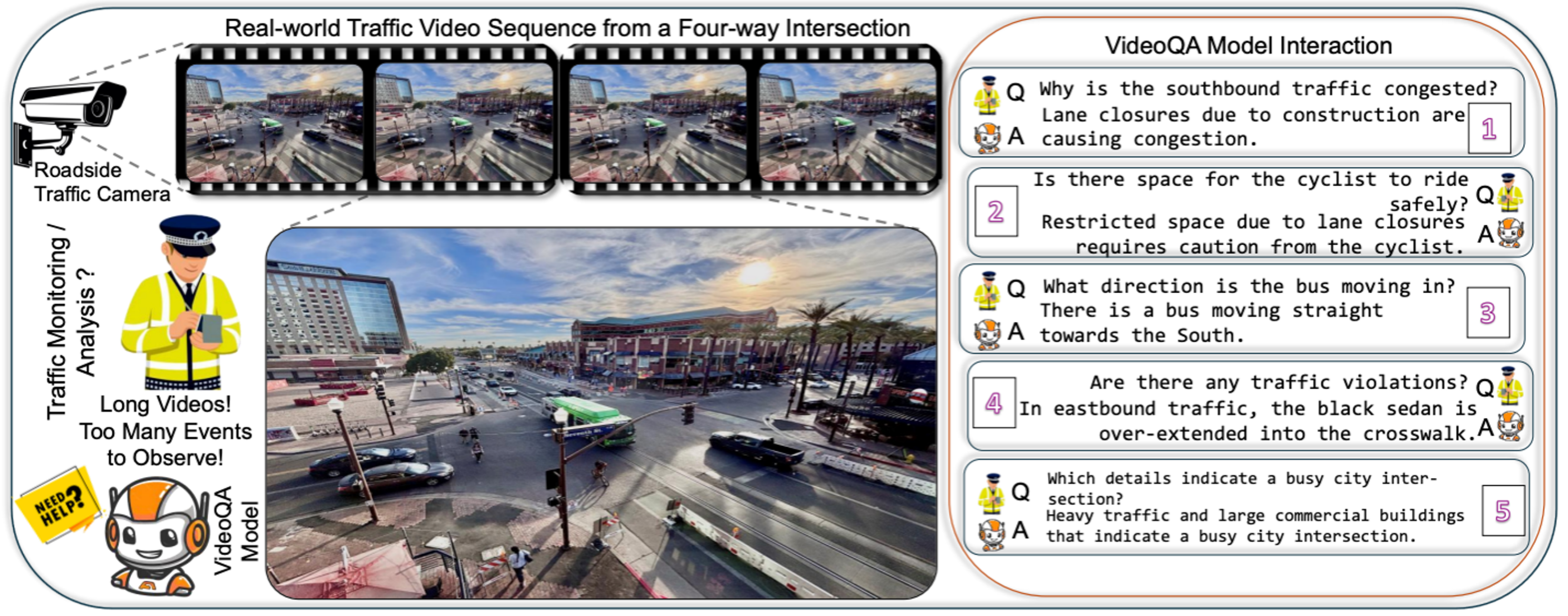
Optimizing Video Question Answering for Traffic Monitoring Systems

Rutuja Patil, BSE Computer Systems Engineering
Bharatesh Chakravarthi,
Assistant Teaching Professor, SCAI

Introduction

Problem Statement: Current models struggle with spatiotemporal understanding, tracking multiple events, and high traffic dynamics, limiting their ability to analyze real-world scenarios. To bridge this gap, the InterAct VideoQA dataset is introduced, providing annotated footage and QA pairs for traffic-specific reasoning.


Real-world Traffic Video Sequence from a Four-way Intersection
Roadside Traffic Camera
Traffic Monitoring / Analysis ?
Long Videos! Too Many Events to Observe!
NEED HELP?
VideoQA Model
VideoQA Model Interaction
1
2
3
4
5

Q Why is the southbound traffic congested?
A Lane closures due to construction are causing congestion.

Q Is there space for the cyclist to ride safely?
A Restricted space due to lane closures requires caution from the cyclist.

Q What direction is the bus moving in?
A There is a bus moving straight towards the South.

