Evaluating Video Language Models

Black Swan: Abductive and Defeasible Video Reasoning in Unpredictable Events

Sahithya Ravi: sahiravi@cs.ubc.ca,
Aditya Chinchure: aditya10@cs.ubc.ca

Can you learn meaning only from text?

A monkey grabbed a plastic bag and jumped out the window of a moving bus.

Can you learn meaning only from text?

A monkey grabbed a plastic bag and jumped out the window of a moving bus.

💭 **Why did the monkey grab the bag?**	(Stealing food? Curious?)
💭 **How did it look while jumping?**	(Was it frantic, playful, or scared?)
💭 **What was inside the bag?**	(Food?)
💭 **What were the humans in the scene doing?**	(Chasing it? Ignoring it?)

Can you learn meaning only from text?

A monkey grabbed a plastic bag and jumped out the window of a moving bus.

🤣😱

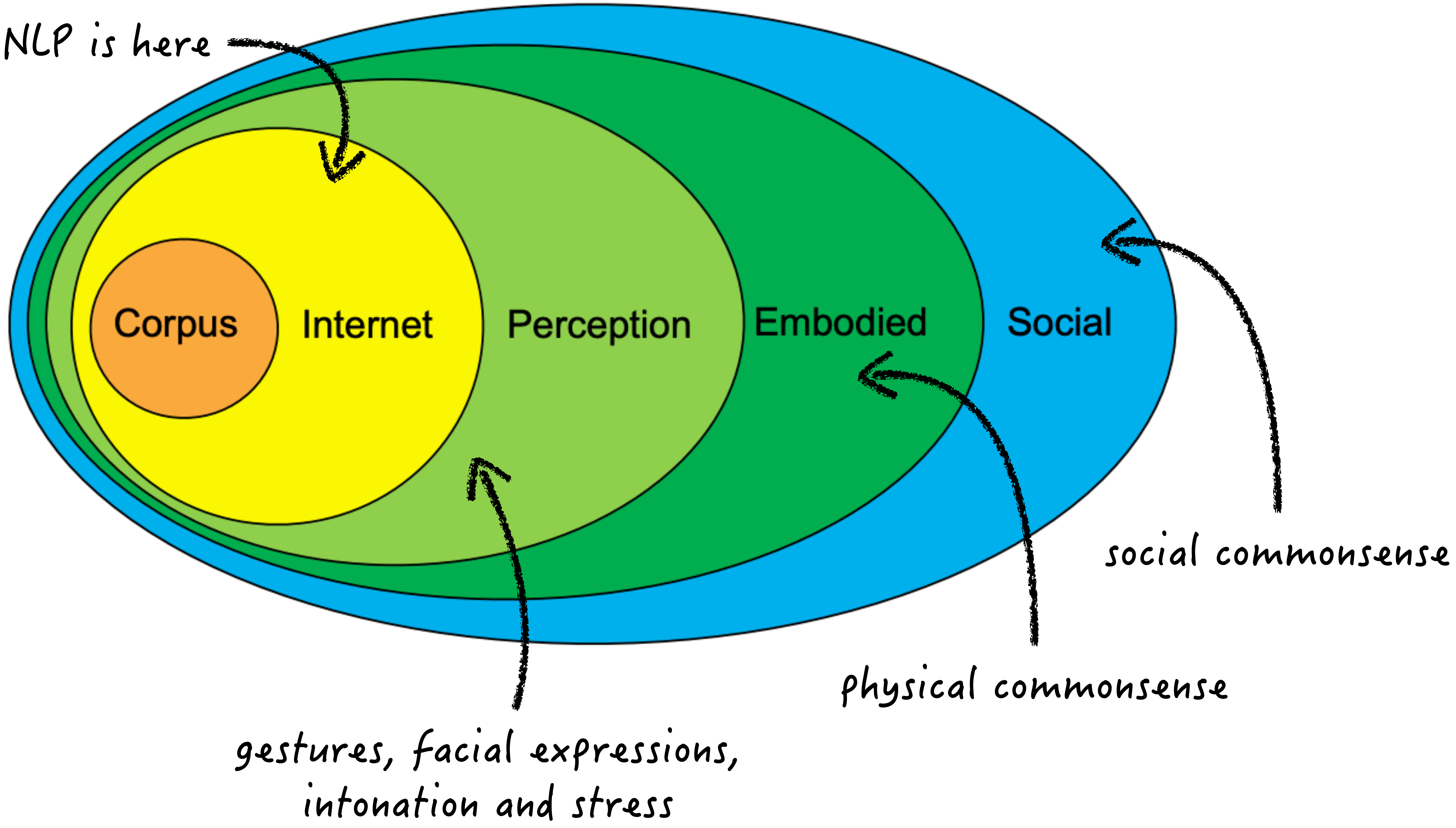
💭 **Why did the monkey grab the bag?**

💭 **How did it look while jumping?**

💭 **What was inside the bag?**

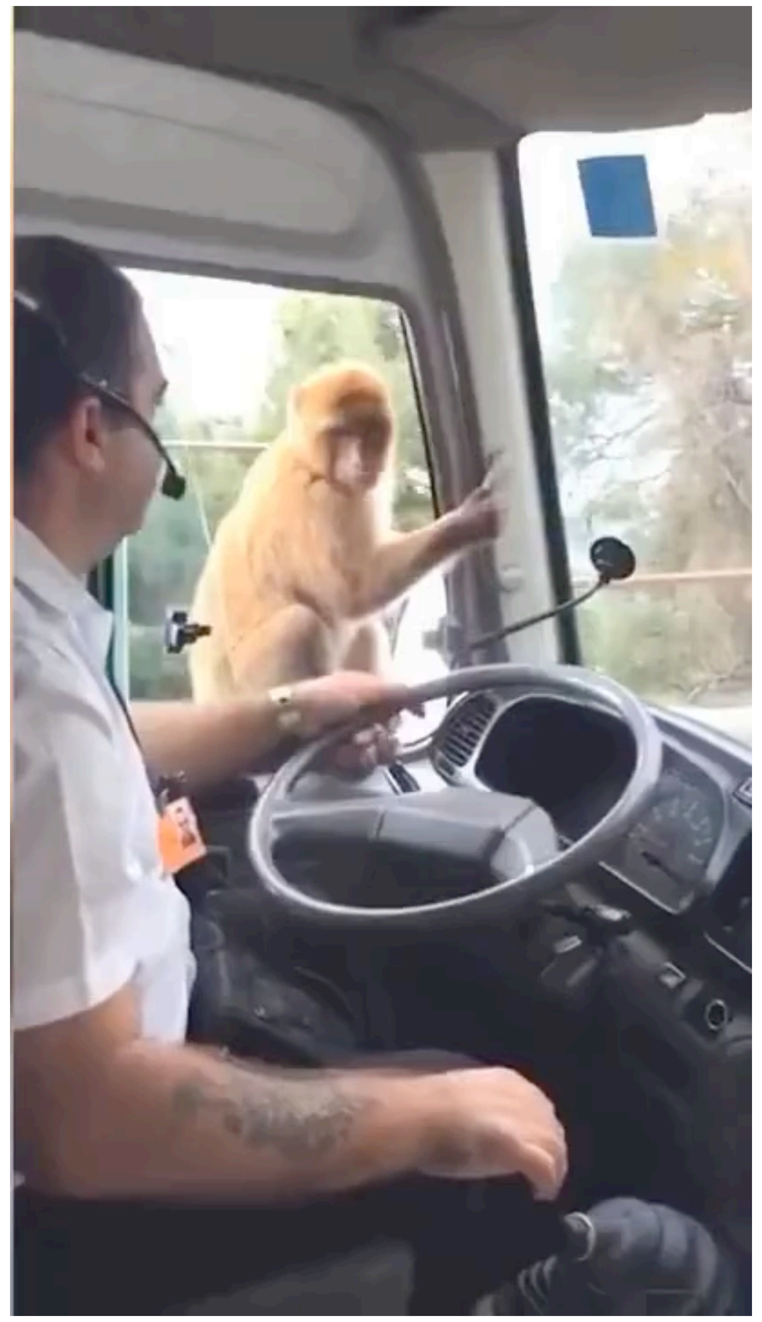
💭 **What were the humans in the scene doing?**

Can you learn meaning only from text?



Experience Grounds Language. Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, Joseph Turian. EMNLP 2020.

Multimodal Model Skills?



- Perception - How many humans are there?
- Causal Reasoning – Why did the monkey jump?
- Temporal Understanding – What happened before and after?
- Physical Intuition – Could a monkey safely jump from a moving bus?
- Social & Commonsense Knowledge – Was the monkey stealing or playing?

