

# Can Video LLMs Refuse to Answer?

## Alignment for Answerability in Video Large Language Models

---

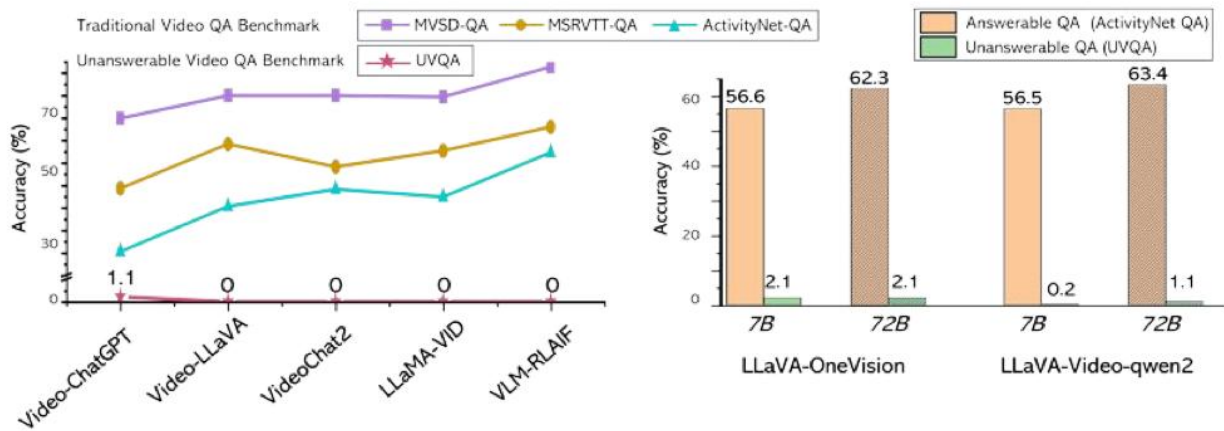
*The Thirteenth International Conference on Learning Representations  
(ICLR 2025)*

Eunseop Yoon<sup>1\*</sup>, Hee Suk Yoon<sup>1\*</sup>, Mark Hasegawa-Johnson<sup>2</sup>, Chang D. Yoo<sup>1</sup>



# Can Video LLMs Refuse to Answer?

**Problem:** Video LLMs are not trained to handle unanswerable questions (only trained with QA pairs from video content)



Answer

A No, there is no cat sitting on the table in the video. The video only shows a person reading a book and drinking coffee.

Video LLM



Video Frames

Q Is there a cat in the video?

Question

Answer

A The cat is a Siamese cat.

