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

Hao Fei [1] **Shengqiong Wu** [1] **Wei Ji** [1] **Hanwang Zhang** [2] **Meishan Zhang** [3] **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 sub-problems, 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. The project is open at `https://haofei.vip/VoT`.

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.

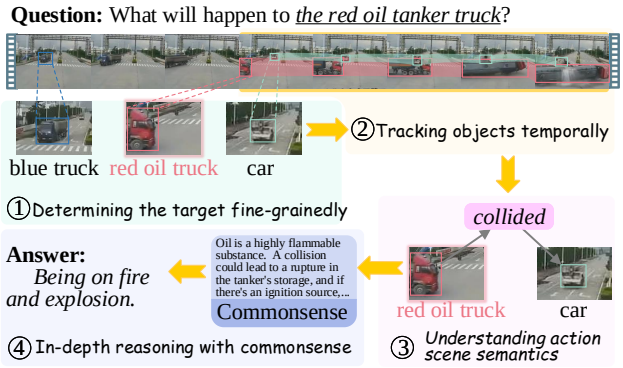

Figure 1: Human-like video reasoning intuitively follows a multi-step procedure, from lower-level perceptive fine-grained pixel grounding and tracking, to higher-level cognitive action scene semantics understanding.

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.

[1]National University of Singapore, Singapore [2]Nanyang Technological University, Singapore [3]Harbin Institute of Technology (Shenzhen), China. Correspondence to: Meishan Zhang <zhangmeishan@hit.edu.cn>.

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

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 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

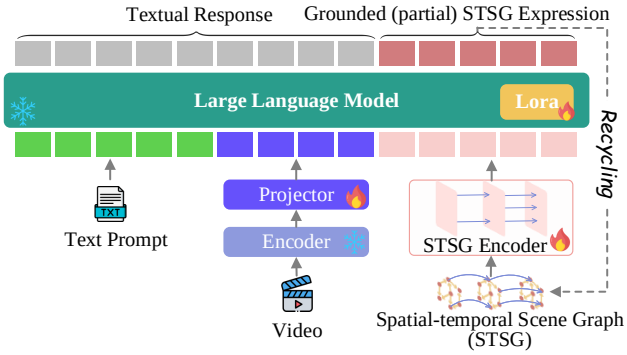

Figure 2: Overview of the MotionEpic video MLLM.

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.

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 very clear margins, 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 datasets (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; Wu et al., 2023c), 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. To enable MLLMs with spatiotemporal modeling, we consider employing the dynamic video scene graph representation (Ji et al., 2020). SGs (Johnson et al., 2018) are characterized by highly structured graph representations (Fei et al., 2022), which intrinsically depict the underlying semantic implications of the data, and thus have been extensively integrated into a wide range of downstream cross-modal tasks (Zhao et al., 2023b; Wu et al., 2024; Fei et al., 2023b; Wu et al., 2023b;a), especially in video modeling (Zhao et al., 2023a; Fei et al., 2023c; 2024).

Meanwhile, recent advancements in CoT technology have made significant strides in enhancing the reasoning capabilities of LLMs (Wei et al., 2022; Zhang et al., 2022; Fei et al., 2023a; Zheng et al., 2024). While there are efforts enhancing multimodal reasoning with multimodal CoT (Lu et al., 2022; Zhang et al., 2023c), we still note a lack of research specifically focused on integrating CoT into video 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.

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.

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., 2023a). 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

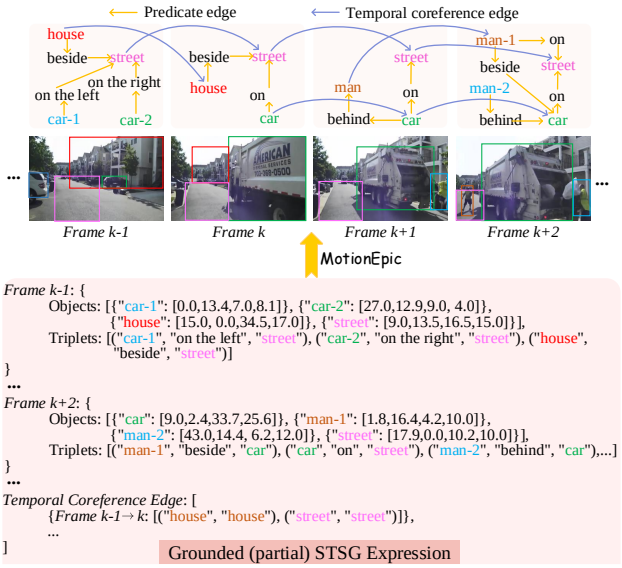

Figure 3: The STSG expression generated by MotionEpic, with its corresponding structural STSG illustration.

