

# Video-of-Thought

## Step-by-Step Video Reasoning from Perception to Cognition



**Hao Fei<sup>1</sup>, Shengqiong Wu<sup>1</sup>, Wei Ji<sup>1</sup>, Hanwang Zhang<sup>2</sup>  
Meishan Zhang<sup>3</sup>, Mong-Li Lee<sup>1</sup>, Wynne Hsu<sup>1</sup>**

<sup>1</sup> National University of Singapore

<sup>2</sup> Nanyang Technological University

<sup>3</sup> Harbin Institute of Technology (Shenzhen)

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

Paper: <https://is.gd/fcfZeO>

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

# 1. Motivation

- Existing research of video understanding mostly falls prey to the type of straightforward perceptual-level understanding, i.e., for **simple videos**.
- 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,
  - but also a profound grasp of the **underlying implications** behind pixels.

# 1. Motivation

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

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

**Question:** What will happen to *the red oil tanker truck*?



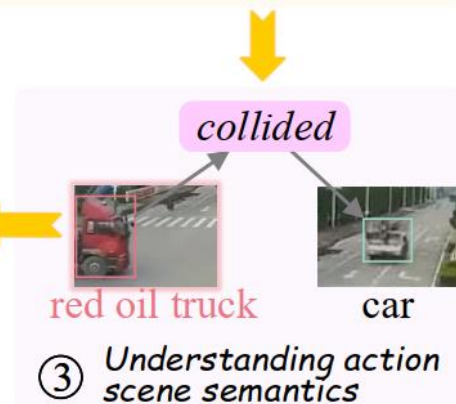
