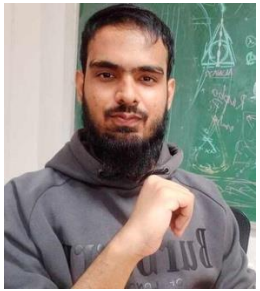
# Towards Neuro-Symbolic Video Understanding

## The 18th European Conference on Computer Vision (ECCV 2024)


Minkyu Choi

Harsh Goel



Mohammad Omama

Yunhao Yang

Sahil Shah

Sandeep Chinchali


# Surge in Video Data Production and its Impact

**500 hours / min**

**500 PB / day**

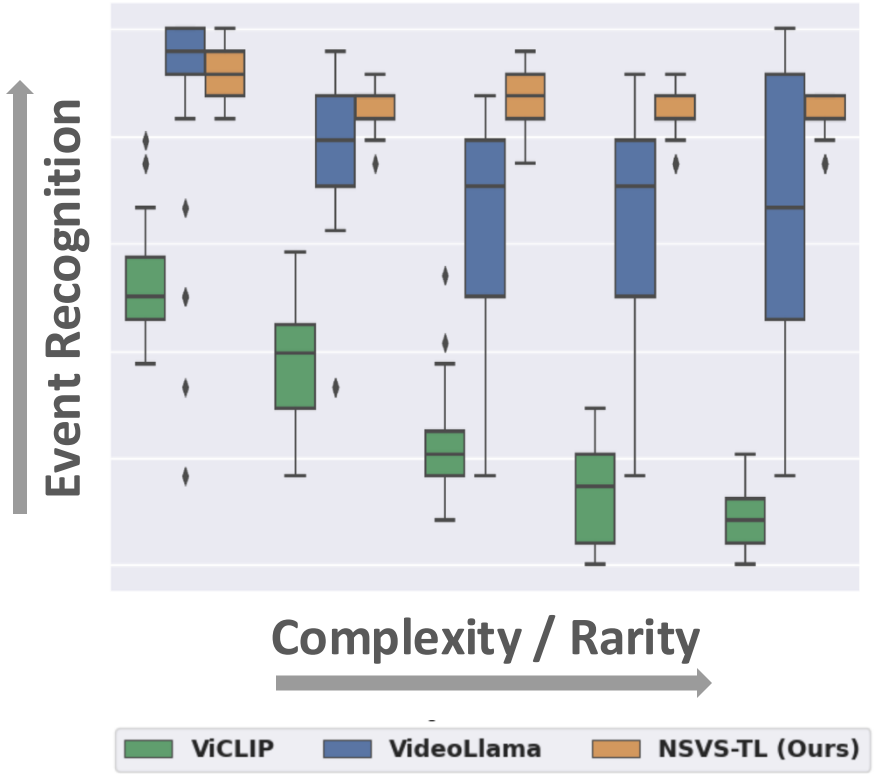

**10 TB / day**

**The computer vision technologies can provide *some levels of perception.***

**Users are asking increasingly complex queries:**

*"Find me all scenes where event A happened, event B did not occur, and event C occurs hours later."*

# Video Foundation Models Fail at Event Recognition when Event is **Rare** or Logically **Complex**




## Why?

- **Our Hypothesis:** Semantics and Reasoning coupled in a single network.
- **Proposed Solution:** Decouple Semantics and Reasoning into a perception module and a **temporal logic** module.

# Reasoning and Understanding over The Formality

$P(\text{Dog}) = 0.40$
$P(\text{Cat}) = 0.60$

$V = F_1 \dots F_n \in \mathbb{R}^{C \times W \times H}$
$P = \{\text{Dog}, \text{Cat}\}$