Multimodal Models need to *see* 👁️, *interpret* 💡, and *reason* 🧐

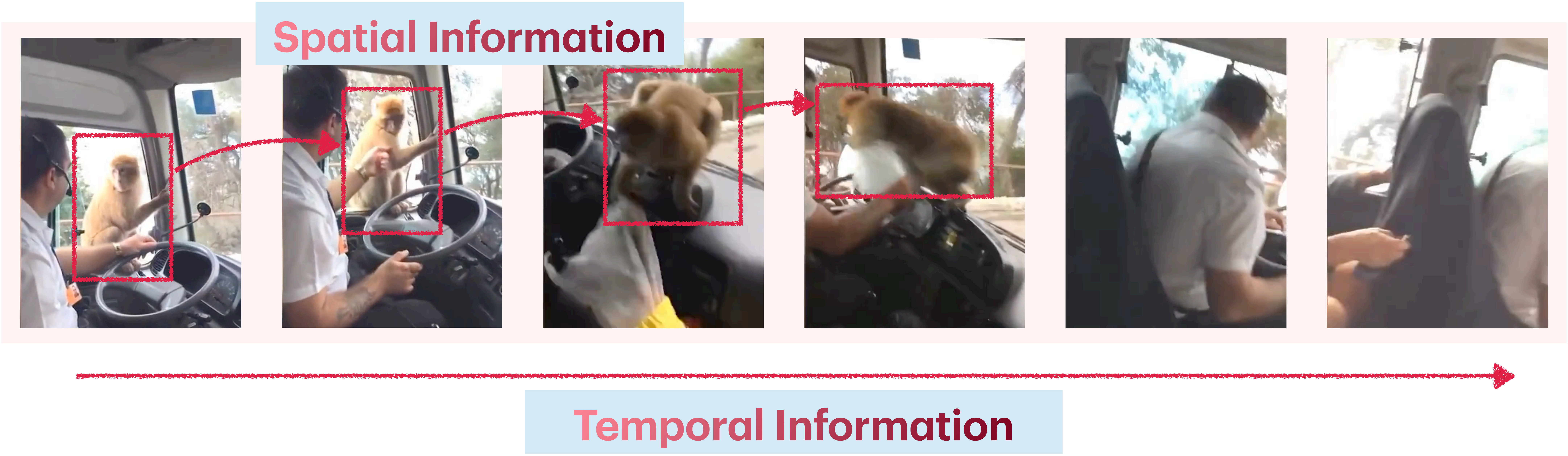
Video Models

From Image to Video

Videos → Collection of **frames** that are **sequentially** inter-related to each other

Video Models

From Image to Video



Spatial Information

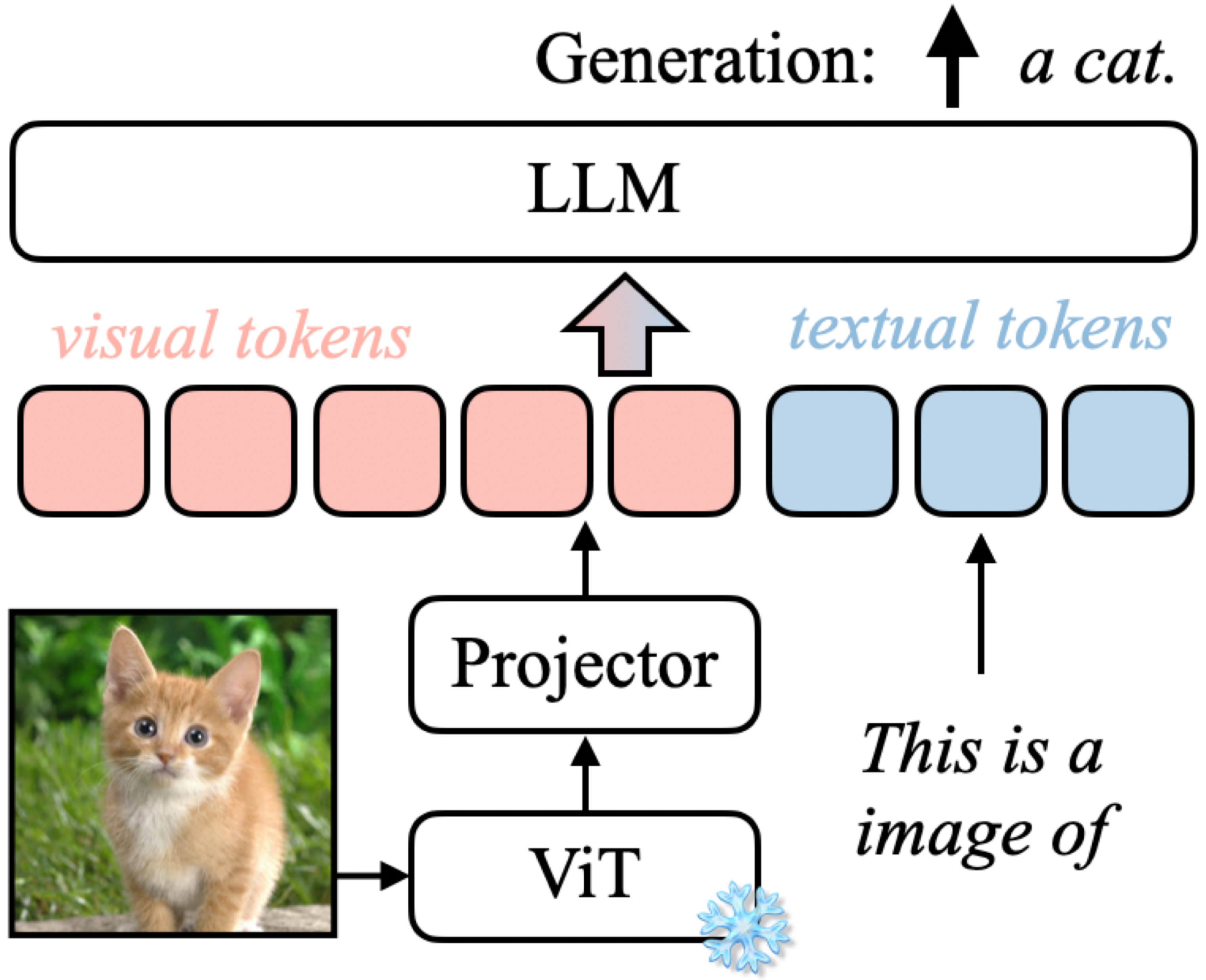
Temporal Information

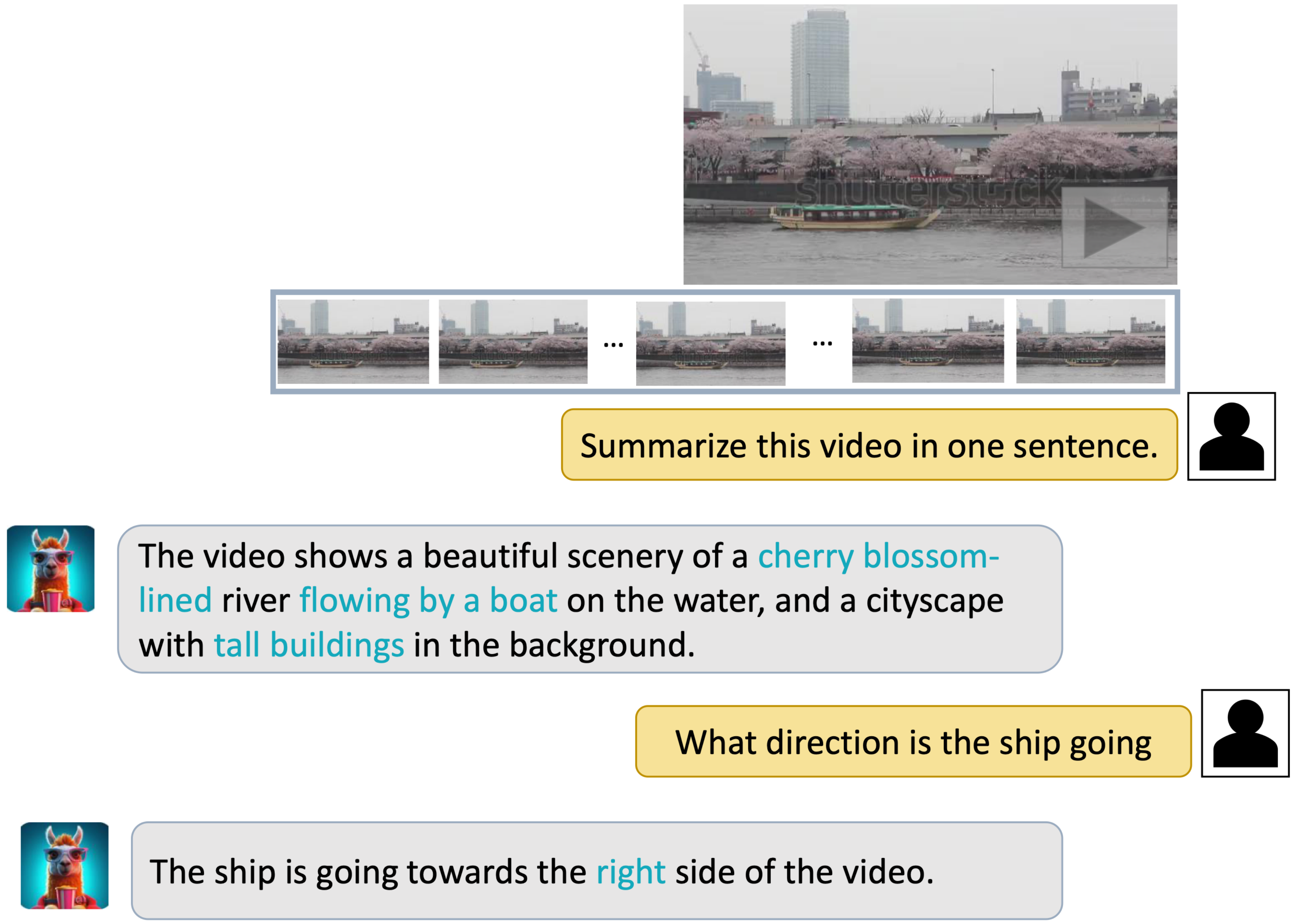
Videos → Collection of **frames** that are **sequentially** inter-related to each other

Challenge: Models must be trained to understand spatiotemporal relationships between frames

Multimodal LLMs

Projecting Visual Inputs to the Text space




Summarize this video in one sentence.

