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]

[1] National University of Singapore [2] Nanyang Technological University [3] Harbin Institute of Technology (Shenzhen)

ICML International Conference On Machine Learning

Project: https://haofei.vip/VoT/

Paper: https://is.gd/fcfZeO

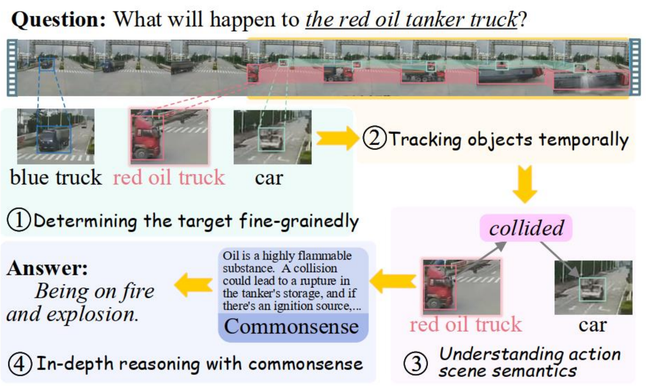
Code: https://github.com/scofield7419/Video-of-Thought

TL;DR

This work for the first time propose the video Chain-of-Thought (CoT) reasoning framework, Video-of-Thought (VoT), for achieving human-level video reasoning. VoT 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.

Motivation

➢ Keynotes of human cognition patterns on video understanding/reasoning:



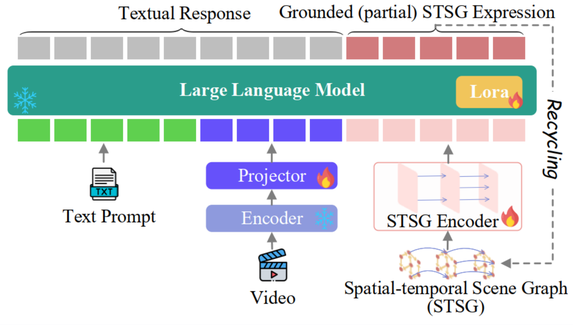
First, *to achieve precise content perception, a fine-grained perceptive pixel understanding of the video movement is necessary*

Second, *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.*

Third, *for humans, video reasoning is not an instantaneous process but follows a multi-hop procedure from lower level to higher level.*

Methodology

➢ **MotionEpic**: Fine-grained Spatial-temporal Grounded Video MLLM



Methodology

➢ **Video-of-Thought** Reasoning Framework

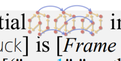
Step-1: Task Definition and Target Identification

Input: 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?

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

Step-2: Object Tracking

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

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

Step-3: Action Analyzing

Input: 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.

Ouput: [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: Question Answering via Ranking

Input: A B C D
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.

Ouput: 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 ...

Ranking

Ouput: 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].

Step-5: Answer Verification

Input: 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.

Ouput: 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.

Experiment

➢ Quantitative Results

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**

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**

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

➢ Analysis

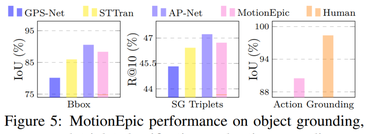


Figure 5: MotionEpic performance on object grounding, scene graph triplet classification, and action grounding.

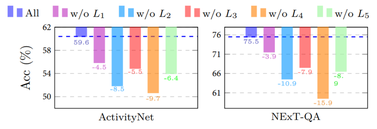


Figure 6: Performance drop (zero-shot) of MotionEpic after ablating different grounding-aware tuning item.

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

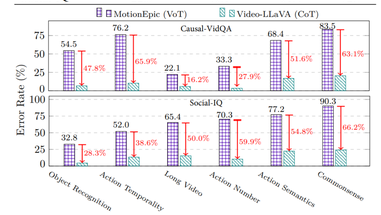


Figure 7: **Above Table**: human evaluation of video QA. **Below Figure**: error rate under various specific categories.

➢ Video Reasoning Visualization