**Start**

$F_1$ $F_2$ $F_3$ $F_n$

$s_1$ — $P(\text{Dog}) \,\&\, P(\text{Cat}) = 0.24$

$s_2$ — $P(\text{Dog}) \,\&\, NOT\ P(\text{Cat}) = 0.16$

$s_3$ — $NOT\ P(\text{Dog}) \,\&\, P(\text{Cat}) = 0.36$

$s_4$ — $NOT\ P(\text{Dog}) \,\&\, NOT\ P(\text{Cat}) = 0.24$

$s_5$ $s_6$ $s_7$ $s_8$ $s_9$ …

# Reasoning and Understanding over The Formality Cont.

**Temporal Logic Specification**

*"Find me all scenes where event A happened, event B did not occur, and event C occurs hours later."*

$$\mathbf{F}(A \wedge \neg B \wedge \mathbf{F}_{[t,t]}(C))$$

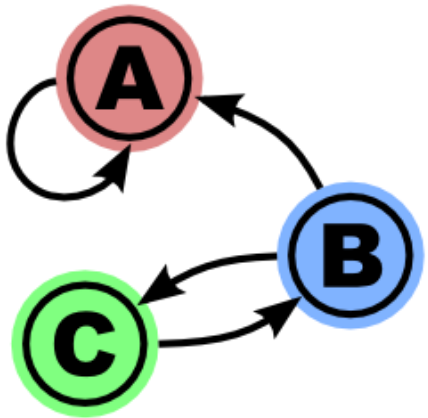
Metric Temporal Logic (MTL)

**Formal Verification**

Satisfy? Not Satisfy?

P(satisfaction)?

**Video Automaton**



# Find the “I’m Flying” Scene from Titanic

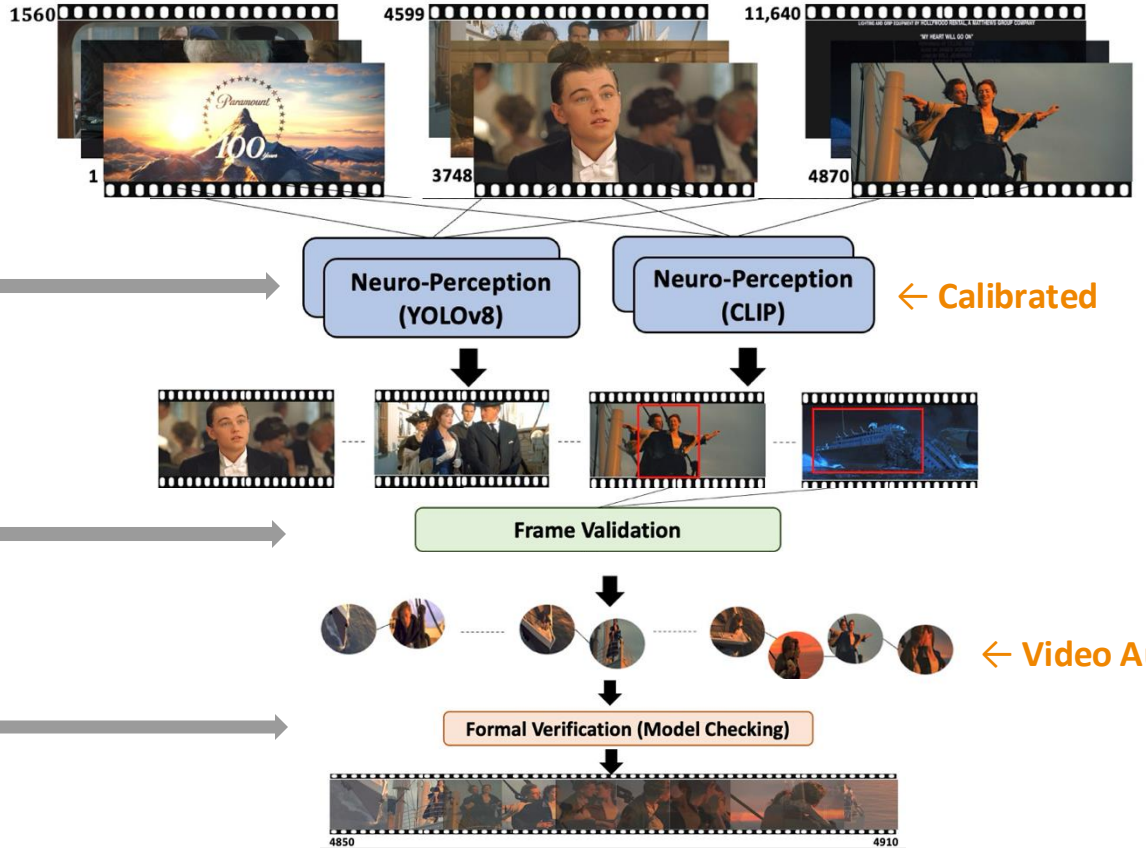
Find the **“I’m Flying”** scene from **Titanic**

