frac{\hat{Q}^{\kappa,l,k} \cdot K^{\kappa,l,k}}{\sqrt{d_\kappa}}) \cdot E^{\kappa,l,k} \,, \tag{2}$$

$$\text{where} \quad \hat{Q}^{\kappa,l,k} = (1 - z_k) \cdot Q^{\kappa,l,k-1} + z_k \cdot Q^{\kappa,l,k} \,, \quad z_t = \sigma(\boldsymbol{W}^z \cdot Q^{\kappa,l,k} \cdot K^{\kappa,l,k}) \,. \tag{3}$$

where the initial node representation $h_{k,i}^0$ is the concatenation of 1) the proposal's neural representation $f_{k,i}$, and the 2) the embedding of the node category label $c_{k,i}$. The region of the neural representation $f_{k,i}$ of object is interpreted via the bbox annotation $b_{k,i}$. $E^{\kappa,l,k} \in \mathbb{R}^{d \times d}$=$W_E\{e_{k,i,j}\}$ is the embedding of edge $e_{k,i,j}$ of time-step $k$ in STSG.

## A.2. Detailed Prompt Construction and System I/O

Here, we provide detailed prompts as well as their inputs and outputs, for each step of the VoT reasoning framework.

► **Step-1:** If the raw question is a multi-choice question, the prompt for Step-1 should be:

Step-1: Task Definition and Target Identification for Multi-choice Question

► **Input:**
<Task Definition>
Now you are an expert in analyzing video data, and you should answer a question based on the given video.
For the question, several candidate answers are provided, where you need to choose [the most suitable option — all possible correct option(s)].
</Task Definition>

<Input Video> </Input Video>

<Question>
Given the question: [`What is the relationship between the white truck and this neighborhood? A. Transportation B. Buildings C. Clean Services D. Entertainment Facilities`], what are the possible targets of the [`Video`] mainly mentioned or involved?
</Question>

► **Output:**
The involved targets are [`the white truck`], [`the neighborhood`]

Otherwise for the open-ended format, the prompt will be:

Step-1: Task Definition and Target Identification for Open-ended Question

► **Input:**

<Task Definition>
Now you are an expert in analyzing video data, and you should answer a question based on the given video.
For the question, you should answer in an open-ended format.
</Task Definition>

<Video> </Video>

<Question>
Given the question: [`What is the relationship between the white truck and this neighborhood?`] what are the possible targets of the [`Video`] mainly mentioned or involved?
</Question>

► **Output:**
The involved targets are [`the white truck`], [`the neighborhood`].

► **Step-2:** The detailed prompt for Step-2 is shown as follows:

Step-2: Object Tracking

► **Input:**

<Question>
Provide the tracklet of involved [`the neighborhood`] and [`the white truck`] by outputting the corresponding partial expression in the [`STSG`].
</Question>

► **Output:**
The partial STSG in tracking [`the neighborhood`] and [`the white truck`] is [Frame 1: Objects: [”car-1”: [0.0,13.4,7.0,8.1], ...], Triplets: [(”car-1”,”on the left”, ”street”), ...]...].

► **Step-3:** The detailed prompt for Step-3 is shown as follows:

Step-3: Action Analyzing

► **Input:**

<Question>
Combining all possible related commonsense, analyze the motion behavior based on the [`the white truck`] and [`the neighborhood`] and the neighbor scenes within [`STSG`]. Describing the action observations and implications.
</Question>

```
<STSG>
Frame 1: {
    Objects: [{”car-1”: [0.0,13.4,7.0,8.1]}, ...],
    Triplets: [(”car-1”, ”on the left”, ”street”), ...]
}
...
</STSG>
```

► **Output:**
The two men are driving the white truck into a neighborhood, and pour the garbage from the roadside trash cans into the white truck. According to commonsense, the white car is used for collecting rubbish...

► **Step-4:** When the raw question is an open-ended QA, we consider prompting the model to output multiple distinct optional answers, such that we unify all QA problems into the Multi-choice type:

Step-4-Pre: Transforming Open-ended Question Answering into Multi-choice one

► **Input:**

<Question>
For the question [`What is the relationship between the white truck and this neighborhood?`], given a candidate answer [`A. Transportation`], please based on the action's [`The two men are driving the white truck into a neighborhood...`] combined with commonsense, output 4 distinct optional answers with the rationality score of this answer with a 1-10 scale.
</Question>

► **Output:**

Answer A: While the white truck is indeed moving through ... but rather the collection of garbage ...
Answer B: ...