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, $e_{k-1\to k}$, 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 objectives, 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. 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.

Before conducting the above grounding-level tuning, we perform conventional video pre-training on Webvid, which serves as the important warming starting for the following video understanding tuning, 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 is necessary. 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)

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 subproblems. 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

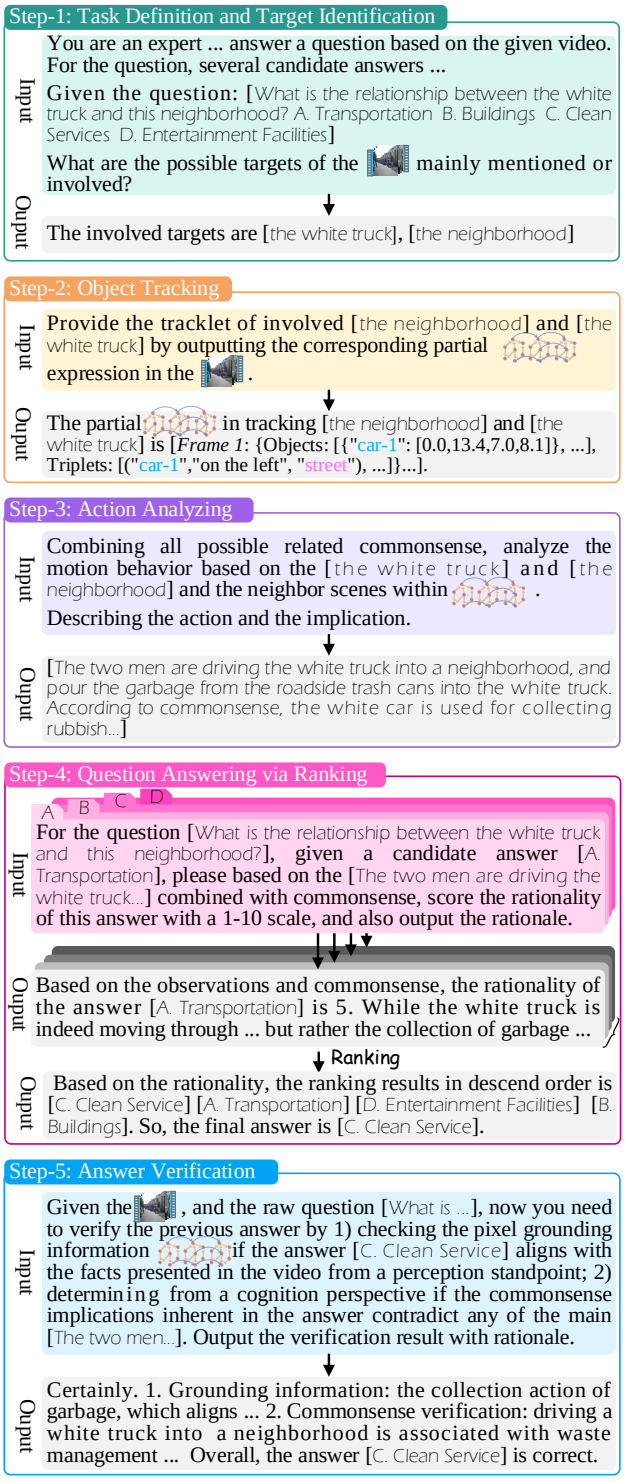

Figure 4: An illustrative view of VoT framework. The complete I/O and prompts are detailed in Appendix.

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.

> 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.

[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.

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 consider using MSR-VTT (Xu et al., 2016) and ActivityNet (Heilbron et al., 2015) datasets. All datasets come with their own splitting, and we follow the prior practice without modification.

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).

Implementations. 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 encoding STSGs, we take a 6-layer architecture with 768-d hidden sizes. The object neural representation f_i is obtained via CLIP, which will be used as node embedding initiation. 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 fine-grained reasoning. We note that too large sampling rate introduces noises (i.e., redundant frames) and huge computation costs, 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

[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 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**.