① Determining the target fine-grainedly

**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

④ In-depth reasoning with commonsense



# 1. Motivation

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

B. 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.

**Question:** What will happen to *the red oil tanker truck*?



# 1. Motivation

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

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

**Question:** What will happen to *the red oil tanker truck*?



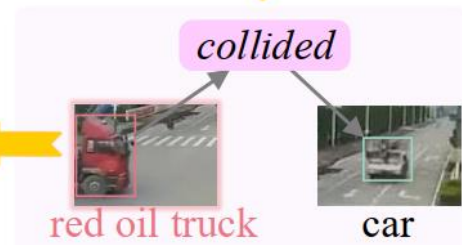
① Determining the target fine-grainedly

**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

④ In-depth reasoning with commonsense



③ Understanding action scene semantics

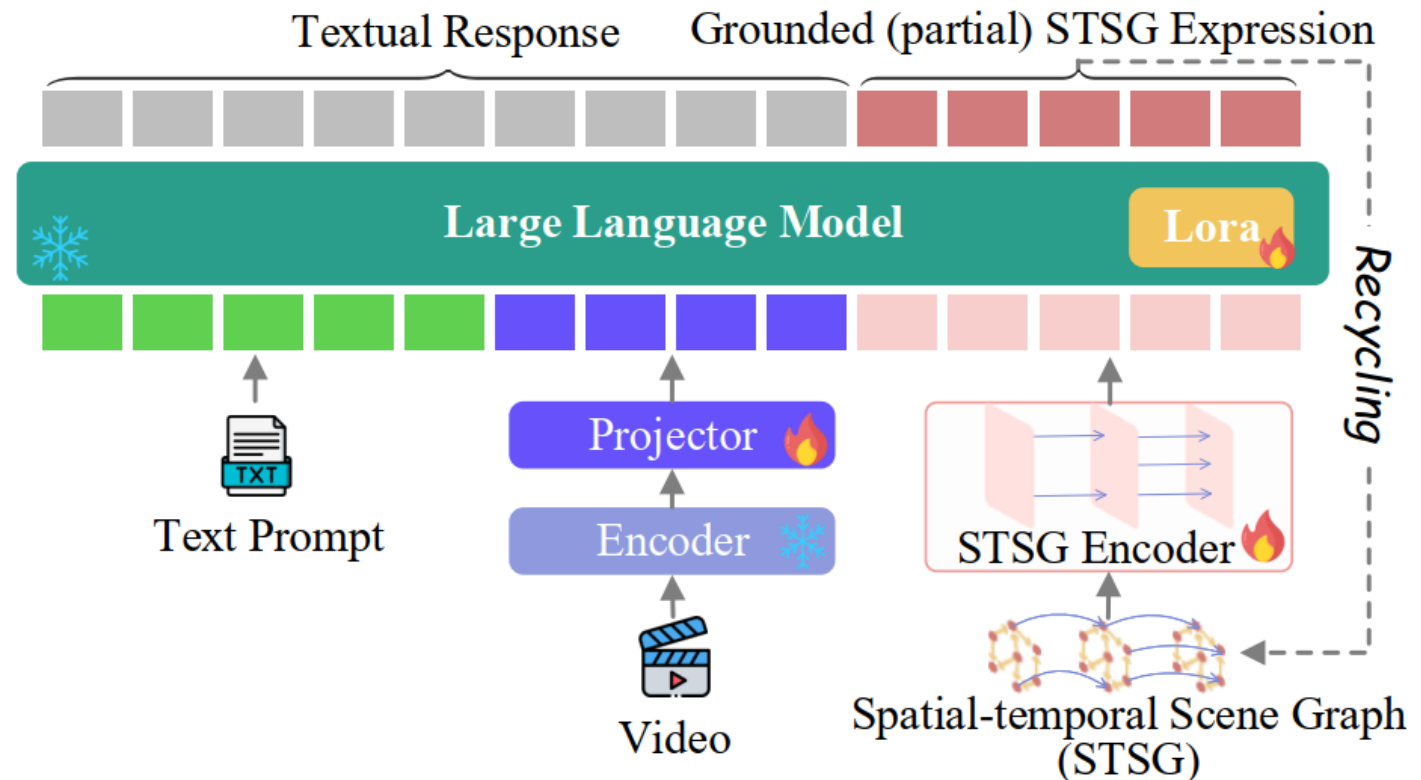


## 2. Our Proposal

### I: MotionEpic: Fine-grained Spatial-temporal Grounded Video MLLM

➤ introduce a novel video LLM, **MotionEpic**

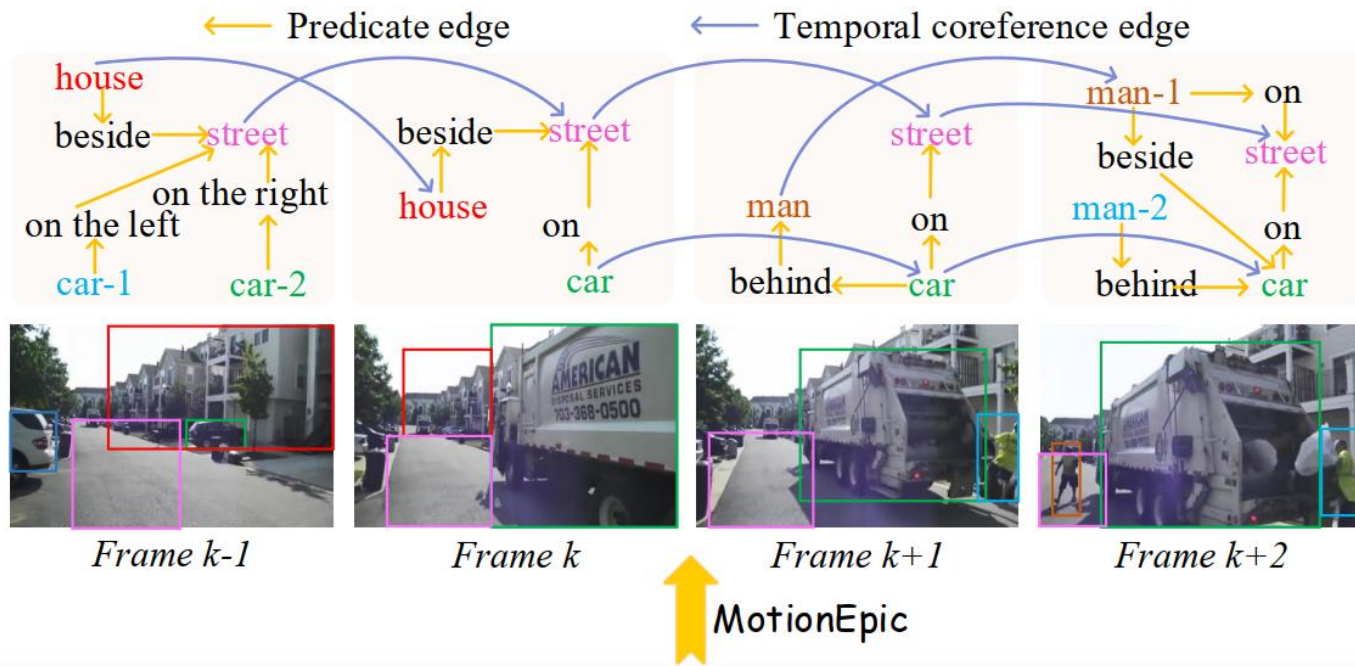
- text prompt,
- video,
- spatial-temporal scene graph (STSG) representation of video