Given the multiple-choice question, we first prompt the model to score its likelihood (from 1 to 10) in conjunction with commonsense knowledge, and provide a corresponding rationale for each candidate answer. Then, we consider a ranking mechanism to determine the final answer.

### Step-4-A: Multi-choice Question Answering via Ranking

► **Input:**

<Question for Answer A>
For the question [What is the relationship between the white truck and this neighborhood? A. Transportation B. Buildings C. Clean Services D. Entertainment Facilities], given a candidate answer [A. Transportation], please based on the action's [The two men are driving the white truck into a neighborhood...] combined with commonsense, score the rationality of this answer with a 1-10 scale, and also output the rationale.
</Question for Answer A>

<Question for Answer B>
For the question [What is the relationship between the white truck and this neighborhood? A. Transportation B. Buildings C. Clean Services D. Entertainment Facilities], given a candidate answer [B. Buildings], please based on the action's [The two men are driving the white truck into a neighborhood...] combined with commonsense, score the rationality of this answer with a 1-10 scale, and also output the rationale.
</Question for Answer B>

<Question for Answer C>
For the question [What is the relationship between the white truck and this neighborhood? A. Transportation B. Buildings C. Clean Services D. Entertainment Facilities], given a candidate answer [C. Clean Services], please based on the action's [The two men are driving the white truck into a neighborhood...] combined with commonsense, score the rationality of this answer with a 1-10 scale, and also output the rationale.
</Question for Answer C>

<Question for Answer D>
For the question [What is the relationship between the white truck and this neighborhood? A. Transportation B. Buildings C. Clean Services D. Entertainment Facilities], given a candidate answer [D. Entertainment Facilities], please based on the action's [The two men are driving the white truck into a neighborhood...] combined with commonsense, score the rationality of this answer with a 1-10 scale, and also output the rationale.
</Question for Answer D>

► **Output:**

Based on the observations and commonsense, the rationality of the answer [A. Transportation] is 5. While the white truck is indeed moving through ... but rather the collection of garbage ...

Step-4-B: Multi-choice Question Answering via Ranking

► **Input:**

<Question for Ranking>
Now, we know the rationale score of the answer [`A. Transportation`] is 5, ... Please rank the candidate the answer based on the rationale score of each candidate's answer.
</Question for Ranking>

► **Output:**

Based on the rationality, the ranking results in descending order is [`C. Clean Service`] [`A. Transportation`] [`D. Entertainment Facilities`] [`B. Buildings`]. So, the final answer is [`C. Clean Service`].

► **Step-5:** The detailed prompt for Step-5 is shown as follows:

Step-5: Answer Verification

► **Input:**

<Video> </Video>

<Question>
Given the [`Video`], and the raw question [`What is the relationship between the white truck and this neighborhood? A. Transportation B. Buildings C. Clean Services D. Entertainment Facilities`], now you need to verify the previous answer by

1) checking the pixel grounding information if the answer [`C. Clean Service`] aligns with the facts presented in the video from a perception standpoint;

2) determining from a cognition perspective if the commonsense implications inherent in the answer [`C. Clean Service`] contradict any of the main [`Observations`] inferred in the 3-$rd$ reasoning step.

Output the verification result with rationale.
</Question>

► **Output:**

Certainly. Let's verify the answer [`C. Clean Service`]:

1. Grounding information: the collection action of garbage, which aligns ...
2. Commonsense verification: driving a white truck into a neighborhood is associated with waste management ...

Overall, the answer [`C. Clean Service`] is correct.

Note that the above demonstration of the entire prompting process is just one example. In practice, we simply need to replace the example instance with any other QA instance. It is also worth noting that the VoT system is not limited to video QA; any other video tasks and scenarios requiring reasoning can be adapted to this framework by transforming the original task into a QA format.