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	63.2	69.5	40.4
Video-LLaVA+STSG	67.0	65.9	68.9	58.7	53.7	64.9	70.4	41.7
MotionEpic	68.2	66.8	69.6	60.6	57.4	66.1	71.7	43.0
• **VoT**								
MotionEpic	**73.4**	**71.5**	**72.6**	**66.6**	**62.7**	**70.8**	**72.8**	**45.0**

training set. All trainings are conducted on 16 NVIDIA A100 GPUs.

Baselines and Evaluations. 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 a vanilla CoT setting. Further, we also implement the Video-LLaVA by integrating the STSG features. We adopt accuracy as the main metric of the QA task performance, following the prior practice.

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. 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 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.

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	72.9	73.9	65.2	63.1	45.9	66.0	62.7	45.8
Video-ChatGPT	73.1	75.1	66.0	63.9	46.0	67.8	63.6	50.0
Video-LLaVA	73.7	74.4	67.6	65.4	47.7	68.0	64.9	51.5
• **CoT**								
Video-LLaVA	74.2	74.8	68.0	65.7	48.1	70.3	65.7	52.9
Video-LLaVA+STSG	75.7	75.9	68.9	67.2	50.0	70.7	67.2	53.6
MotionEpic	78.5	77.2	70.1	70.8	52.4	71.2	69.1	55.0
• **VoT**								
MotionEpic	**81.2**	**83.0**	**74.3**	**73.7**	**54.7**	**74.5**	**73.8**	**58.6**

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	75.7
Video-LLaVA	66.3	67.7	63.8	75.9
• **CoT**				
Video-LLaVA	67.7	69.0	65.9	76.5
Video-LLaVA+STSG	68.0	71.6	67.6	78.9
MotionEpic	72.2	73.4	69.1	80.7
• **VoT**				
MotionEpic	**76.0**	**75.8**	**74.6**	**83.3**

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. 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

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	60.0	46.9	59.5	52.0	54.6
Video-LLaVA+STSG	61.5	48.4	60.6	52.7	55.8
MotionEpic	63.1	50.0	61.9	56.5	57.8
• **VoT**					
MotionEpic	**66.2**	**54.6**	**66.5**	**61.7**	**62.3**
w/o Verify-G	63.6	51.4	62.0	59.1	59.0
w/o Verify-C	65.1	53.4	62.8	58.8	60.1

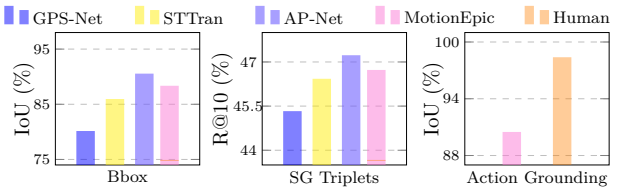

Figure 5: MotionEpic performance on object grounding, scene graph triplet classification, and action grounding.

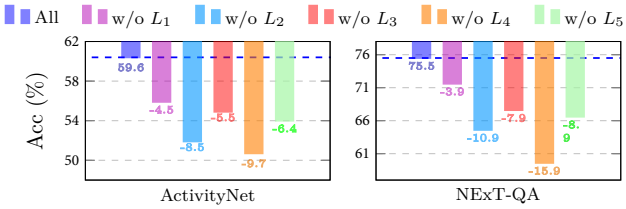

Figure 6: Performance (zero-shot) drop of MotionEpic after ablating different grounding-aware tuning items.

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

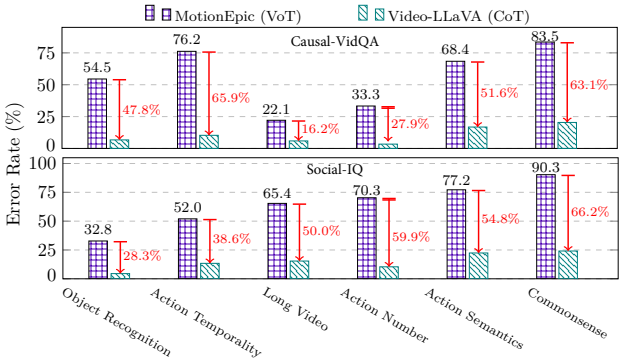

Figure 7: **Above Table**: human evaluation of video QA. **Below Figure**: error rate under various specific categories.

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.

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 on two datasets 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 are directly associated with the subsequent reasoning process, i.e., understanding STSG from video, and generating (partial) STSG given objects.