# 2. Our Proposal

## I: MotionEpic: Fin

- introduce a novel ...
- Integrating



```

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

### I: MotionEpic: Fine-grained Spatial-temporal Grounded Video MLLM

- introduce a novel video LLM, **MotionEpic**
  - Integrating STSG Representation, advantage:
    - ❑ STSG intuitively depicts the underlying core semantics representations of videos while filtering the less-informative background information, aiding the perceptive understanding of videos.
    - ❑ STSG, a type of fine-grained structural feature helps effectively model the compositional spatiotemporal semantics.



### I: MotionEpic: Fine-grained Spatial-temporal Grounded Video MLLM

➤ introduce a novel video LLM, **MotionEpic**

- Fine-grained Video-Scene Grounding-aware Tuning

#### 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.

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



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


**Input** 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?


**Output** 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 .

**Output** 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"}, ...]}...].

### 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.

**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: Question Answering via Ranking



**Input** 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.

**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 ...

↓ Ranking

**Output** 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.

**Output** 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.

## 2. Our Proposal

### II: Video-of-Thought Reasoning Framework


- Problem decomposition:

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

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

Input

**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?

Output

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



## 2. Our Proposal

### II: Video-of-Thought Reasoning Framework


- Problem decomposition:

#### Step-2: Object Tracking

Input

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

Output

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"}, ...]}...].

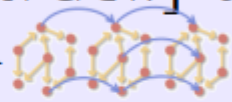
## 2. Our Proposal

### II: Video-of-Thought Reasoning Framework

- Problem decomposition:

#### 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.



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...]



## 2. Our Proposal



### II: Video

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

➤ Pro

**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.**

**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 ...

↓ Ranking

**Output**

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].**



## 2. Our Proposal

### II: Video-of-Thought Reasoning Framework

➤ Pro

#### 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.



Output

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.

# 3. Experiment

- Settings

- Task: *complex video QA*

- Grounding-aware Tuning Corpus: Action Genome (10K) & WebVid (350K)

- Baselines and Implementations: ...

# 3. Experiment

## ➤ Main Performance on Video QA Reasoning

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
<b>• SoTA baselines</b>								
InternVideo	63.9	62.7	65.6	54.9	51.9	-	-	-
LLaMA-VQA	<u>71.0</u>	66.2	67.9	57.2	52.7	-	-	-
VLAP	69.6	<u>70.0</u>	<u>70.4</u>	<u>65.9</u>	<u>62.2</u>	-	-	-
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
<b>• CoT</b>								
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	<u>66.1</u>	<u>71.7</u>	<u>43.0</u>