Video LLM



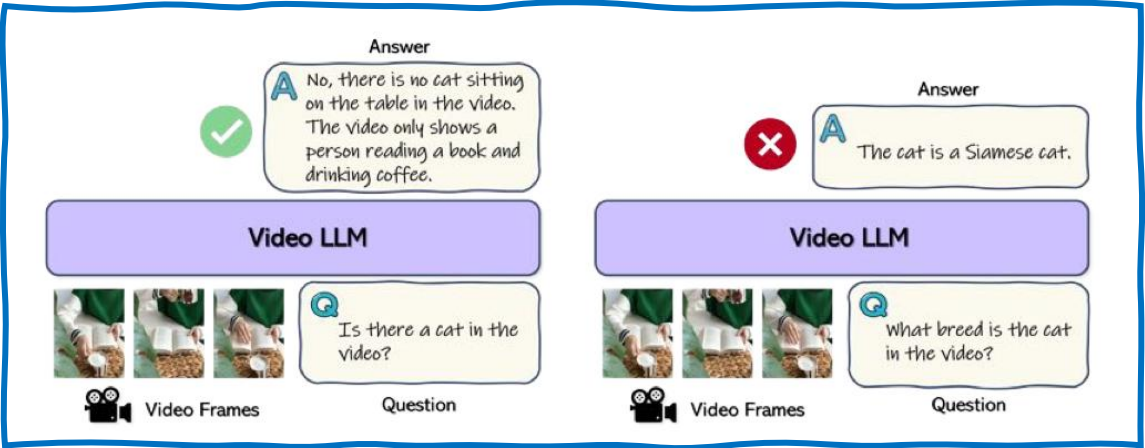
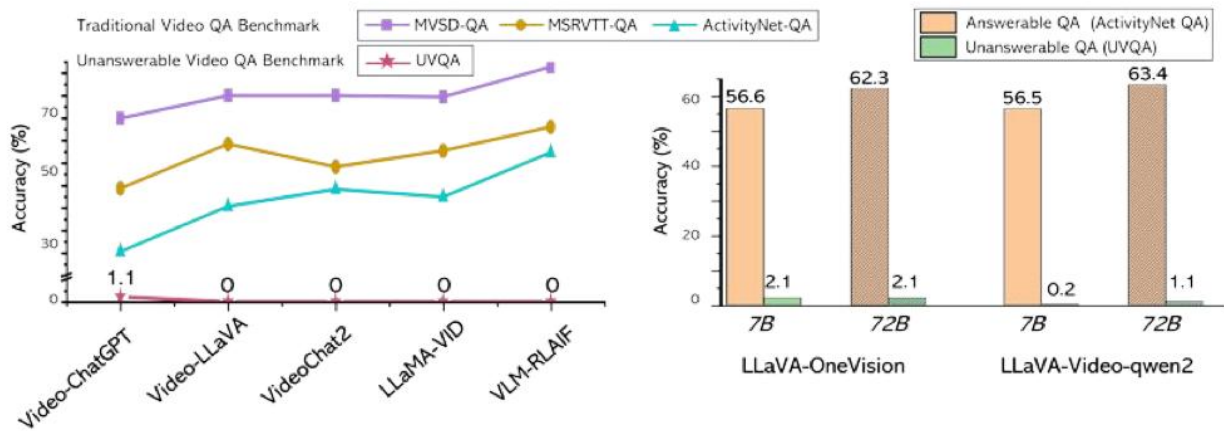
Video Frames

Q What breed is the cat in the video?

Question

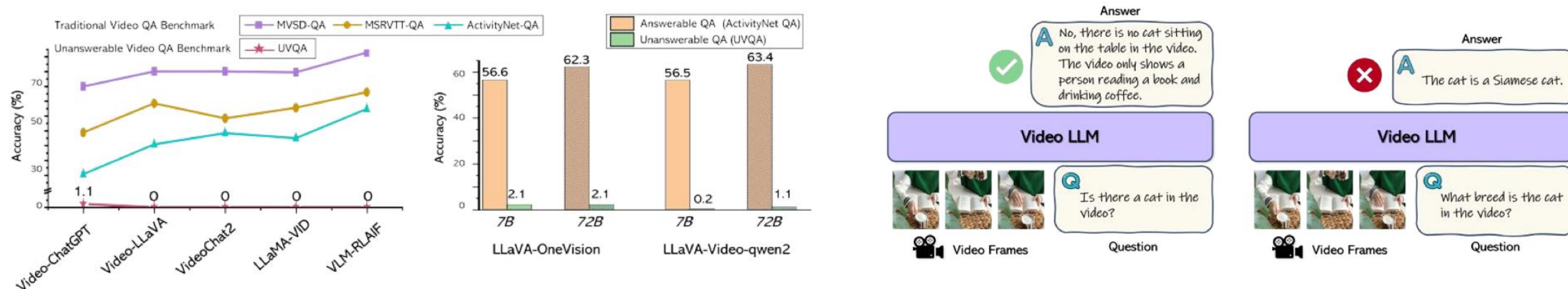
# Can Video LLMs Refuse to Answer?