## Atomic Propositions

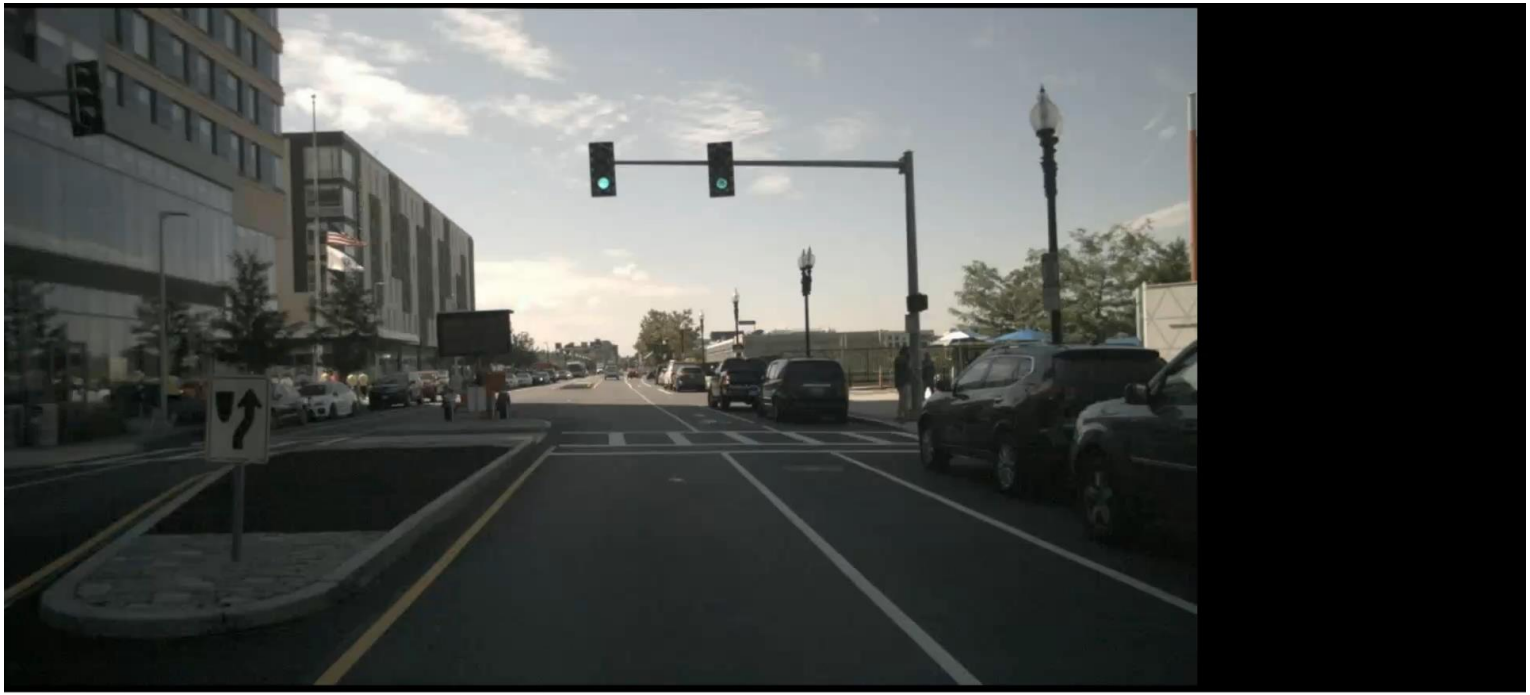
- Man hugging woman
- Ship on the sea
- Kiss