<b>• VoT</b>								
MotionEpic	<b>73.4</b>	<b>71.5</b>	<b>72.6</b>	<b>66.6</b>	<b>62.7</b>	<b>70.8</b>	<b>72.8</b>	<b>45.0</b>

# 3. Experiment

➤ Main Performance on Video QA Reasoning

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
<b>• SoTA baselines</b>								
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
<b>• CoT</b>								
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	<u>78.5</u>	<u>77.2</u>	<u>70.1</u>	<u>70.8</u>	<u>52.4</u>	<u>71.2</u>	<u>69.1</u>	<u>55.0</u>
<b>• VoT</b>								
MotionEpic	<b>81.2</b>	<b>83.0</b>	<b>74.3</b>	<b>73.7</b>	<b>54.7</b>	<b>74.5</b>	<b>73.8</b>	<b>58.6</b>



# 3. Experiment

➤ Main Performance on Video QA Reasoning

Table 3: Results on NExT-QA data.

Model	Acc@All	Acc@C	Acc@T	Acc@D
<b>• SoTA baselines</b>				
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	<u>75.5</u>	<u>74.9</u>	<u>72.3</u>	<u>82.1</u>
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
<b>• CoT</b>				
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
<b>• VoT</b>				
MotionEpic	<b>76.0</b>	<b>75.8</b>	<b>74.6</b>	<b>83.3</b>

# 3. Experiment

## ➤ Zero-shot Performance

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.
<b>• Zero-shot SoTA baselines</b>					
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	<u>59.0</u>	56.0
<b>• CoT</b>					
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	<u>63.1</u>	<u>50.0</u>	<u>61.9</u>	56.5	<u>57.8</u>
<b>• VoT</b>					
MotionEpic	<b>66.2</b>	<b>54.6</b>	<b>66.5</b>	<b>61.7</b>	<b>62.3</b>
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

# 3. Experiment

- Analyses on MotionEpic Video MLLM
  - Probing Video Grounding Ability

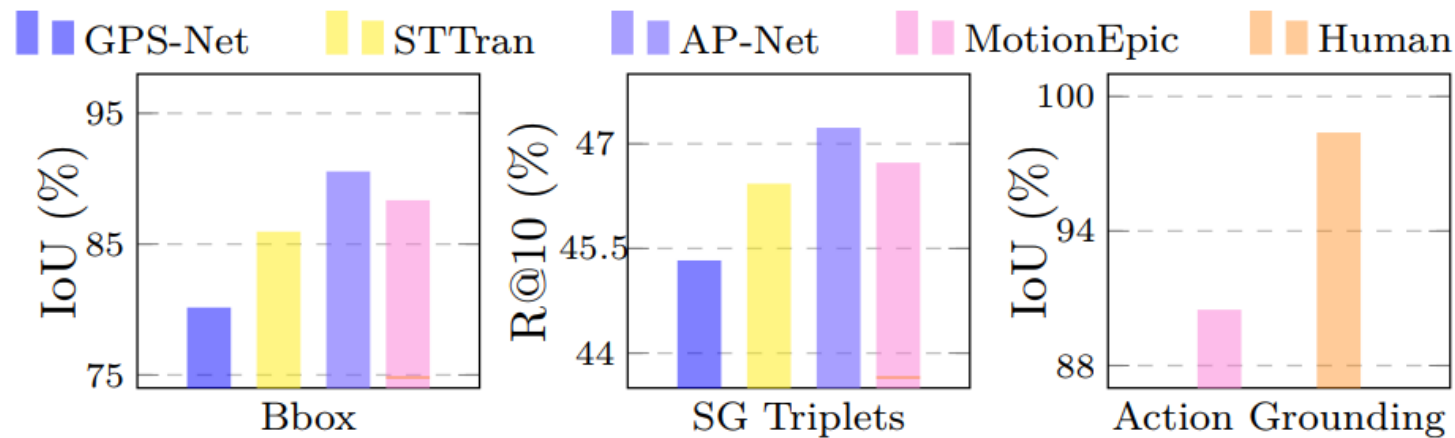


Figure 5: MotionEpic performance on object grounding, scene graph triplet classification, and action grounding.