## B. Extended Experiment Configurations

### B.1. Extended Implementation

MotionEpic uses the Vicuna-7B (v1.5)[2] as the backbone LLM. We adopt the ViT-L/14[3] as the video encoder, and use the Q-Former[4] as the projector. All the modules take the default configurations without much modification. For our Recurrent Graph Transformer, we take a 6-layer architecture with 768-d hidden sizes. The text tokenizer is sourced from LLaMA, with approximately 32,000 classes. For each video, we uniformly sample certain frames with a sampling rate of 8 fps for

[2] https://huggingface.co/lmsys/vicuna-7b-v1.5

[3] https://huggingface.co/sentence-transformers/clip-ViT-L-14

[4] https://huggingface.co/spaces/Salesforce/BLIP2

Table 5: Zero-shot Video QA results comparing with GPT-4V. In the brackets are the video frame sampling rates.

| **Model** | **ActivityNet** | **NExT-QA** | **STAR** | ***AVG.*** |
|---|---|---|---|---|
| **• CoT** | | | | |
| GPT-4V (8 fps) | 29.7 | 35.1 | 32.5 | 32.4 |
| GPT-4V (32 fps) | 40.6 | 51.6 | 46.4 | 46.2 |
| Video-LLaVA | 48.9 | 60.5 | 54.0 | 54.5 |
| **• VoT** | | | | |
| MotionEpic | **59.6** | **75.5** | **68.4** | **67.8** |

Table 6: Zero-shot Video QA results of Video-LLaVA under VoT.

| **Model** | **ActivityNet** | **NExT-QA** | **STAR** | ***AVG.*** |
|---|---|---|---|---|
| **• CoT** | | | | |
| Video-LLaVA | 48.9 | 60.5 | 54.0 | 54.5 |
| **• VoT** | | | | |
| Video-LLaVA | 50.1 | 63.8 | 57.6 | 57.2 |
| MotionEpic | **59.6** | **75.5** | **68.4** | **67.8** |

fine-grained reasoning. We note that too large sampling rate introduces noises (i.e., redundant frames) and huge computation cost, while too small one will cause important information loss. Here we use the 8 fps, as in our preliminary study we verified that it helps achieve the best trade-off. For the fine-tuning setting of end tasks, we will tune the MotionEpic based on the training set using the setting as prior baselines, i.e., data split and evaluation methods. For the zero-shot setting, we will directly perform video QA without using the in-domain training set.

### B.2. Video-Language Modeling

For the training of grounding-level tuning, we take a specific order from coarse-grained one to fine-grained one. That is, we first train the system by optimizing $\mathcal{L}_1$+$\mathcal{L}_2$. When the system learns to converge, we then perform the fine-grained grounding, i.e., by optimizing $\mathcal{L}_3$+$\mathcal{L}_4$+$\mathcal{L}_5$. We train the system for these objectives on 16 A100 GPUs for around 50 hours.

Also besides our main grounding-level tuning, we conducted two important types of pre-training and tuning. First, we performed conventional video pre-training on Webvid, which serves as the important warming starting for the following video understanding tuning, such as the masked language modeling (Sun et al., 2022), and the overall video-text alignment learning (Xue et al., 2022). This enables the overall MLLM to have a fundamental understanding between video and language.

Despite aligning the encoding modules with LLM, there remains a gap towards the goal of enabling the overall system to faithfully follow and understand users' instructions and generate the desired outputs. To address this, further instruction tuning (IT) is necessary. IT involves additional training of overall MM-LLMs using '*(INPUT, OUTPUT)*' pairs, where '*INPUT*' represents the user's instruction, and '*OUTPUT*' signifies the desired model output that conforms to the given instruction. After the grounding-level tuning, we utilized existing video instruction tuning data for instruction tuning of the model, which includes the dataset from VideoChat (Li et al., 2023c) and Video-ChatGPT (Maaz et al., 2023)

## C. Extended Experiments

### C.1. Comparison with GPT-4V on Video Reasoning

Table 5 presents the comparisons with OpenAI GPT-4V (OpenAI, 2022b) on zero-shot video QA. Since GPT-4V doesn't support direct video input, we thus extract the video frames as image inputs. We extract the frames with two different sampling rates: 8 fps and 32 fps, where larger rate leads to less information loss but meanwhile more redundant noises. We see that GPT-4V shows quite lower performance than the Video-LLaVA. This indicates that relying solely on video frame analysis without supporting the entire video input and understanding significantly limits performance. This is because analyzing frames in isolation can actually compromise the original video's temporal integrity. Particularly for data that requires understanding at the cognition level, i.e., NExT-QA and STAR, if the video itself is not well understood, any subsequent cognitive interpretation is even less feasible.