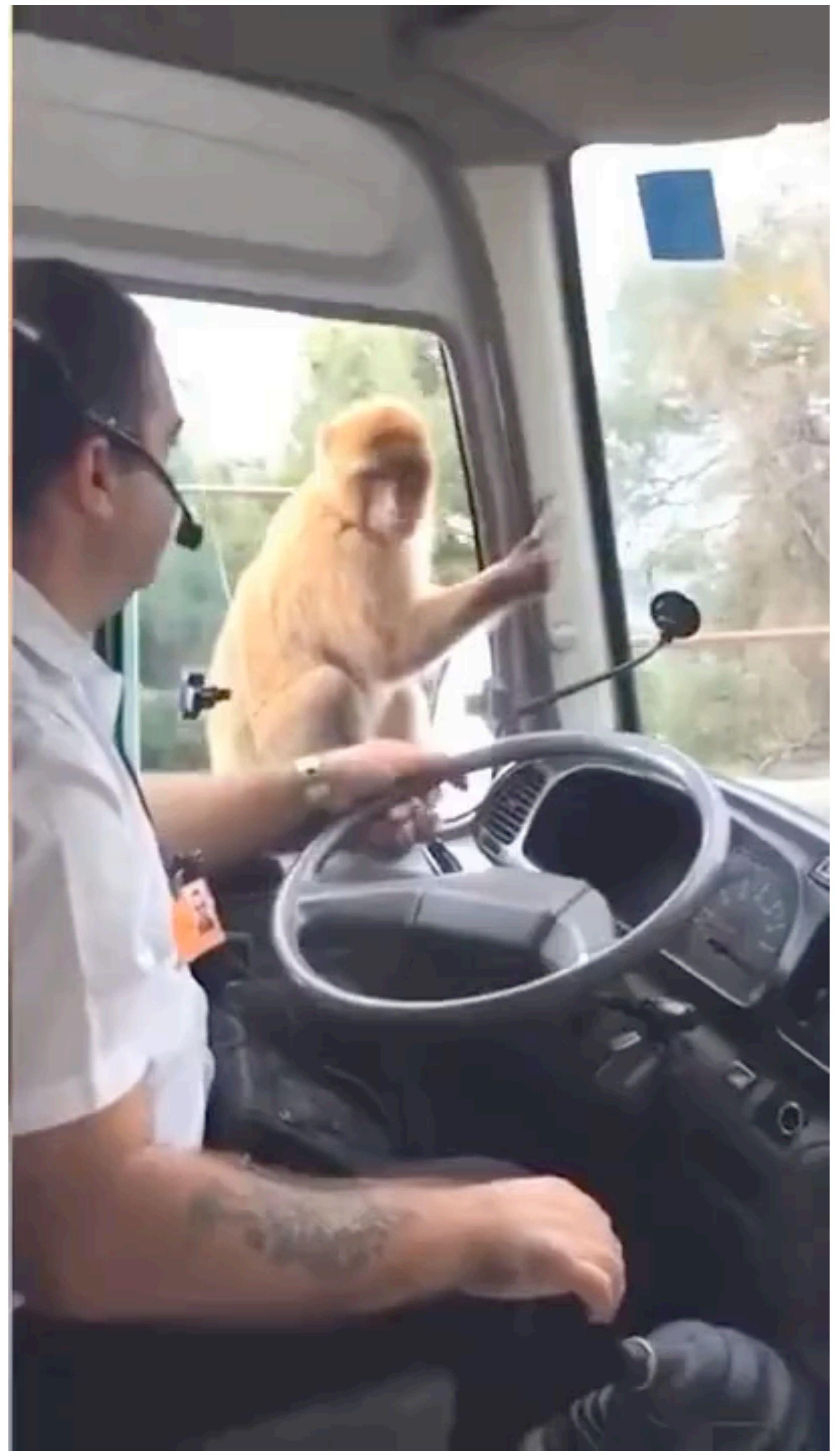
The video shows a beautiful scenery of a cherry blossom-lined river flowing by a boat on the water, and a cityscape with tall buildings in the background.

What direction is the ship going

The ship is going towards the right side of the video.

VILA: On Pre-training for Visual Language Models (CVPR 2024)
Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding (EMNLP 2023 Demo)

How well do VLMs reason about unpredictable events?



GPT 4o:

A monkey rides inside a vehicle with a driver, **explores the dashboard**, and eventually hops out of the vehicle.

Llava-Video:

A monkey is **sitting on the dashboard** of a bus and interacting with the driver.

VideoChat2:

A monkey is seen sitting on the driver's lap and **steering the vehicle** while the driver is wearing a headset and appears to be in a state of surprise...

Unexpected events grab human attention & push AI models beyond their training data.
We investigate how well do VLMs reason about these critical, novel scenarios in our paper **BlackSwan**.

Black Swan: Abductive and Defeasible Video Reasoning in Unpredictable Events

Aditya Chinchure*, Sahithya Ravi*, Raymond Ng, Vered Shwartz, Boyang Li, Leonid Sigal

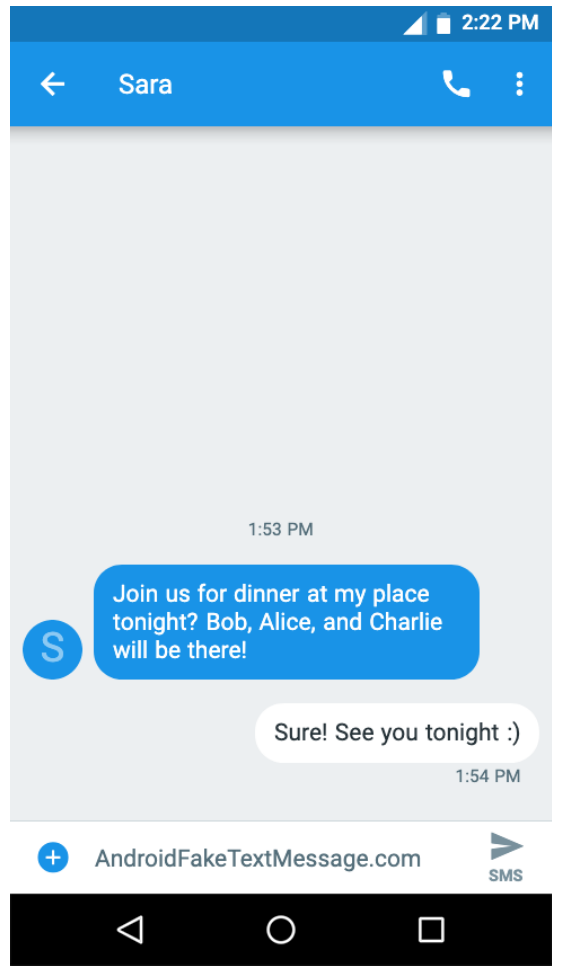
CVPR 2025

Website

UBC VECTOR INSTITUTE | INSTITUT VECTEUR

Abductive Reasoning

Reason about the most plausible explanation for incomplete observations.



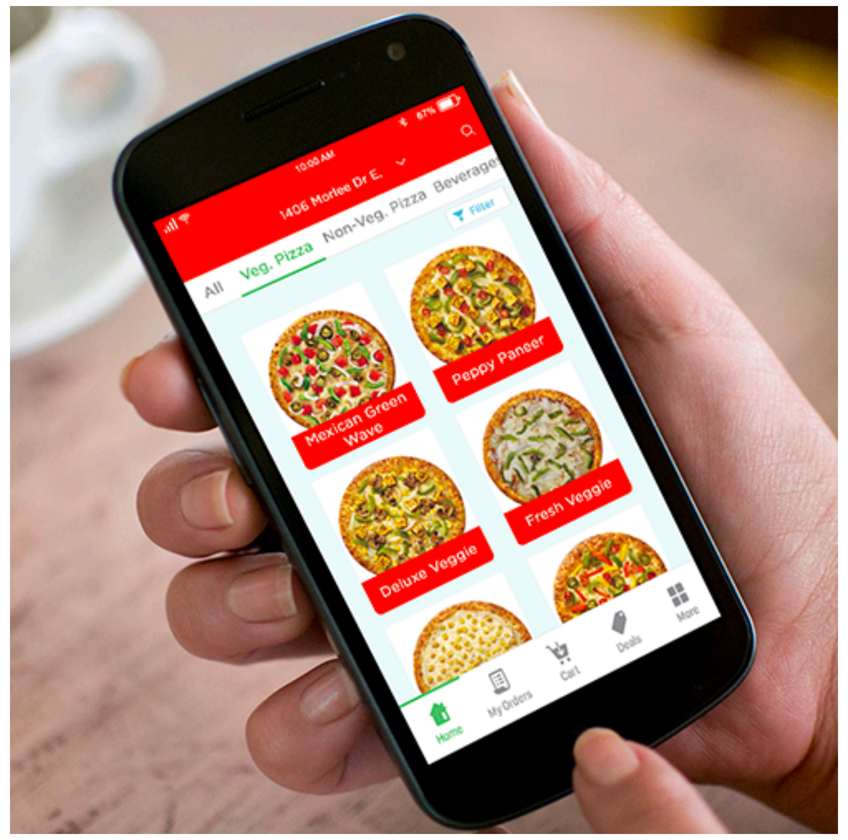
1:53 PM

Join us for dinner at my place tonight? Bob, Alice, and Charlie will be there!

Sure! See you tonight :)

1:54 PM

Sara wanted to make dinner for some guests.

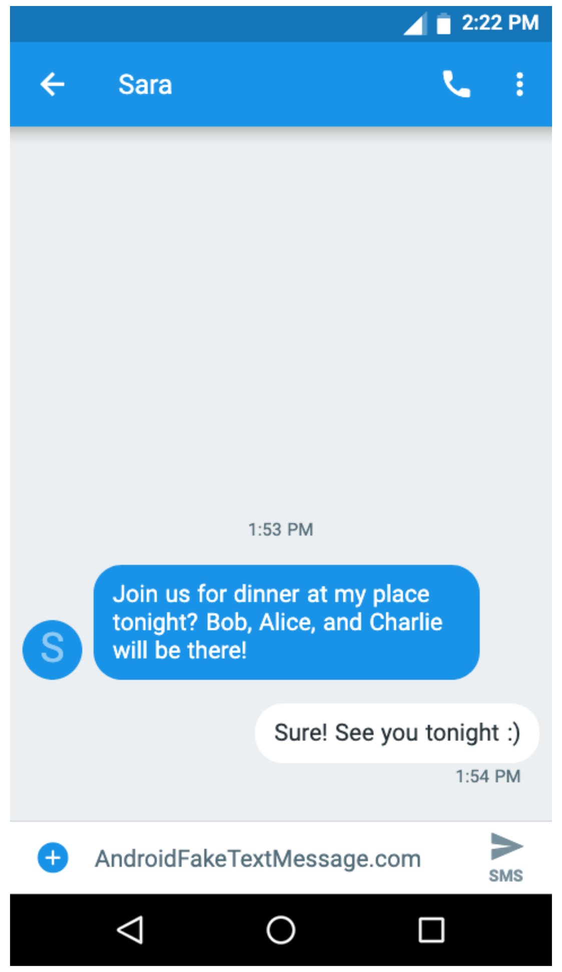


She had to order pizza for her friends instead.