# 3. Experiment

- Analyses on MotionEpic Video MLLM
  - Influence of Various Grounding-aware Tuning Strategies

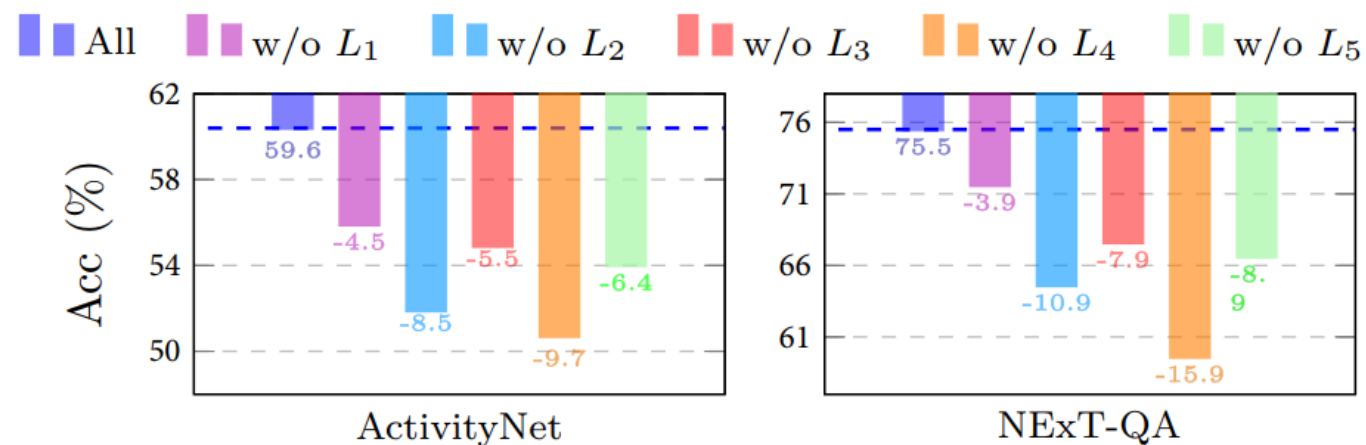


Figure 6: Performance drop (zero-shot) of MotionEpic after ablating different grounding-aware tuning item.

# 3. Experiment

- Analyses on VoT Video Reasoning Framework
  - Reasoning Ability Breaking-down

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

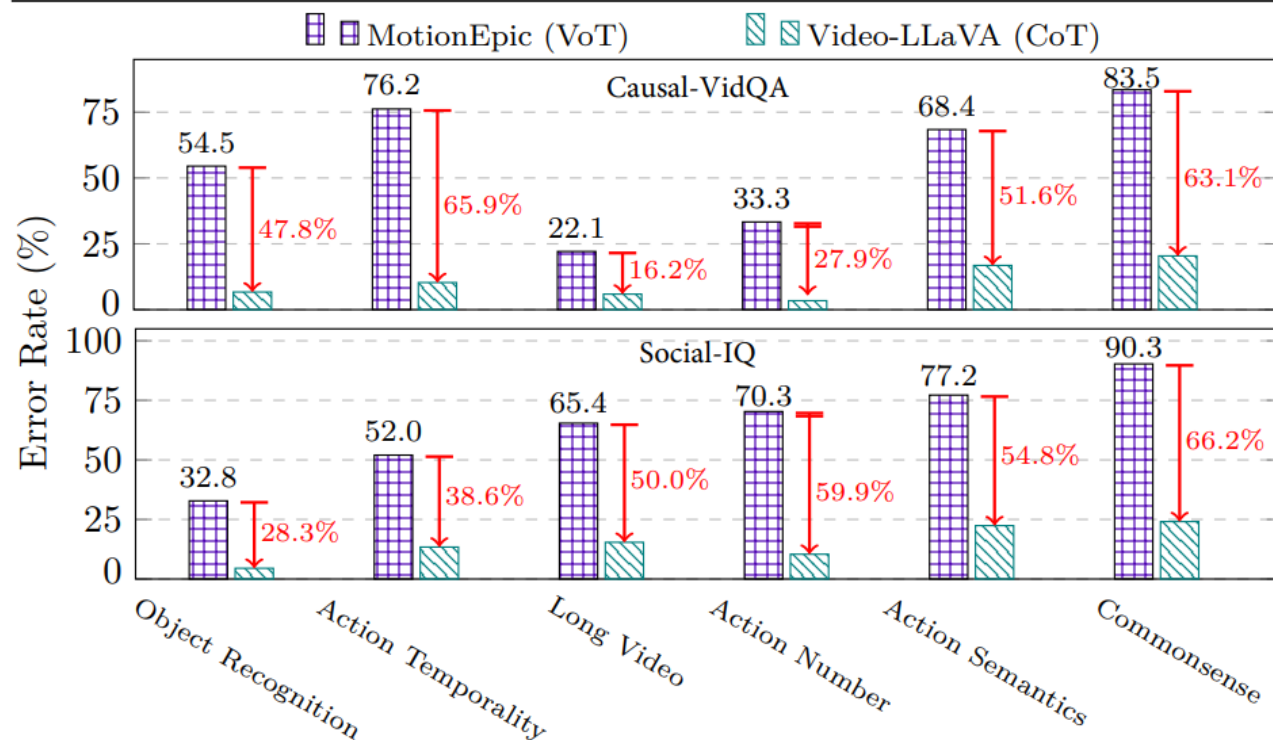


Figure 7: **Above Table:** human evaluation of video QA. **Below Figure:** error rate under various specific categories.

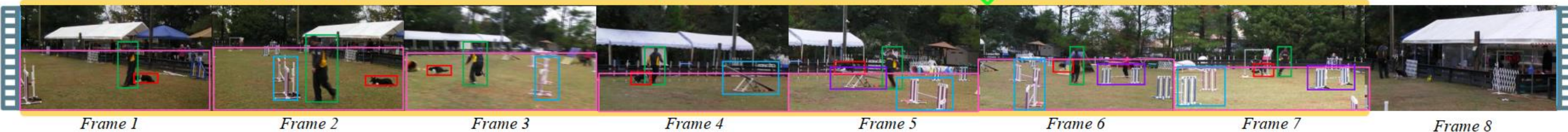


# 3. Experiment

## ➤ Analyses on VoT Video Reasoning Framework

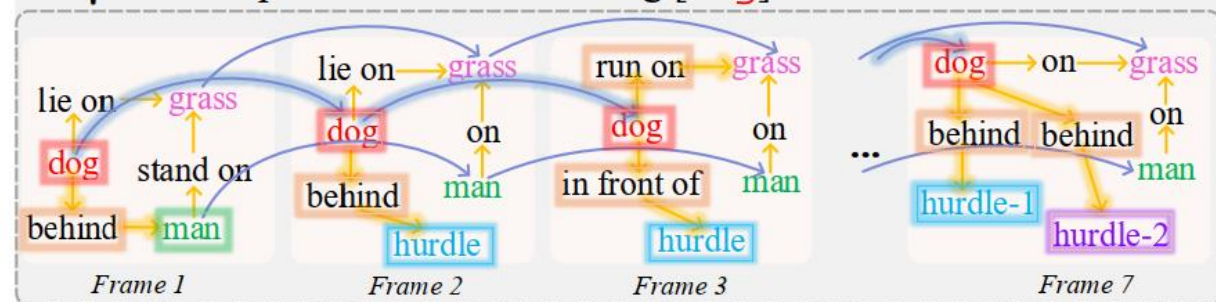
**Question:** What is the least likely category for the animal in this video?

A. Police Dog ~~X~~ Competitive Animal <sup>Video-LLaVA</sup> ~~X~~ Circus Performer <sup>Video-ChatGPT</sup> ~~X~~ Companion Pet <sup>MotionEpic</sup>  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 than 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's 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 chosen answer.

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



# 3. Experiment

- Analyses on VoT Video Reasoning Framework
  - Video Reasoning Visualization

**Question (a):** How many people are wearing white clothes?

- ~~X~~ Two Video-LLaVA
- B. Three MotionEpic
- C. Five
- ~~X~~ 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
- ~~X~~ C. Communicating with a woman Video-ChatGPT
- D. Playing beside the sofa



# 3. Experiment

- Analyses on VoT Video Reasoning Framework
  - Video Reasoning Visualization

**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





Thanks  
Q&A