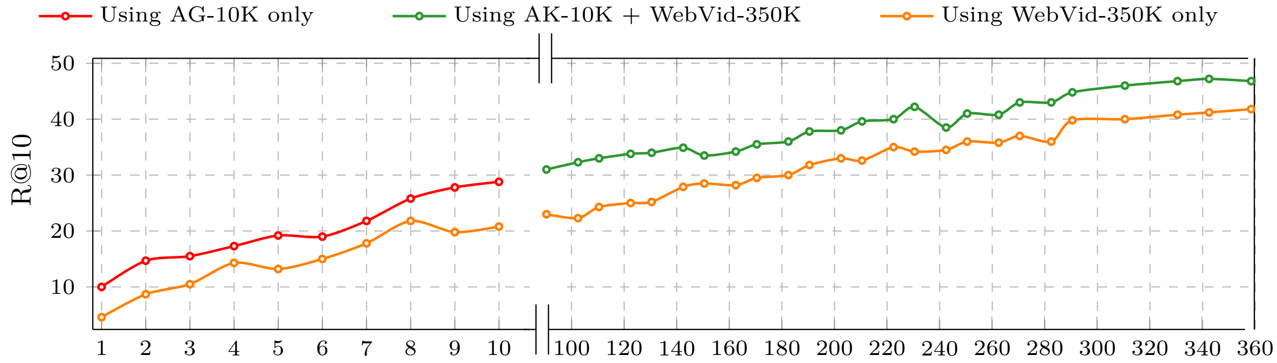


Figure 10: STSG parsing performance using different amounts of data (k).

### C.2. Equipping Video-LLaVA with VoT

Our VoT framework requires pixel grounding ability of the backbone video MLLM, e.g., MotionEpic. For Video-LLaVA, although it doesn't support temporal grounding with STSG interpretation in the 2-$nd$ and 3-$rd$ reasoning step in VoT framework, we consider equipping it with a simplified version of VoT where the STSG-based temporal grounding part is removed. In other words, we also practice the main 6 reasoning steps in VoT, but we don't ask Video-LLaVA for outputting the STSG expressions. As alternative, we ask Video-LLaVA to give the bbox of target objects and the time steps of the mentioned frames in the 2-$nd$ step. The rest parts of VoT are kept unchanged. Table 6 shows the performance of Video-LLaVA + VoT. We see that there are clear improvements for Video-LLaVA.

### C.3. Study of STSG Impact

We here carry out evaluation on the impact of the STSG data used in our system, in terms of the STSG data amount and STSG parsing quality. To tune our MotionEpic, we use the video-STSG pairs. We leverage the Action Genome (AG) data (Ji et al., 2020), which contains 10K high-quality manual annotated STSGs of videos. We also use part of WebVid-10M videos (Bain et al., 2021), where we select 350K videos with clear and rich actions, and parse the video STSGs via SoTA parser, AP-Net (Li et al., 2022b). Now we study the performances of our system on scene graph parsing (i.e., grounding ability) if we use 1) AG 10K only, 2) AG 10K + WebVid 350K, and 3) WebVid 350K only. Fig. 10 shows the trends. From 'AG 10K only' vs. 'WebVid 350K only' at 10K data, we see that with the high-quality STSG data, i.e., AG 10K, the model tends to converge at higher performance. From 'AG 10K + WebVid 350K' vs. 'WebVid 350K only', we learn that with AG for initiating the tuning, using more datasets even from automatic parsing, the performance can be further boosted.

### C.4. Discussion of STSG Parsing with MotionEpic

In the previous experiment, i.e., Fig. 5, we conducted an automatic evaluation of the STSG grounding/parsing capability. However, this mode of automatic evaluation significantly underestimates the capabilities of MotionEpic: as an LLM, it boasts a more powerful capacity in terms of open vocabulary. Consequently, we consider conducting a human evaluation on STSG Parsing. Similarly, we focus solely on scene graph triplet classification, namely measuring the categories of objects and relation predicates between entities. We randomly select 200 pieces of data from the AG test set for testing. Table 7 presents the performance of manual evaluation in terms of objects and relations, as well as overall performance. It is evident that our MotionEpic model outperforms the other two state-of-the-art models, exhibiting a lower error rate across the board. Notably, MotionEpic significantly reduces the error rate for both wrong objects and wrong relations to the minimum. This advantage stems entirely from the open vocabulary benefits of the LLM. It is clear that MotionEpic's vocabulary for objects and relations is more extensive.