Charles Sanders Peirce. Collected papers of Charles Sanders Peirce, volume 5. Harvard University Press, 1965.

Abductive Reasoning

Reason about the most plausible explanation for incomplete observations.



1:53 PM

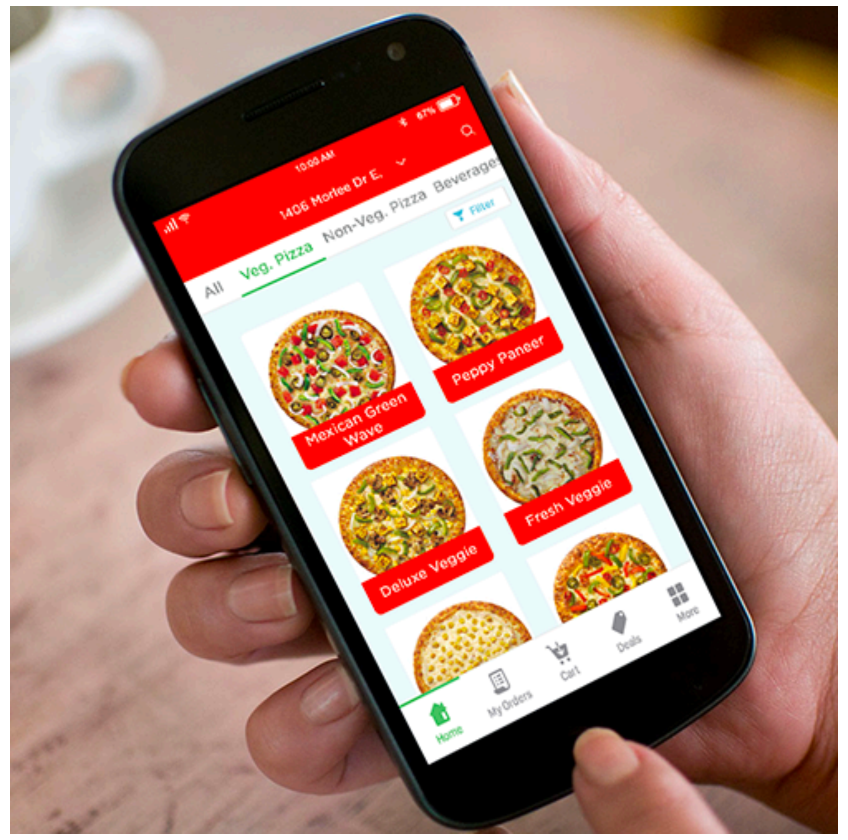
Join us for dinner at my place tonight? Bob, Alice, and Charlie will be there!

Sure! See you tonight :)

1:54 PM

Sara wanted to make dinner for some guests.

But she didn't know how to cook.



She had to order pizza for her friends instead.

Charles Sanders Peirce. Collected papers of Charles Sanders Peirce, volume 5. Harvard University Press, 1965.

Defeasible Inference in Natural Language

An update U is called a **weakener** if, given a premise P and hypothesis H, a human would most likely find H *less likely to be true* after learning U; if they would find H *more likely to be true*, then we call U a **strengthener**.

P: Tweety is a bird.

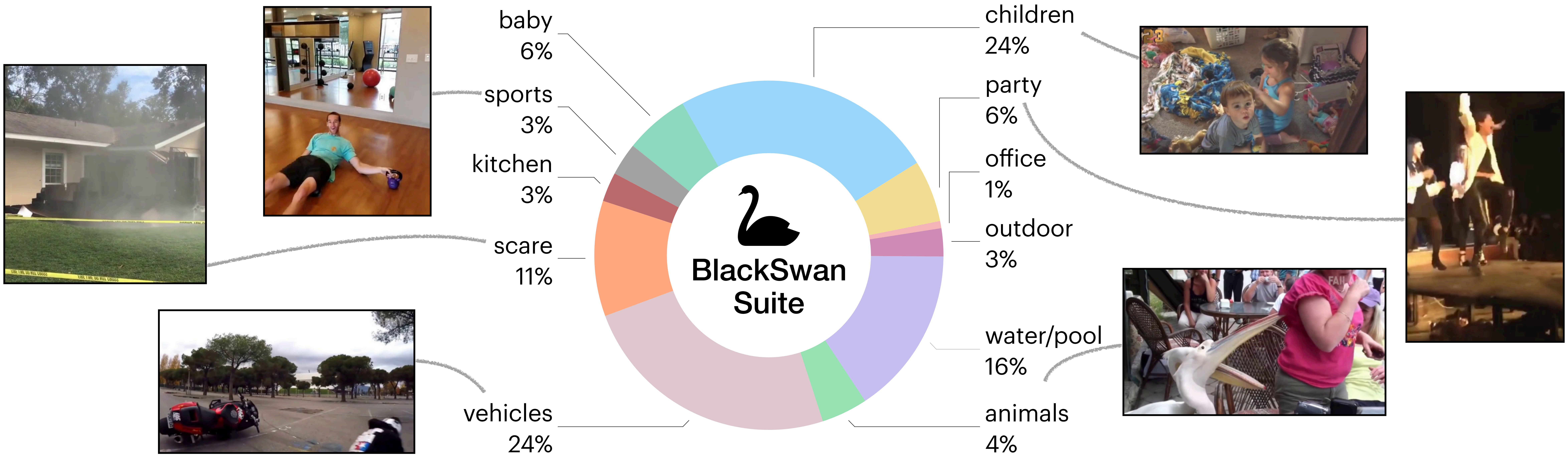
H: Tweety flies.

Weakener: Tweety is a penguin.

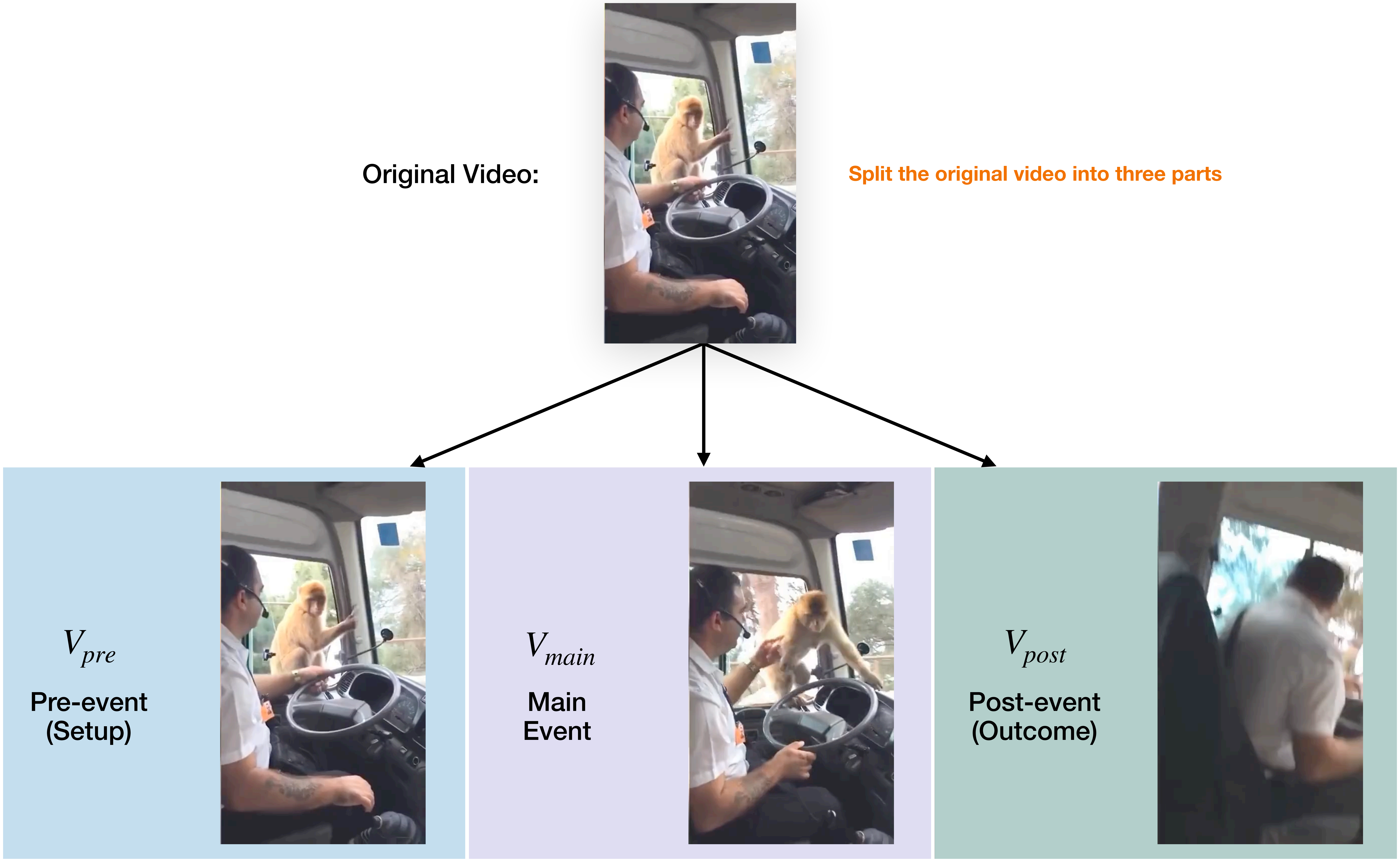
Strengthener: Tweety is on a tree.

Thinking Like a Skeptic: Defeasible Inference in Natural Language. Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Findings of EMNLP 2020.

BlackSwanSuite



We collect 3,800 MCQ, 4,900 generative and 6,700 yes/no tasks, spanning 1,655 videos.