**Problem:** Video LLMs are not trained to handle unanswerable questions (only trained with QA pairs from video content)



# Can Video LLMs Refuse to Answer?

**Problem:** Video LLMs are not trained to handle unanswerable questions (only trained with QA pairs from video content)

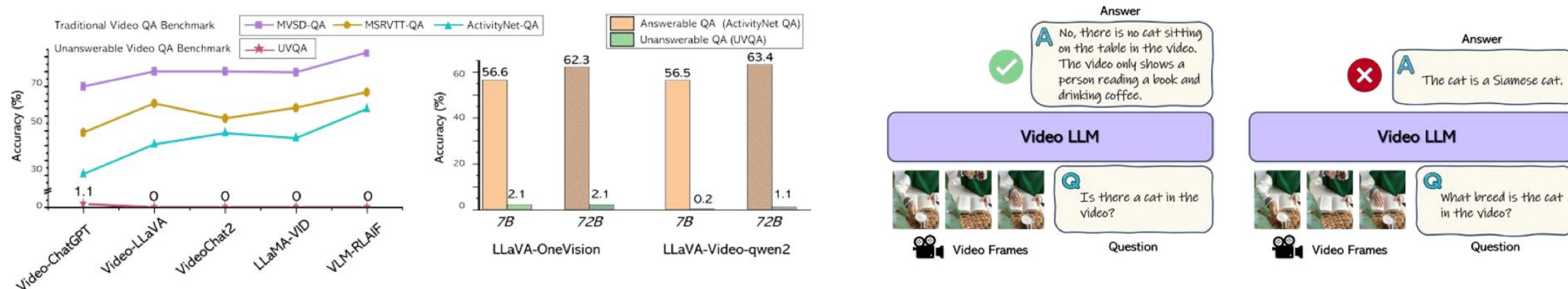


Explicit prompting such as *"If the question cannot be answered using the video content, state that it is unanswerable and provide a reason"* does not solve the problem.

Base Model	$f(\cdot)$	Answerability	Alignment Performance				Absolute Performance	
		F1	$S_{\text{ex-ref.}} \downarrow$	$S_{\text{permis.}} \uparrow$	$S_{\text{disc.}} \uparrow$	$S_{\text{align}} \uparrow$	$S_{\text{acc.}} \uparrow$	$\text{LLM}_{\text{score}} \uparrow$
VLM-RLAIF (7B)	unaligned	0.00	0	0	0	0.33	0.25	2.36
	prompt-aligned	0.08	0.02	0	0.06	0.35	0.27	1.66
	DPO (ours)	<b>0.66</b>	0.08	<b>0.5</b>	<b>0.64</b>	<b>0.69</b>	<b>0.53</b>	<b>2.93</b>
LLaVA-Video-Qwen2 (7B)	unaligned	0.00	0	0	0	0.33	<b>0.38</b>	2.26
	prompt-aligned	<b>0.04</b>	0.01	0	<b>0.03</b>	<b>0.34</b>	<b>0.38</b>	<b>2.33</b>
LLaVA-Video-Qwen2 (72B)	unaligned	0.02	0	0	0	0.33	0.38	2.35
	prompt-aligned	<b>0.12</b>	0	0	<b>0.07</b>	<b>0.36</b>	<b>0.41</b>	<b>2.50</b>

# Can Video LLMs Refuse to Answer?

**Problem:** Video LLMs are not trained to handle unanswerable questions (only trained with QA pairs from video content)



Explicit prompting such as *"If the question cannot be answered using the video content, state that it is unanswerable and provide a reason"* does not solve the problem.

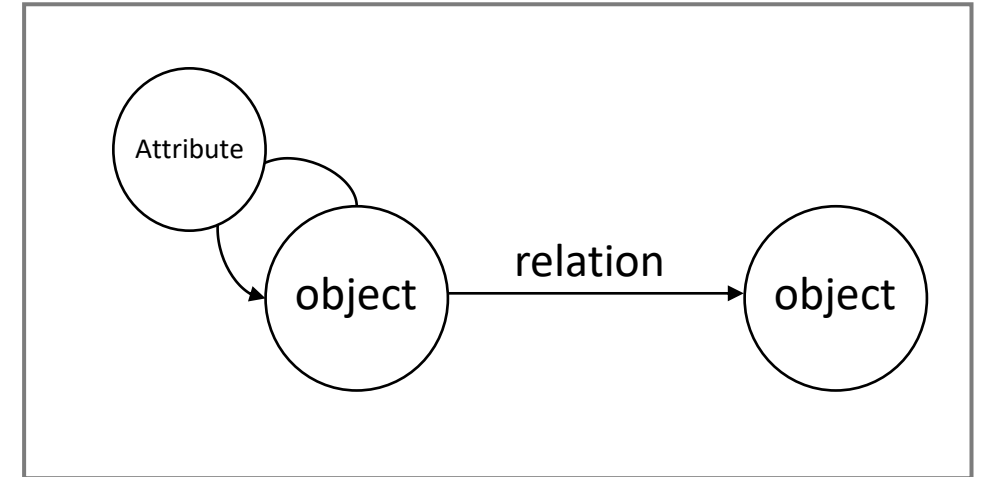
Base Model	$f(\cdot)$	Answerability	Alignment Performance				Absolute Performance	
		F1	$S_{\text{ex-ref.}} \downarrow$	$S_{\text{permis.}} \uparrow$	$S_{\text{disc.}} \uparrow$	$S_{\text{align}} \uparrow$	$S_{\text{acc.}} \uparrow$	$\text{LLM}_{\text{score}} \uparrow$
VLM-RLAIF (7B)	unaligned	0.00	0	0	0	0.33	0.25	2.36
	→ prompt-aligned	0.08	0.02	0	0.06	0.35	0.27	1.66
	DPO (ours)	<b>0.66</b>	0.08	<b>0.5</b>	<b>0.64</b>	<b>0.69</b>	<b>0.53</b>	<b>2.93</b>
LLaVA-Video-Qwen2 (7B)	unaligned	0.00	0	0	0	0.33	<b>0.38</b>	2.26
	→ prompt-aligned	<b>0.04</b>	0.01	0	<b>0.03</b>	<b>0.34</b>	<b>0.38</b>	<b>2.33</b>
LLaVA-Video-Qwen2 (72B)	unaligned	0.02	0	0	0	0.33	0.38	2.35
	→ prompt-aligned	<b>0.12</b>	0	0	<b>0.07</b>	<b>0.36</b>	<b>0.41</b>	<b>2.50</b>