## Symbolic Operations

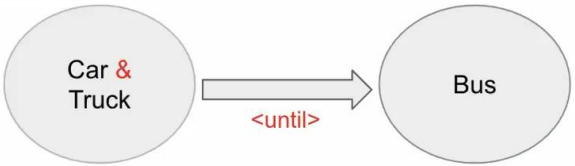
- Temporal Symbol
- Condition Symbol



# Demonstration



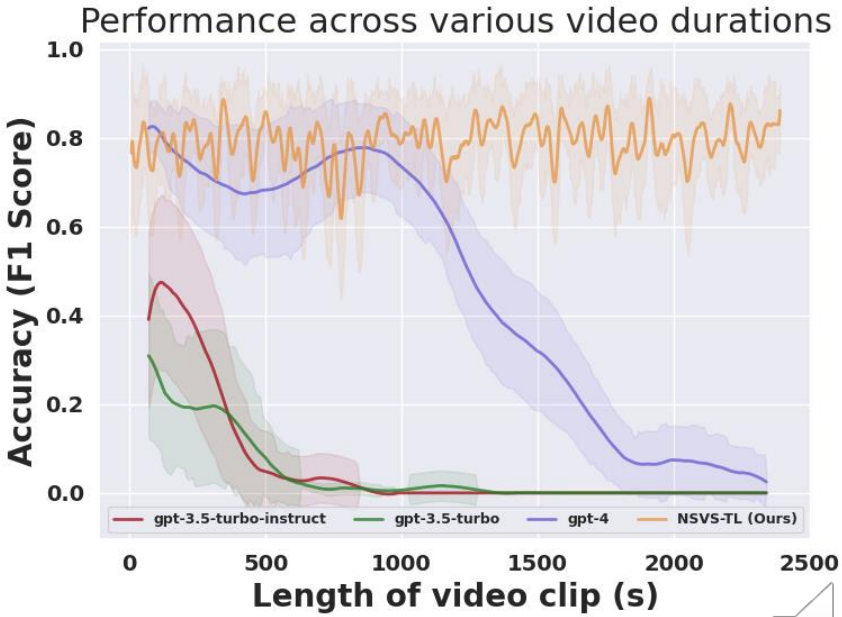
**Specification:**
(Car & Truck) U
Bus



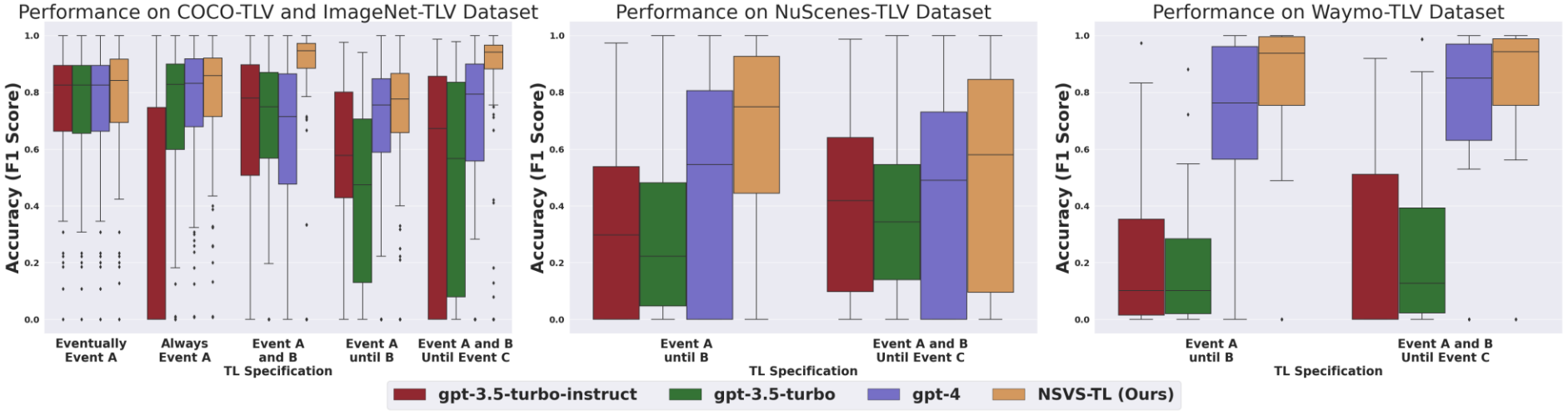
# Frame Retrieval Performance with Increasing Video Length