Original Video:

Split the original video into three parts

V_{pre}

Pre-event (Setup)

V_{main}

Main Event

V_{post}

Post-event (Outcome)

V_{pre}
Pre-event (Setup)
V_{main}
Main Event
V_{post}
Post-event (Outcome)
Hidden!
Forecaster
V_{pre}

Given V_{pre}, what could happen next?

🤖/🤔 “The monkey will run across the dashboard and out the other window.”

V_{pre}
Pre-event (Setup)
V_{main}
Main Event
V_{post}
Post-event (Outcome)
Hidden!
Forecaster
V_{pre}
Detective
V_{pre}
V_{post}
Validate
← New Answer

Given V_{pre}, what could happen next?

🤖/🤔 “The monkey will run across the dashboard and out the other window.”

Given V_{pre} and V_{post}, what could happen in the middle?

❌ “The monkey will run across the dashboard and out the other window.”

🤖/🤔 “The monkey will slap the driver in his face.”

V_{pre}
Pre-event (Setup)
V_{main}
Main Event
Hidden!
V_{post}
Post-event (Outcome)
Forecaster
Detective
Reporter
V_{pre}
V_{main}
V_{post}
Validate
New Answer

Given V_{pre}, what could happen next?

🤖/🤔 "The monkey will run across the dashboard and out the other window."

Given V_{pre} and V_{post}, what could happen in the middle?

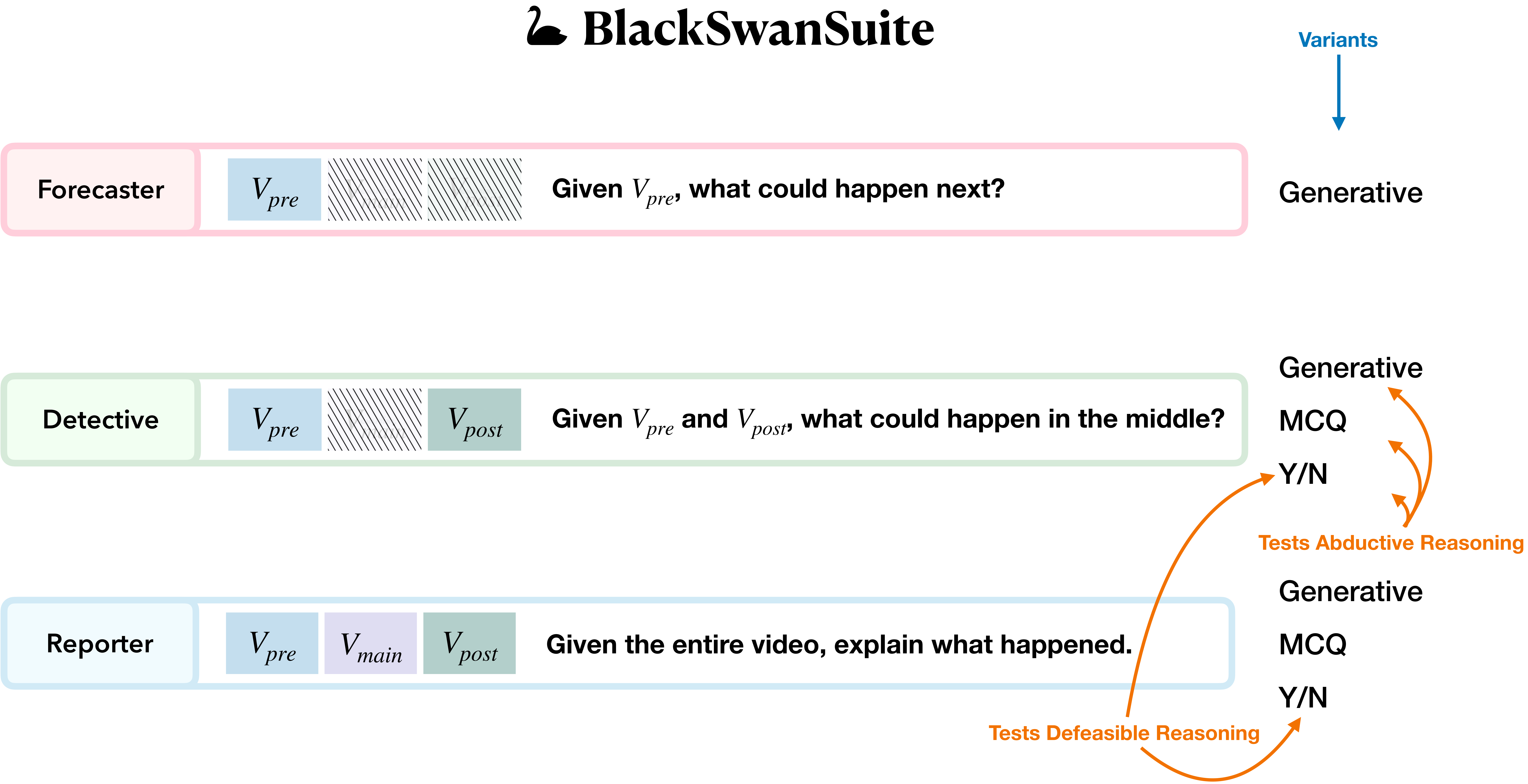
❌ "The monkey will run across the dashboard and out the other window."

🤖/🤔 "The monkey will slap the driver in his face."

Given the entire video, explain what happened.

❌ "The monkey will slap the driver in his face."

🤖/🤔 "The monkey enters the vehicle, steals a bag, and then run out."



BlackSwanSuite

Variants

Forecaster

V_{pre}

Given V_{pre}, what could happen next?

Generative

Detective

V_{pre} V_{post}

Given V_{pre} and V_{post}, what could happen in the middle?

Generative

MCQ

Y/N

Tests Abductive Reasoning

Reporter

V_{pre} V_{main} V_{post}

Given the entire video, explain what happened.