# Category for unanswerability

Inspired by structured vision understanding through **scene graphs**,



[example video]



- **case1.** object in the question is not contained in the video

Q: What breed is the **cat** in the video?

A: It is unanswerable because there is no **cat** in the video

- **case2.** objects in the question are contained in the video but the relation between the objects is not matched

Q: What is the color of the cup that the person is **cleaning**?

A: It is unanswerable because the person in the video is not cleaning any cup. The person picks up the cup to drink.

- **case3.** object category in the question is contained in the video but the attribute is not matched

Q: What is the person in the **red** shirt doing?

A: It is unanswerable because there is no person in the **red** shirt. The person in the video is wearing a green shirt



# Category for unanswerability

Inspired by structured vision understanding through **scene graphs**,



[example video]

- **case1.** object in the question is not contained in the video

Q: What breed is the **cat** in the video?

A: It is unanswerable because there is no **cat** in the video

- **case2.** objects in the question are contained in the video but the relation between the objects is not matched

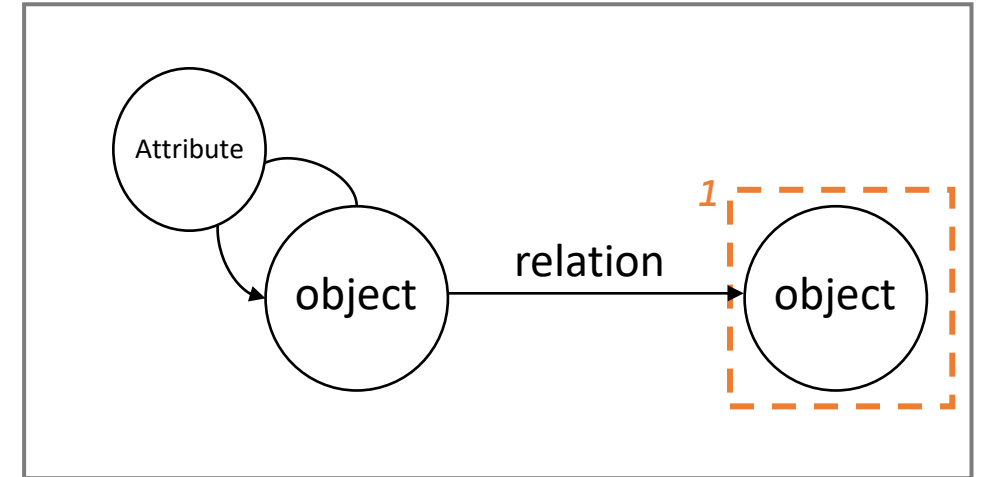
Q: What is the color of the cup that the person is **cleaning**?

A: It is unanswerable because the person in the video is not cleaning any cup. The person picks up the cup to drink.

- **case3.** object category in the question is contained in the video but the attribute is not matched

Q: What is the person in the **red** shirt doing?