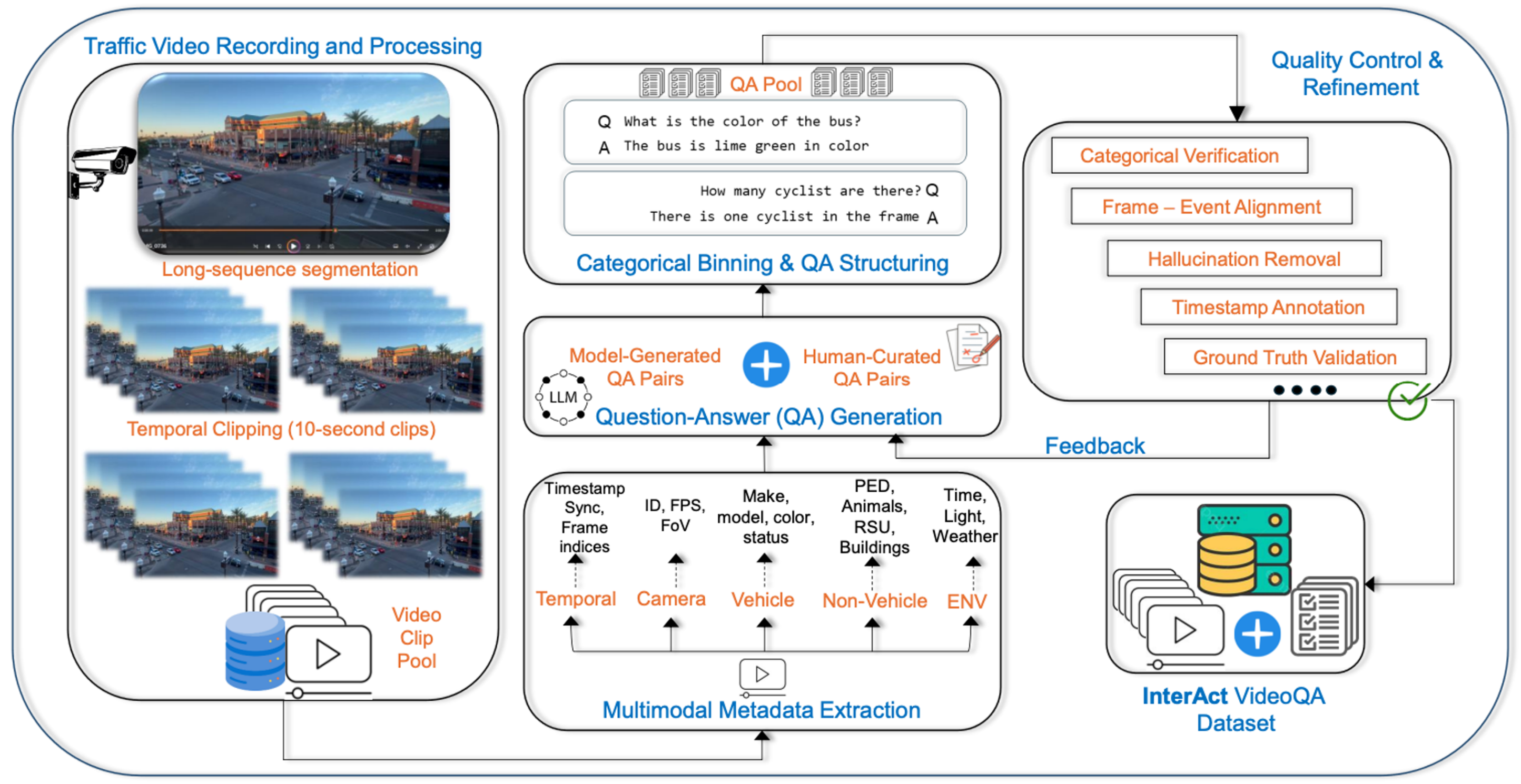
Q Are there any traffic violations?
A In eastbound traffic, the black sedan is over-extended into the crosswalk.

Q Which details indicate a busy city inter-section?
A Heavy traffic and large commercial buildings that indicate a busy city intersection.

Contributions

- ❑ **Comprehensive Data Collection:** 8 hours of real-world traffic footage, segmented into 10-second clips with over 25,000 QA pairs covering critical traffic situations.
- ❑ **Evaluation of SOTA Video QA Models:** Revealing challenges in spatio-temporal reasoning.
- ❑ **Fine Tuning and Performance Improvements:** Highlights significant gains in accuracy and interoperability by fine-tuning models for traffic-related tasks.