5.5. Analyses on VoT 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. 7, show that MotionEpic with VoT reasoning

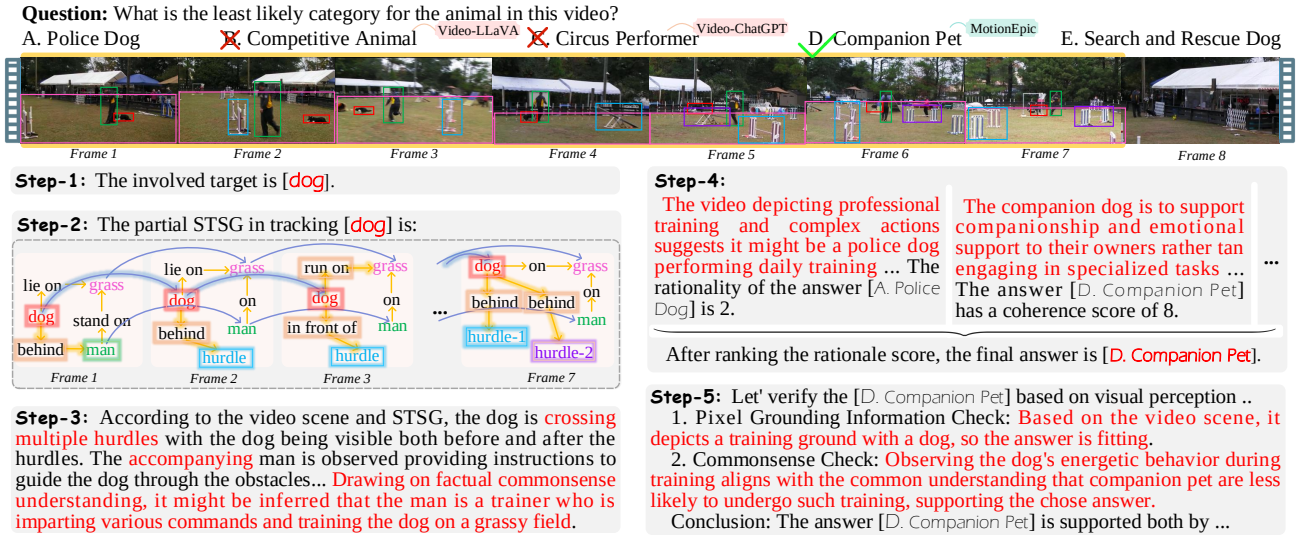

Figure 8: Visualization of qualitative example showcasing how our VoT framework achieves successful video reasoning.

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 part of the 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 aid an intuitive understanding of the superiority of our system. We randomly select an instance where our model gives the correct answer. As shown in Fig. 8, the video displays a complex scene, and the given question is abstract and complex, not directly answerable through the 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. In the Appendix, we show more qualitative visualizations of examples.

6. Conclusion

In this work, we for the first time introduce an innovative solution for complex video reasoning, the Video-of-Thought (VoT) framework. To accomplish the reasoning framework, also a novel video MLLM, MotionEpic, is proposed. 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 sub-problems, 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 the relevant community.

Acknowledgements

This work is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001), CCF-Baidu Open Fund, and the National Natural Science Foundation of China (NSFC) Grant (No. 62336008).

Impact Statement

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 90 2023.

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.

Fei, H., Wu, S., Ren, Y., and Zhang, M. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pp. 6373–6391, 2022.

Fei, H., Li, B., Liu, Q., Bing, L., Li, F., and Chua, T.-S. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023a.

Fei, H., Liu, Q., Zhang, M., Zhang, M., and Chua, T.-S. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5980–5994, 2023b.

Fei, H., Wu, S., Ji, W., Zhang, H., and Chua, T.-S. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 2023c.

Fei, H., Wu, S., Zhang, M., Zhang, M., Chua, T.-S., and Yan, S. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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.

Johnson, J., Gupta, A., and Fei-Fei, L. Image generation from scene graphs. In *Proceedings of the CVPR*, pp. 1219–1228, 2018.

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.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

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.

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.

Wu, S., Fei, H., Cao, Y., Bing, L., and Chua, T.-S. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14734–14751, 2023a.

Wu, S., Fei, H., Ji, W., and Chua, T.-S. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2593–2608, 2023b.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023c.

Wu, S., Fei, H., Zhang, H., and Chua, T.-S. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

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.

Zhang, Z., Zhang, A., Li, M., and Smola, A. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023c.

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, 2023a.

Zhao, Y., Fei, H., Ji, W., Wei, J., Zhang, M., Zhang, M., and Chua, T.-S. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7960–7977, 2023b.