A: It is unanswerable because there is no person in the **red** shirt. The person in the video is wearing a green shirt



# Category for unanswerability

Inspired by structured vision understanding through **scene graphs**,



[example video]

- **case1.** object in the question is not contained in the video

Q: What breed is the **cat** in the video?

A: It is unanswerable because there is no **cat** in the video

- **case2.** objects in the question are contained in the video but the relation between the objects is not matched

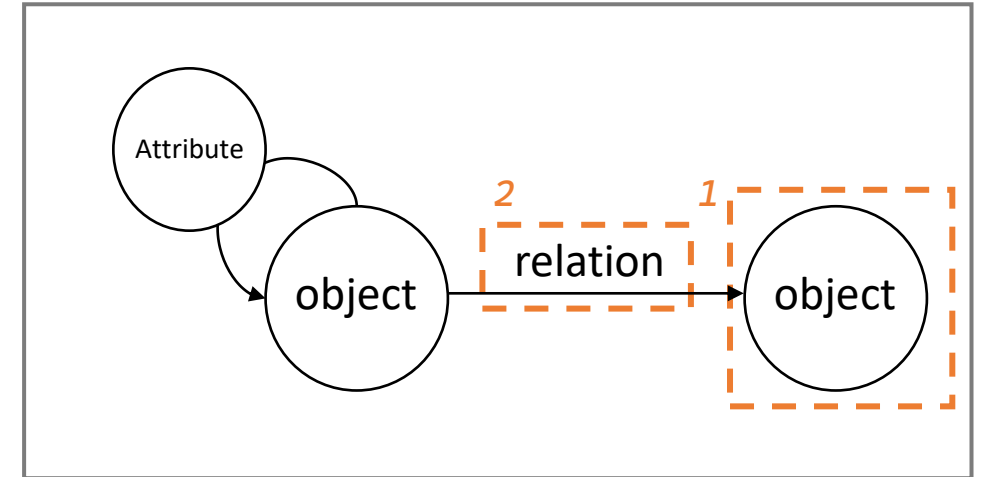
Q: What is the color of the cup that the person is **cleaning**?

A: It is unanswerable because the person in the video is not cleaning any cup. The person picks up the cup to drink.

- **case3.** object category in the question is contained in the video but the attribute is not matched

Q: What is the person in the **red** shirt doing?

A: It is unanswerable because there is no person in the **red** shirt. The person in the video is wearing a green shirt





# Category for unanswerability

Inspired by structured vision understanding through **scene graphs**,



[example video]

- **case1.** object in the question is not contained in the video

Q: What breed is the **cat** in the video?

A: It is unanswerable because there is no **cat** in the video

- **case2.** objects in the question are contained in the video but the relation between the objects is not matched

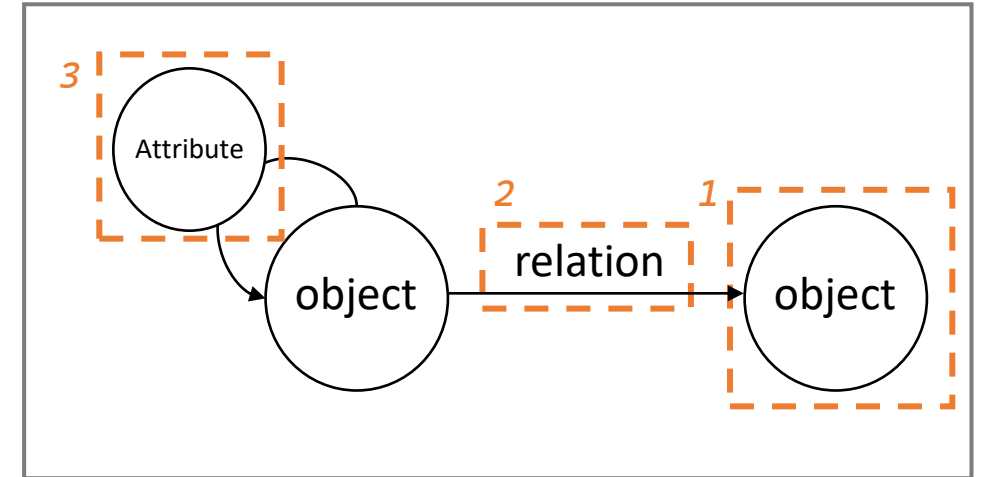
Q: What is the color of the cup that the person is **cleaning**?