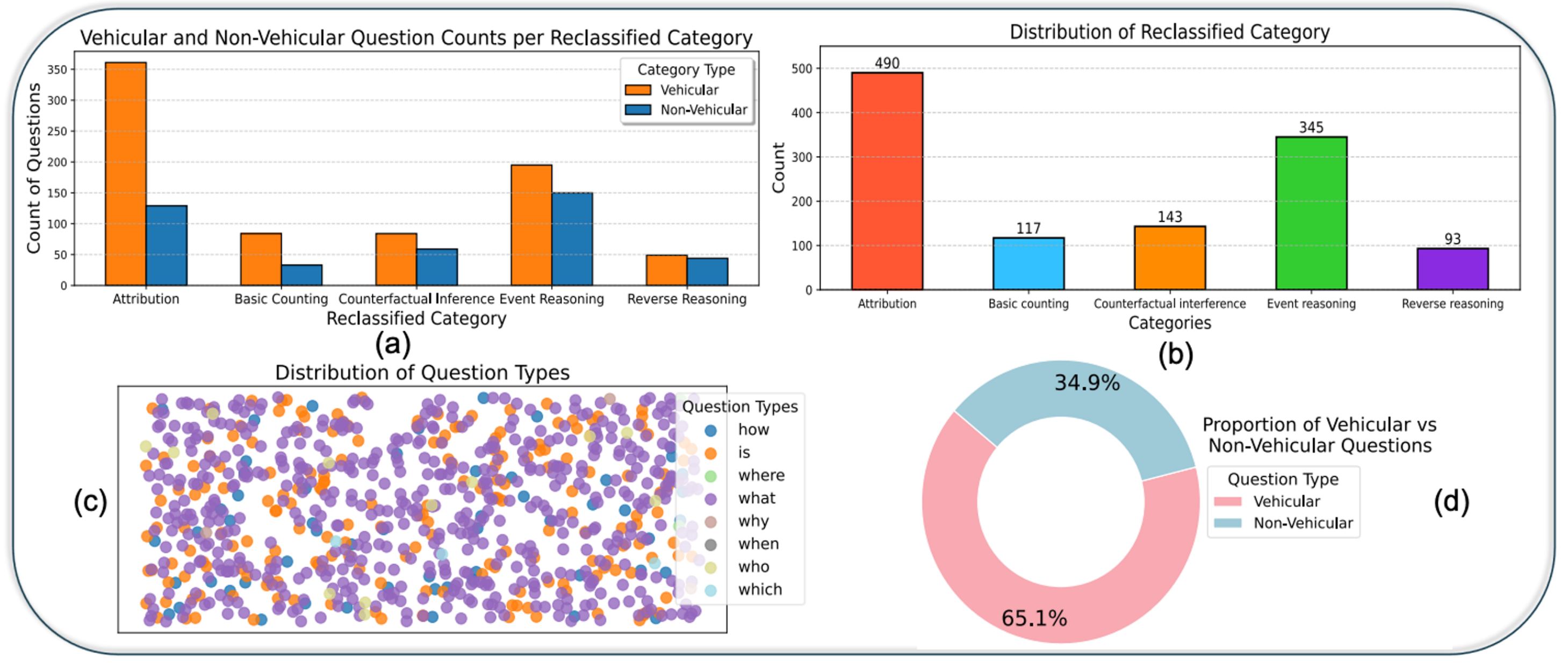
InterAct VideoQA Pipeline


Traffic Video Recording and Processing
Long-sequence segmentation
Temporal Clipping (10-second clips)
Video Clip Pool
QA Pool
Categorical Binning & QA Structuring
LLM
Model-Generated QA Pairs
Human-Curated QA Pairs
Question-Answer (QA) Generation
Timestamp Sync, Frame indices
ID, FPS, FoV
Make, model, color, status
PED, Animals, RSU, Buildings
Time, Light, Weather
Temporal
Camera
Vehicle
Non-Vehicle
ENV
Multimodal Metadata Extraction
Quality Control & Refinement
Categorical Verification
Frame – Event Alignment
Hallucination Removal
Timestamp Annotation
Ground Truth Validation
Feedback
InterAct VideoQA Dataset