### C.5. Impact of Video Length

Here, we continue to analyze the performance of different video reasoning models across various video lengths. Fig. 11 displays the comparative results of the three methods. It is evident that our MotionEpic system exhibits the strongest performance under the VoT framework. Most notably, even in the case of excessively long videos, such as those in the top

Table 7: Human evaluation on STSG parsing.

| Model | Object | | | Relation | | | Overall Error Rate |
|---|---|---|---|---|---|---|---|
| | Missing Object | Wrong Object | Object Vocab | Missing Relation | Wrong Relation | Relation Vocab | |
| STTran | 20.3 | 35.0 | 523 | 33.7 | 45.1 | 203 | 38.2 |
| AP-Net | **18.6** | 29.6 | 640 | 28.3 | 40.4 | 245 | 30.8 |
| MotionEpic | 23.5 | **8.4** | **2,882** | **25.4** | **13.5** | **863** | **19.5** |

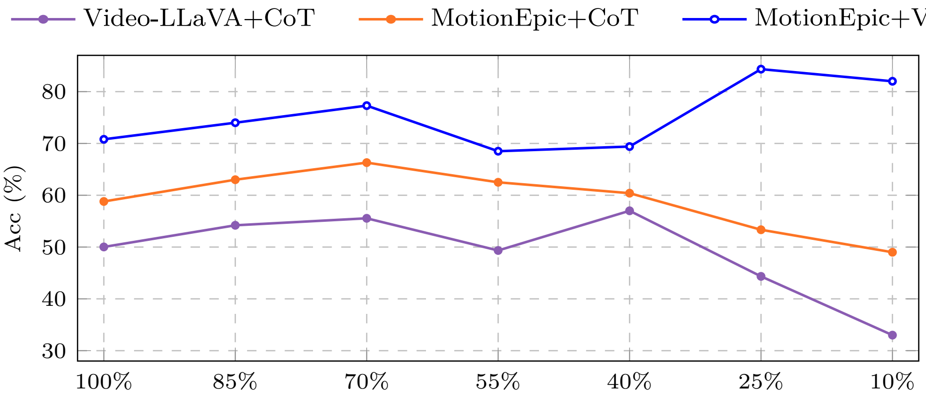


Figure 11: Accuracy (zero-shot on NExT-QA) by video length. X-xis is the top percentage by video length (from longer to shorter).

Table 8: Zero-shot Video QA results on NExT-QA by object number.

| Model | Object Number ≥5 | Object Number <5 |
|---|---|---|
| • **CoT** | | |
| Video-LLaVA | 44.0 | 75.9 |
| MotionEpic | 44.3 | 79.7 |
| • **VoT** | | |
| MotionEpic | **70.4** | **84.0** |

10% of length, it manages to achieve superior performance in counteracting the challenges posed by video length.

### C.6. Performance of Video with Different Number of Objects

We evaluate the reasoning performance of the video by varying the number of objects in the video, as shown in Table 8. MotionEpic demonstrates a substantial performance advantage, specifically when the object count is less than 5, with accuracy scores of 84.0 compared to 75.9/79.7 of baselines. Even when the number of objects exceeds 5 and the same inference mechanism (i.e., COT) is employed, our MotionEpic consistently outperforms the leading baseline, Video-LLaVA. Notably, this performance enhancement is further greatly accentuated when incorporating the VoT mechanism, i.e., increasing to 70.4.

### C.7. More Visualizations for Qualitative Results

Finally, we provide two sets of cases for qualitative analyses. We observe that different Video QA datasets exhibit varying biases. Some datasets lean more towards content recognition, relying heavily on perceptual abilities without necessitating much cognitive understanding; others are more inclined towards cognitive-level comprehension, such as physical, cultural or humanities knowledge, where the video content itself is relatively straightforward. We consider cases from both these perspectives.

Fig. 12 presents two sets of QA cases at the video perception level. For the first question, which requires counting the number of people in the video, it is observed that both baselines provided incorrect answers. However, thanks to our