A: It is unanswerable because the person in the video is not cleaning any cup. The person picks up the cup to drink.

- **case3.** object category in the question is contained in the video but the attribute is not matched

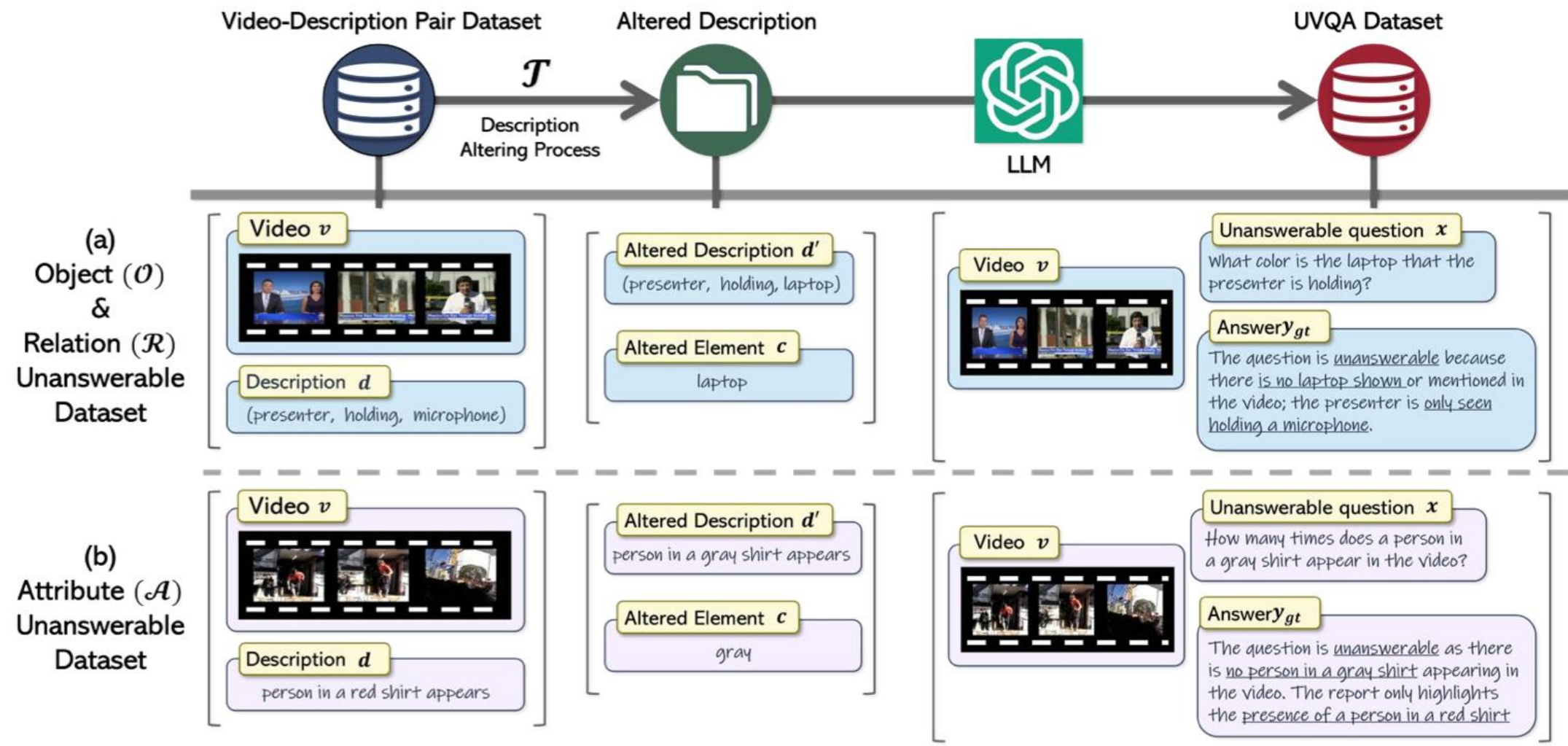
Q: What is the person in the **red** shirt doing?