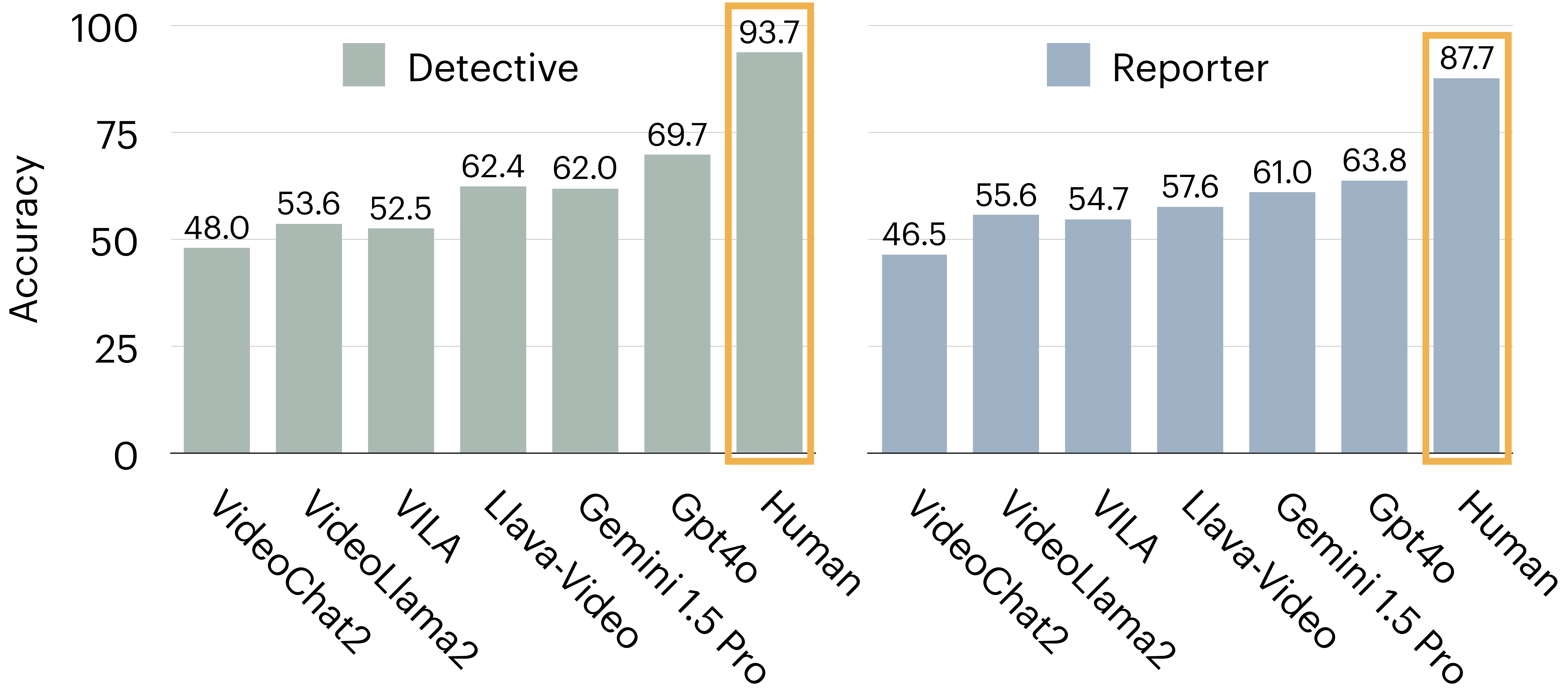
Generative

MCQ

Y/N

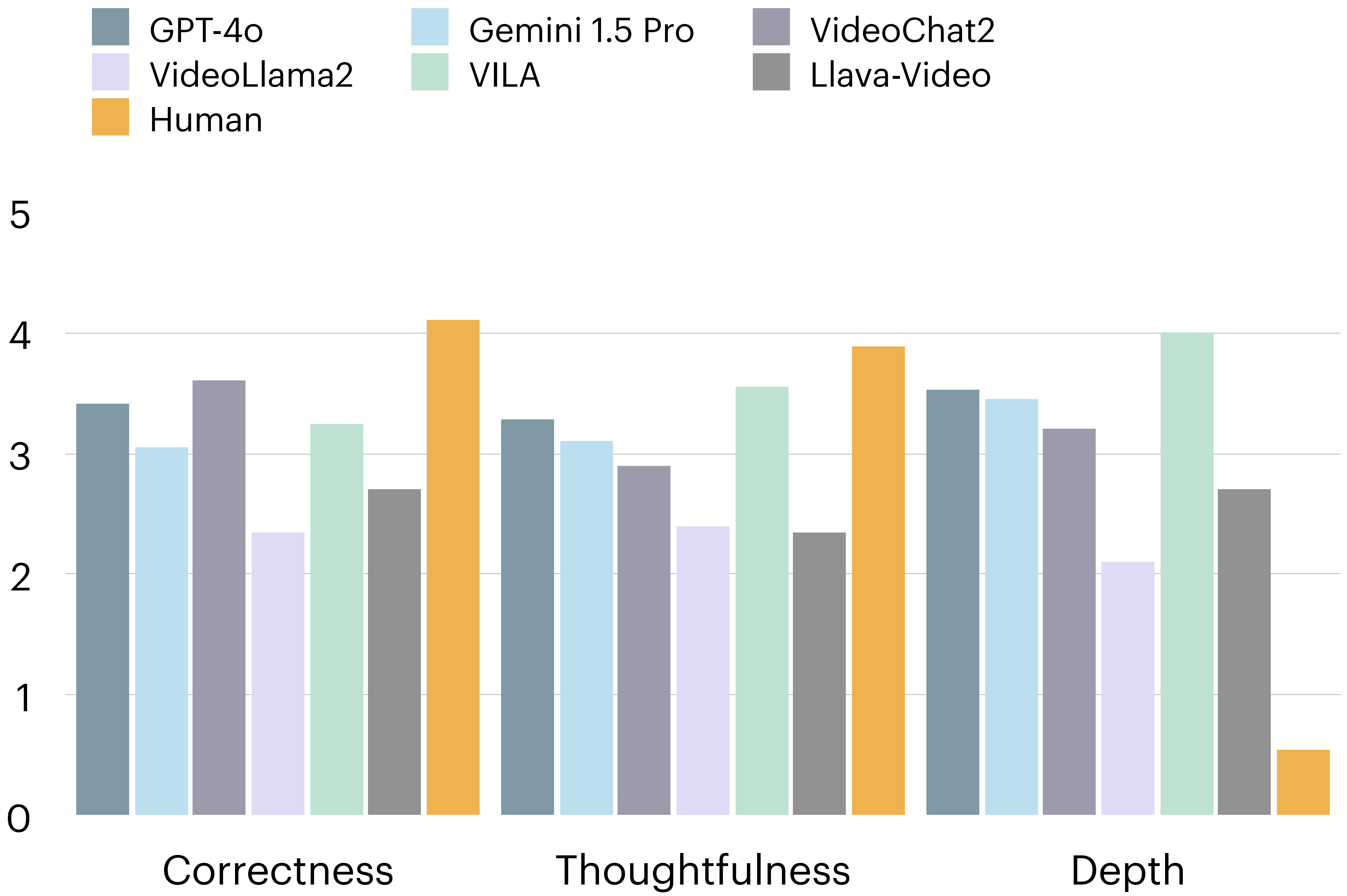
Tests Defeasible Reasoning

Quantitative Results: Video-LLMs on MCQ & Y/N Questions



Models lag behind humans by **~25-30%** in accuracy

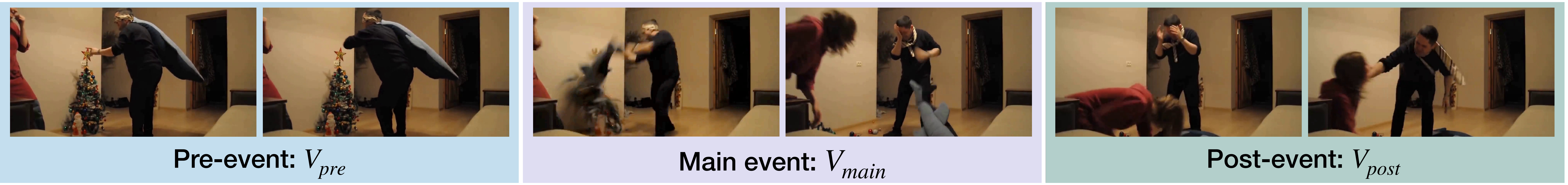
Quantitive Results: Video LLMs on Generative Questions



Models lag behind on correctness & thoughtfulness, but provide (unnecessary) depth

Qualitative Examples

The man swings a pillow and hits the Christmas tree. The tree falls down. Ornaments from the tree fall on the woman, and the man checks on her.

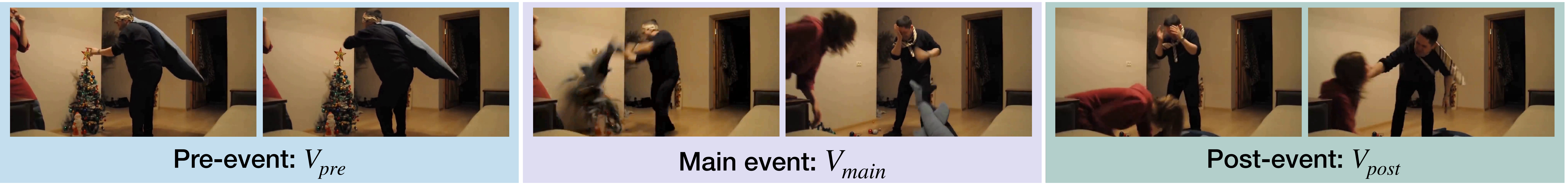


Pre-event: V_{pre}

Main event: V_{main}

Post-event: V_{post}

Qualitative Examples



Pre-event: V_{pre}

Main event: V_{main}

Post-event: V_{post}

Sample evaluation tasks for the above video:

Detective—MCQ:
Given: V_{pre} **&** V_{post} **,** V_{main} **hidden**
What happened in between?

A. The man swings the object and twists around, causing himself to fall to the ground

B. The man swings the object and hits the other person in the visual, as well as the Christmas tree.

C. The man will stand in a room with a Christmas tree while wearing a cape.

Ground Truth: B
Predicted: A — all models incorrect

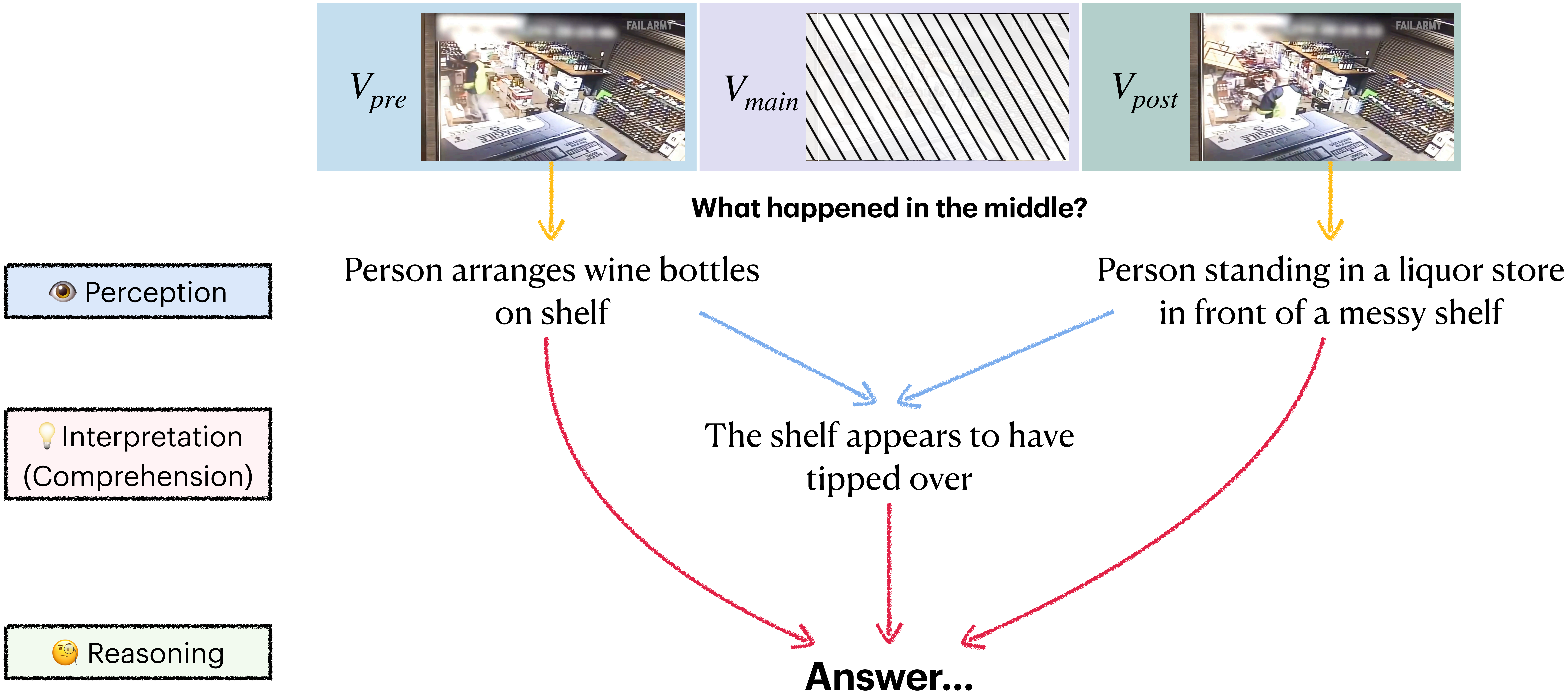
The models miss the fact that **the man does not fall to the ground** in the outcome of the video

Reporter—Y/N: **Given** V_{pre}, V_{main}, V_{post}
Validate the Hypothesis: "The man swings the object and hits the other person in the visual as well as the Christmas tree."