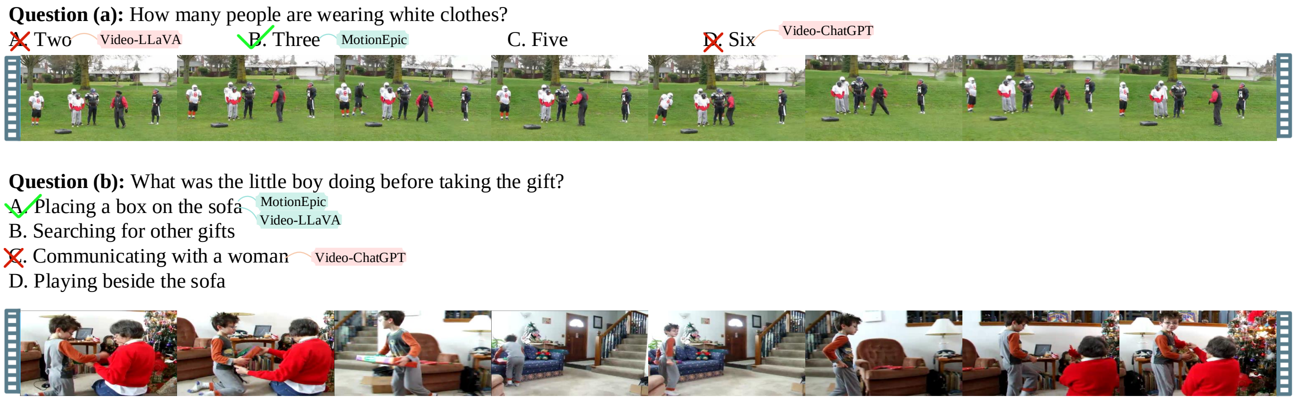

Question (a): How many people are wearing white clothes?
A. Two
Video-LLaVA
B. Three
MotionEpic
C. Five
D. Six
Video-ChatGPT
Question (b): What was the little boy doing before taking the gift?
A. Placing a box on the sofa
MotionEpic
Video-LLaVA
B. Searching for other gifts
C. Communicating with a woman
Video-ChatGPT
D. Playing beside the sofa

Figure 12: Qualitative examples of perception-level reasoning. The correct answer is marked with a green checkmark, and the wrong answer is marked with a red cross.

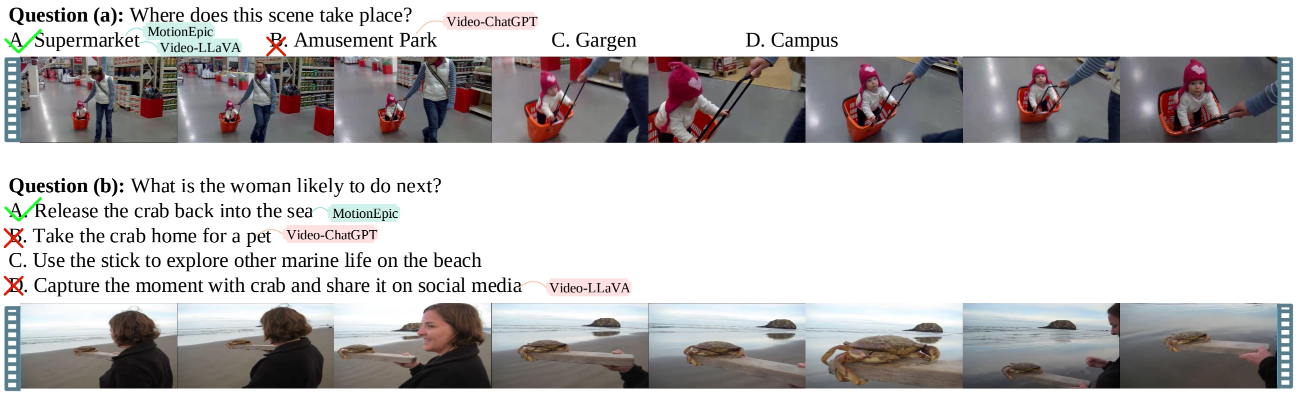

Question (a): Where does this scene take place?
A. Supermarket
MotionEpic
Video-LLaVA
B. Amusement Park
Video-ChatGPT
C. Gargen
D. Campus
Question (b): What is the woman likely to do next?
A. Release the crab back into the sea
MotionEpic
B. Take the crab home for a pet
Video-ChatGPT
C. Use the stick to explore other marine life on the beach
D. Capture the moment with crab and share it on social media
Video-LLaVA

Figure 13: Qualitative examples of cognitive-level reasoning.

MotionEpic's utilization of the STSG structured representation, it can accurately ground the number of objects, thereby providing the correct result. In the second case, a straightforward understanding of the temporal information in the video suffices to answer the question. It is shown that both MotionEpic and Video-LLaVA answered correctly.

Fig. 13 showcases two cases at the cognitive level. For the first case, the question "Where does this scene take place?" can be answered by understanding the scene's content and combining it with common sense to conclude: a supermarket. For the second case, merely observing the video content "a woman holds a crab with a stick" makes it challenging to grasp the implicit intention. However, integrating some cultural commonsense, it can be understood that the girl is releasing the crab back into the sea.