- Video foundation models are **insufficient** for long term temporal reasoning

- Hence, we design baselines where we pass per-frame annotations to LLMs like ***GPT4*** and ask them to reason on the annotations



# Frame Retrieval Performance for Varying Event Complexity



# Temporal Logic Video (TLV) Dataset

## 1. Synthetic TLV Dataset

$\Phi$ = (person $\wedge$ frisbee) $\mathbb{U}$ boat
Total Number of frames: 25

Data source: COCO
Frames of Interest = [[0,4,6,8,15],[22,23]]

COCO dataset

ImageNet dataset

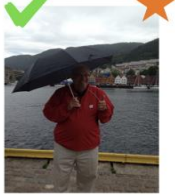
**Frame of Interest Set 1:**


**Frame 0**: Person and frisbee


**Frame 4**: Person and frisbee


**Frame 6**: Person and frisbee


**Frame 8**: Person and frisbee


**Frame 15**: Person and boat

[0,4,6,8,15] satisfies $\Phi$

**Frame of Interest Set 2:**


**Frame 18**: Oven


**Frame 20**: Orange


**Frame 22**: Person and frisbee


**Frame 23**: Boat


**Frame 24**: Dog

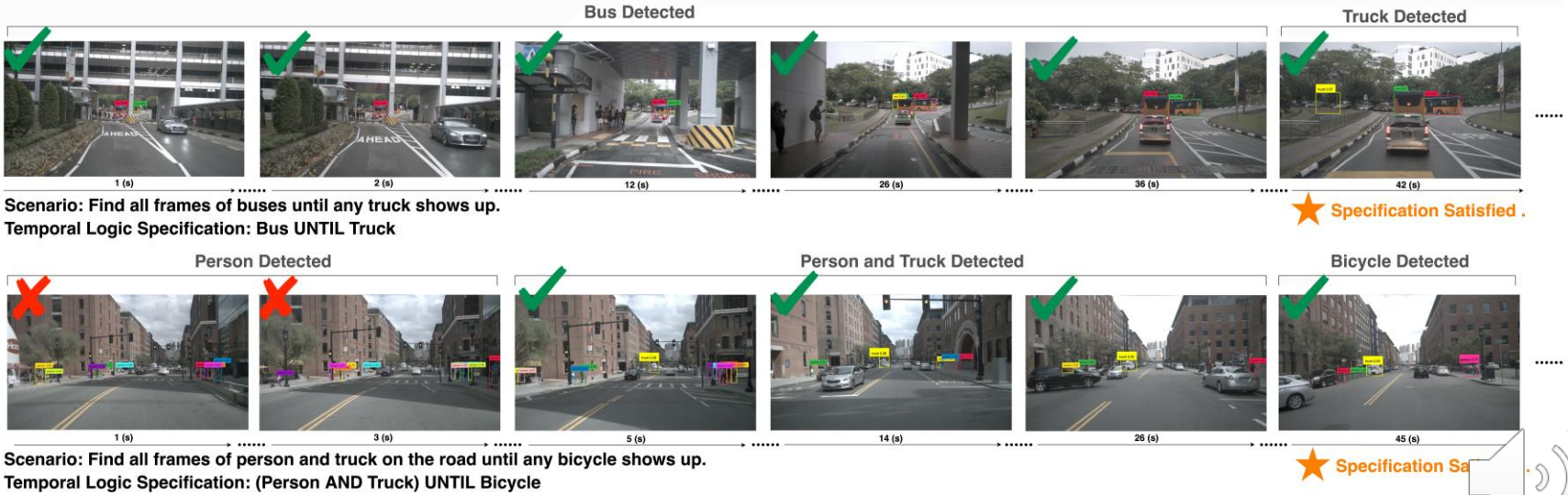
[22,23] satisfies $\Phi$ tlv_synthetic_dataset-source:coco-number_of_frames:25-a21eb.pkl

# Temporal Logic Video (TLV) Dataset Cont.

## 2. Real TLV Dataset

- Annotate temporal logic specification ground truth label.

NuScenes Waymo



# Thank you!

**See you at the poster session (#145)**

**Find Our Work!**

Paper

Code

Dataset