Ground Truth: "No"
❌ **Predicted "Yes":** All models, all are incorrect

The main event shows that the man hits the Christmas tree **but *not* the woman.**

What happens when humans assist with Perception & Comprehension?



What happened in the middle?

👁 Perception

💡 Interpretation (Comprehension)

🧐 Reasoning

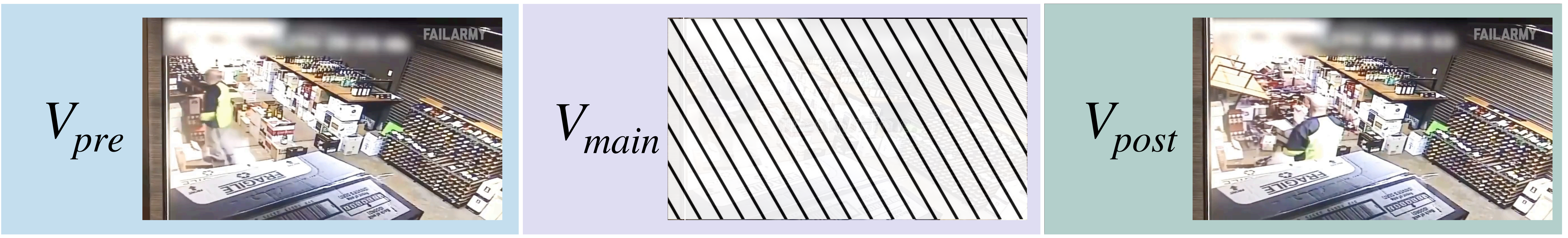
Person arranges wine bottles on shelf

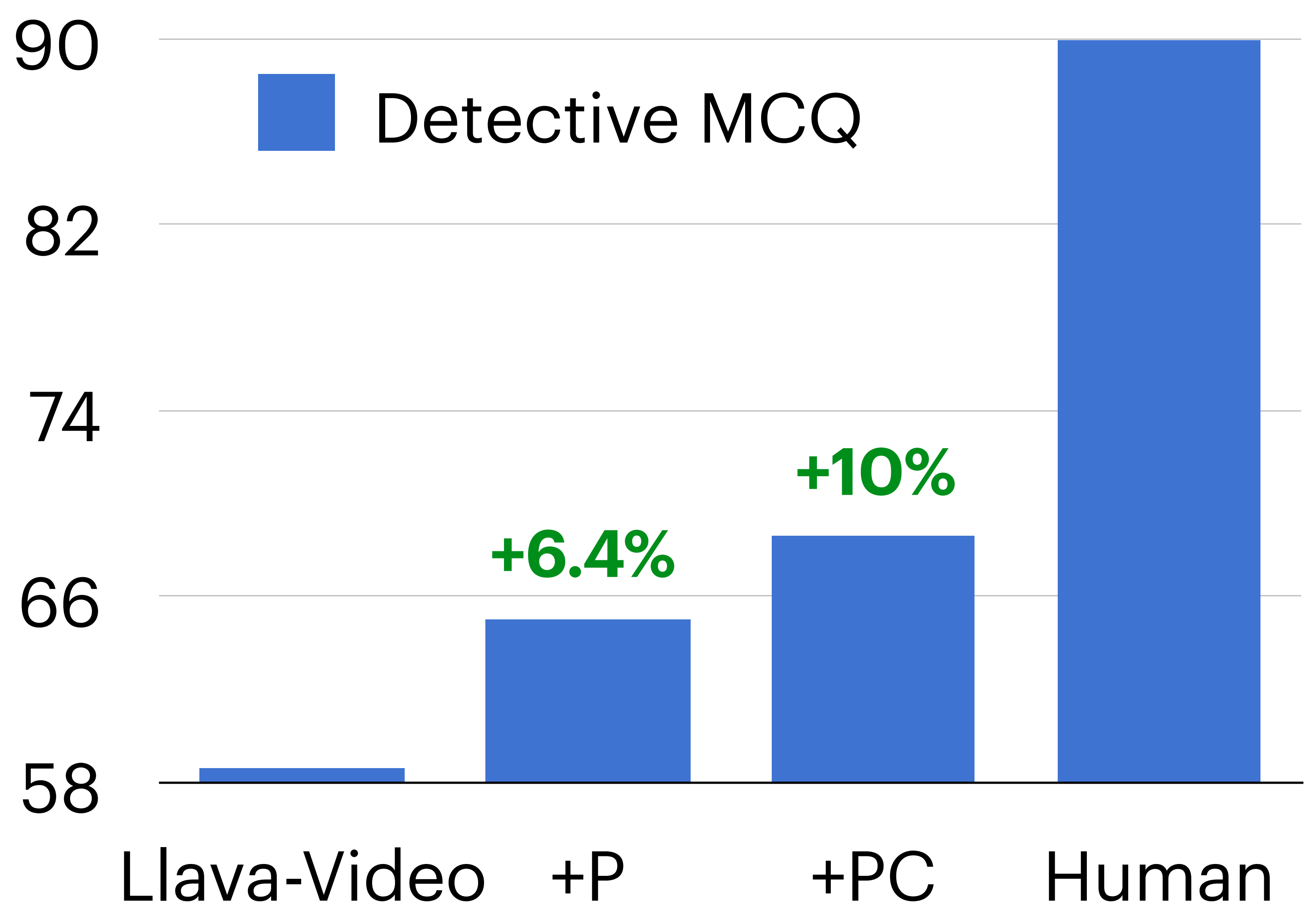
Person standing in a liquor store in front of a messy shelf

The shelf appears to have tipped over

Answer...

What happens when humans assist with Perception & Comprehension?





A. As the guy carries the box of wine bottles, he begins to slip around while still carrying them.

B. The guy throws the box of wine bottles in the air out of frustration and lets the bottles crash onto the floor all around him.

C. As the man removes a box of wine bottles from the table, the table starts to wobble, causing the other boxes still on the table to start falling to the floor.

Perception:

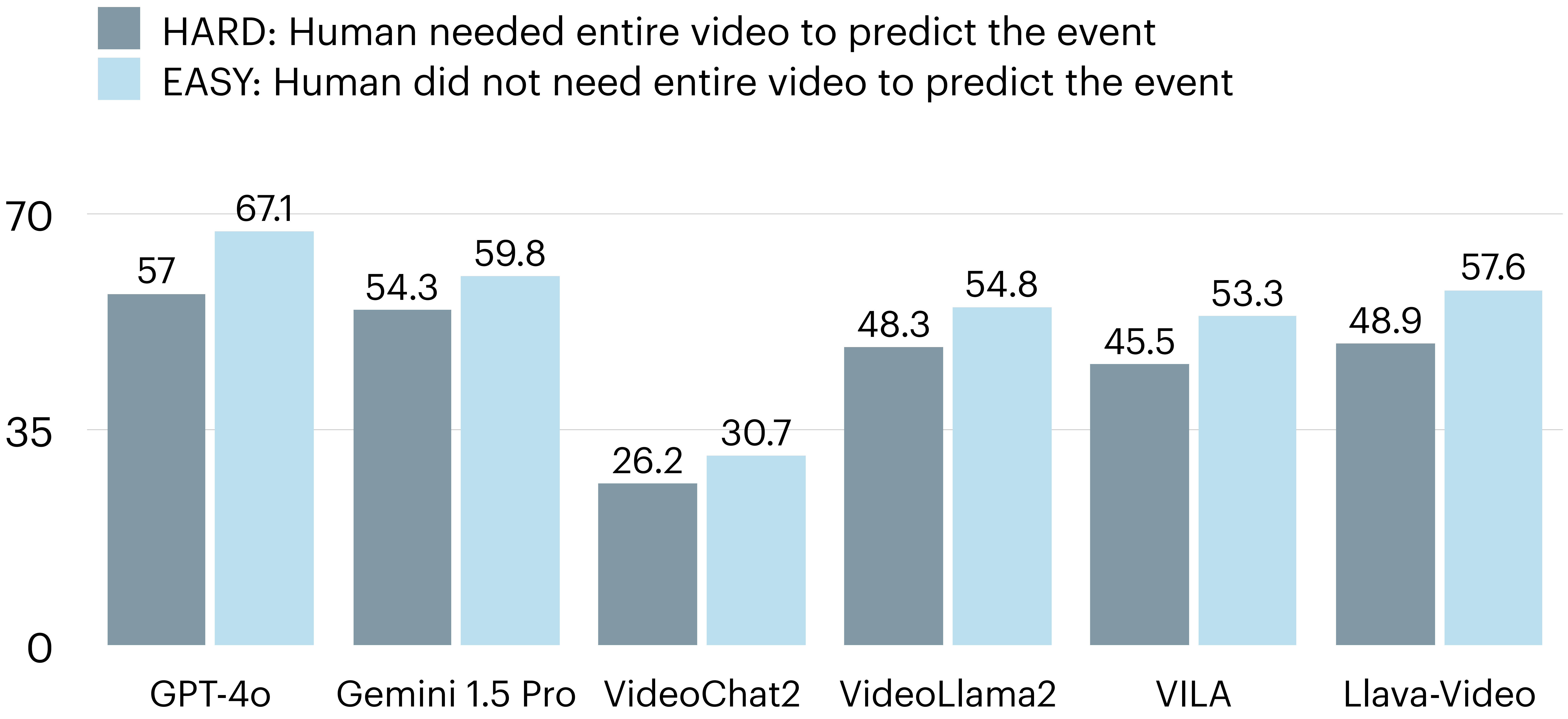
V_{pre}: A man is removing a box of wine bottles from a shelf in a liquor storage area or liquor store. The area is closed up and presumably not open to the public or not a retail store.

V_{post}: A man is standing with his back to the camera. Surrounding him are many shelves and boxes with what appear to be wine and liquor bores. Directly behind the man is a box labeled "Fragile".

Comprehension: In the beginning, a bald man wearing tan pants, a black shirt, and a yellow vest appears to be taking boxes off a shelf on the left-side wall of a warehouse or brewery. In end, the man is seen facing away from the camera looking at the shelf he originally took the box from. The shelf appears to have tipped, as it's leaning sideways and its contents are all over the floor.