Zheng, L., Fei, H., Li, F., Li, B., Liao, L., Ji, D., and Teng, C. Reverse multi-choice dialogue commonsense inference with graph-of-thought. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19688–19696, 2024.

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. More Configuration Details

A.1. 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?`], 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.

A.2. Implementations

MotionEpic uses the Vicuna-7B (v1.5)[5] as the backbone LLM. We adopt the ViT-L/14[6] as the video encoder, and use the Q-Former[7] 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 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

[5] https://huggingface.co/lmsys/vicuna-7b-v1.5
[6] https://huggingface.co/sentence-transformers/clip-ViT-L-14
[7] https://huggingface.co/spaces/Salesforce/BLIP2

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. More Qualitative Visualizations

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. 9 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 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. 10 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.

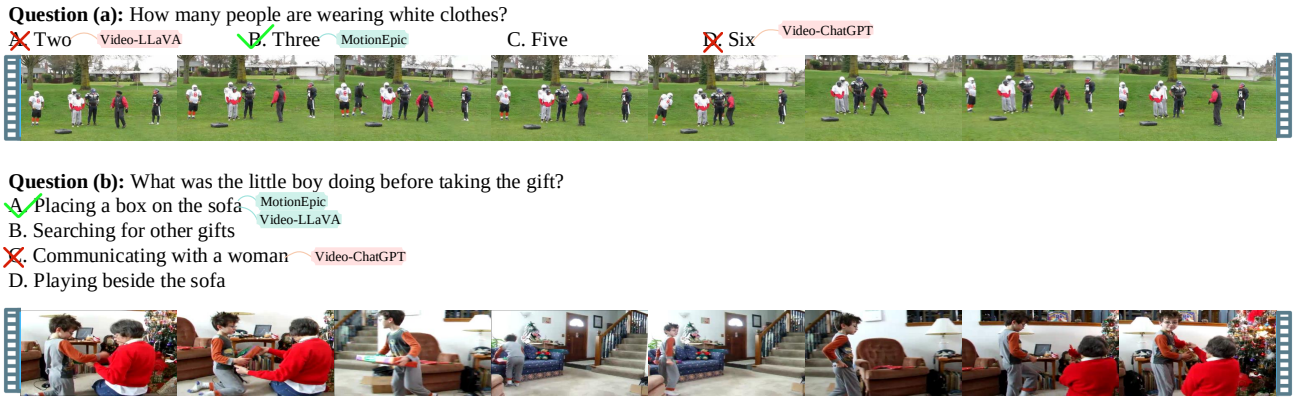

Figure 9: Qualitative examples of perception-level reasoning. The correct answer is marked with a green checkmark, and the wrong answer is marked with a red cross.

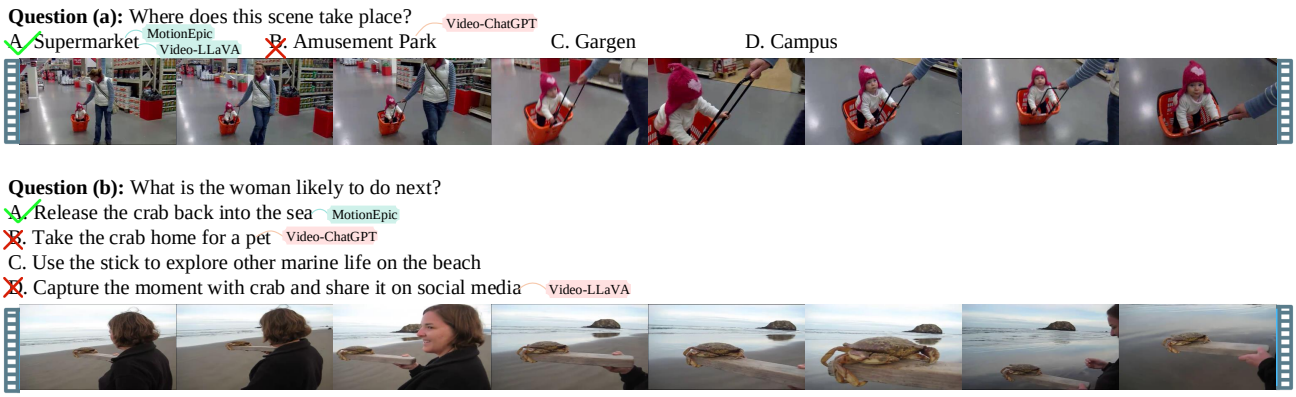

Figure 10: Qualitative examples of cognitive-level reasoning.