A: It is unanswerable because there is no person in the **red** shirt. The person in the video is wearing a green shirt



# UVQA: Unanswerable Video Question Answering

Solution: **Synthetic Dataset – QA pairs that goes beyond the video content**




# Can Video LLMs Refuse to Answer? Alignment for Answerability in Video Large Language Models

## Result:

- possible to refuse answer when the question is including information beyond the video content

Video  $\mathcal{V}$



Unanswerable Question  $X$   
Why is the police officer using chopsticks to investigate the scene?

Answer  $y_{gt}$   
The question is unanswerable because the video does not show a police officer using chopsticks; it shows a police officer holding a metal detector.

Unaligned Model

The police officer is using chopsticks to investigate the scene because they are a tool commonly used in Asian cuisine.

SFT Aligned Model


The question is unanswerable because the video does not show a police officer using chopsticks to investigate the scene.

DPO Aligned Model

The question is unanswerable because the video does not show or mention the presence of chopsticks. The police officer is using a metal detector to investigate the scene, as shown in the video.

(a) Object-related ( $\mathcal{O}$ ) unanswerable QA prediction

Video  $\mathcal{V}$



Unanswerable Question  $X$   
How does the child's action of pressing the bicycle affect its shape?

Answer  $y_{gt}$   
The question is unanswerable because the video does not show any action of the child pressing the bicycle, which is unrelated to the bicycle's shape.

Unaligned Model

The child's action of pressing the bicycle causes it to change shape, as shown in the video. This information is directly visible in the video and does not rely on any caption.

SFT Aligned Model


The question is unanswerable because the video does not show the child pressing the bicycle, and therefore, it is impossible to determine how pressing the bicycle would affect its shape.

DPO Aligned Model

The question is unanswerable because the video does not show the child pressing the bicycle, nor does it show any changes in the bicycle's shape.

(b) Relation-related ( $\mathcal{R}$ ) unanswerable QA prediction

Video  $\mathcal{V}$



Unanswerable Question  $X$   
How does the person in blue-green react after they stop looking down?

Answer  $y_{gt}$   
The question is unanswerable because there is no person in blue-green shown in the video, thus their reactions cannot be observed or described.

Unaligned Model

After stopping to look down, the person in blue-green continues climbing up the rock. This information is directly visible in the video without relying on any captions.

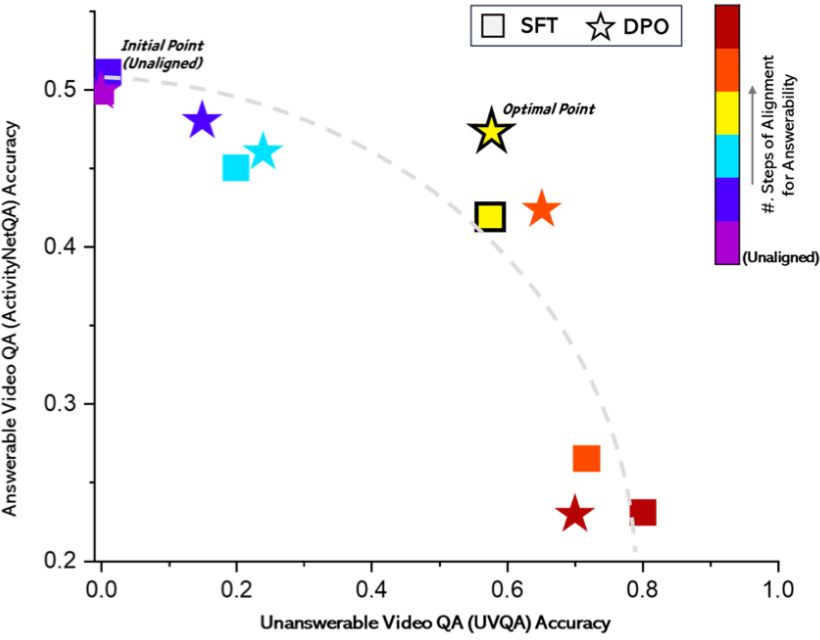
SFT Aligned Model

The question is unanswerable because the video does not provide information about the person in blue-green's reaction after they stop looking down.

DPO Aligned Model

The video does not show the person in blue-green looking down, so it is not possible to determine how they react after stopping.

(c) Attribute-related ( $\mathcal{A}$ ) unanswerable QA prediction



: trade-off between answerable QA accuracy and unanswerable QA accuracy

**Thank you**