GT Ans: C **Baseline:** B ❌ | **+Perception:** B ❌ | **+Perception+Comprehension:** C ✅

Influence of Predictability on Performance



Up to **10%** drop in performance on the hard subset compared to the easy subset, suggesting models may struggle more when events are unpredictable.

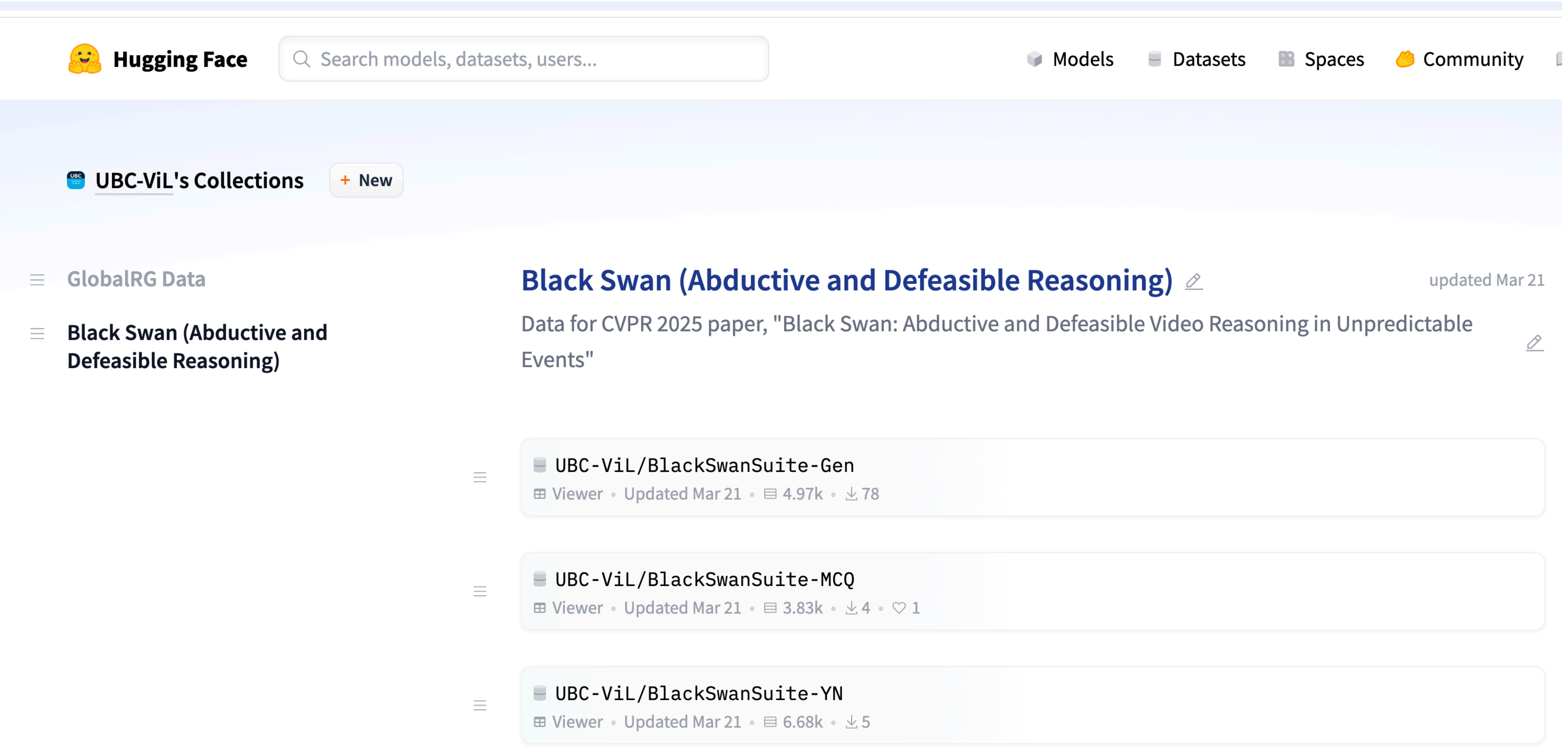
Takeaways: Abilities in Multimodal Models of the Future

2️⃣
Ability to change prior hypotheses as new information is available

1️⃣
Recognize and Adapt to real-world unpredictable scenarios

3️⃣
Deeper perception and comprehension of the visual input

🔗 blackswan.cs.ubc.ca



Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Community
UBC-ViL's Collections
+ New
GlobalRG Data
Black Swan (Abductive and Defeasible Reasoning)

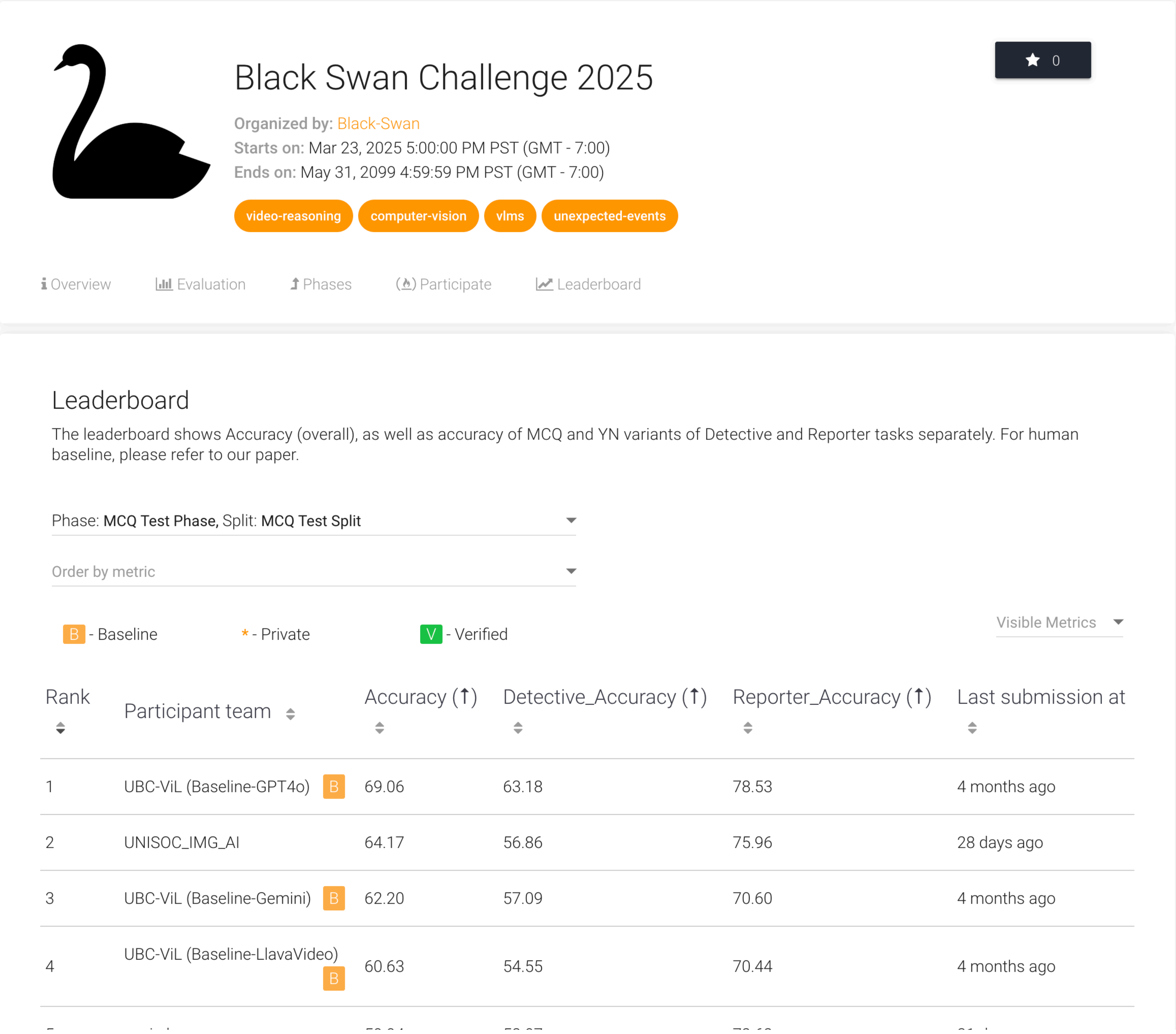
Black Swan (Abductive and Defeasible Reasoning) updated Mar 21

Data for CVPR 2025 paper, "Black Swan: Abductive and Defeasible Video Reasoning in Unpredictable Events"

UBC-ViL/BlackSwanSuite-Gen
Viewer • Updated Mar 21 • 4.97k • 78

UBC-ViL/BlackSwanSuite-MCQ
Viewer • Updated Mar 21 • 3.83k • 4 • 1

UBC-ViL/BlackSwanSuite-YN
Viewer • Updated Mar 21 • 6.68k • 5



Black Swan Challenge 2025

Organized by: Black-Swan
Starts on: Mar 23, 2025 5:00:00 PM PST (GMT - 7:00)
Ends on: May 31, 2099 4:59:59 PM PST (GMT - 7:00)

Leaderboard

The leaderboard shows Accuracy (overall), as well as accuracy of MCQ and YN variants of Detective and Reporter tasks separately. For human baseline, please refer to our paper.

Phase: **MCQ Test Phase,** Split: **MCQ Test Split**

Order by metric

B - Baseline   * - Private   V - Verified

Rank	Participant team	Accuracy (↑)	Detective_Accuracy (↑)	Reporter_Accuracy (↑)	Last submission at
1	UBC-ViL (Baseline-GPT4o) B	69.06	63.18	78.53	4 months ago
2	UNISOC_IMG_AI	64.17	56.86	75.96	28 days ago
3	UBC-ViL (Baseline-Gemini) B	62.20	57.09	70.60	4 months ago
4	UBC-ViL (Baseline-LlavaVideo) B	60.63	54.55	70.44	4 months ago

Validation Set

Hidden Test Set

Thank You!