Q What is the color of the bus?
A The bus is lime green in color

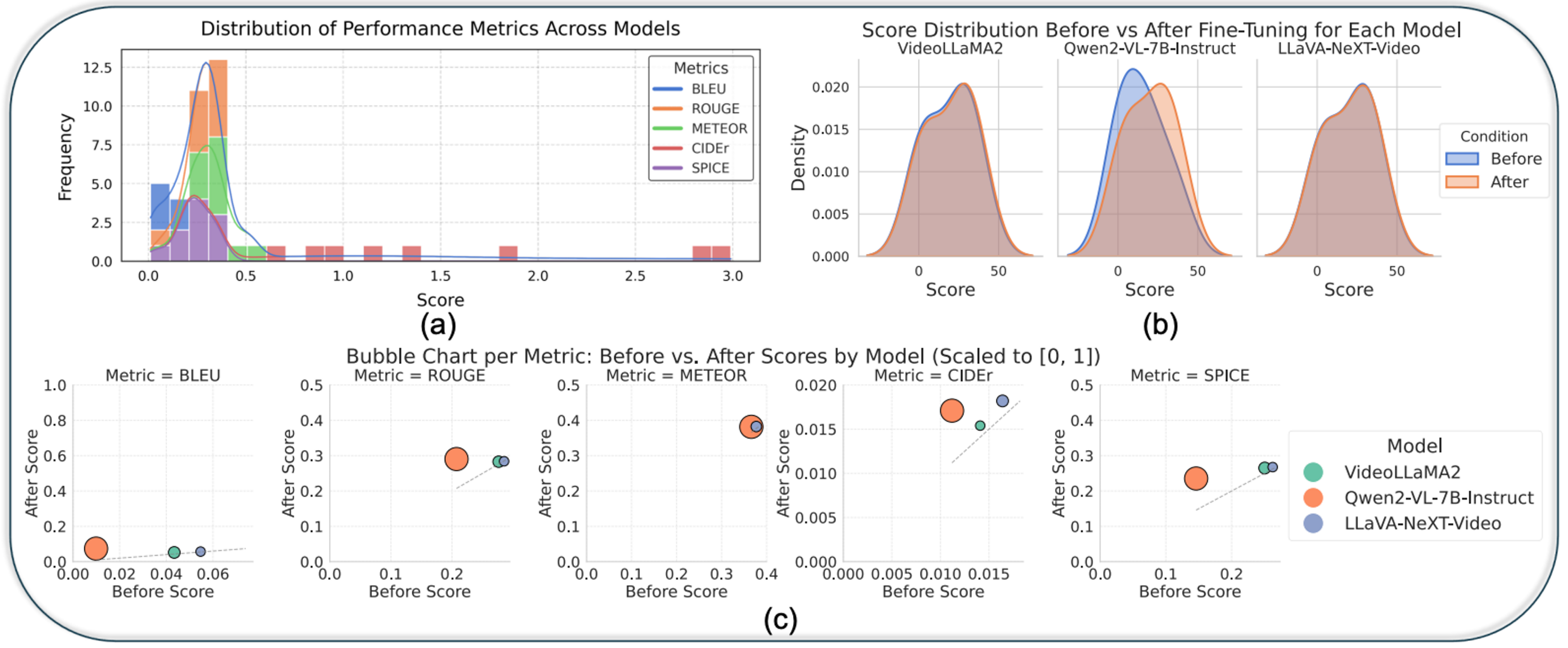
Q How many cyclist are there?
A There is one cyclist in the frame

Question Distribution



InterAct: The dataset features counterfactual, reverse, and event reasoning questions, focusing on spatiotemporal queries and multi-event interactions in traffic scenarios. "What" and "Is" questions dominate, breaking down complex situations into manageable components. This structure helps models interpret overlapping events and distinguish real from hallucinated occurrences.

Quantitative evaluation



Benchmark performance metrics

FineTuning	Questions	VideoLlama2					Llava-NexT-Video					Qwen2-VL-7B-hf				
		BLUE	ROUGE	METEOR	CIDEr	SPICE	BLUE	ROUGE	METEOR	CIDEr	SPICE	BLUE	ROUGE	METEOR	CIDEr	SPICE
After	Basic Counting	-	27.78	29.82	117.52	32.50	-	10.53	22.55	62.04	4.35	-	27.78	29.82	117.52	32.50
	Attribution	1.08	26.01	30.59	84.41	24.25	-	26.27	34.06	91.24	23.83	2.70	27.48	36.09	108.44	27.16
	Event Reasoning	15.15	37.70	51.61	279.44	34.30	15.15	36.14	50.00	298.99	34.40	10.14	34.94	45.45	250.95	31.04
	CounterFactual	-	31.25	31.20	-	18.18	-	31.25	31.20	-	18.18	-	31.25	38.75	-	19.05
	Reverse Reasoning	2.65	24.21	38.22	136.53	23.98	7.39	31.24	39.57	182.07	22.25	16.83	28.44	39.03	253.70	23.82
Before	Basic Counting	-	27.78	29.82	117.20	32.50	-	27.78	29.82	117.20	32.50	-	16.89	37.42	109.07	32.50
	Attribution	0.89	25.66	32.54	87.96	24.40	2.70	27.73	35.59	104.20	24.20	1.67	14.95	47.35	148.11	24.46
	Event Reasoning	-	12.57	47.66	241.75	29.98	10.43	33.68	45.89	235.47	29.98	0.64	12.52	22.05	74.29	8.08
	CounterFactual	-	31.25	31.20	-	18.18	-	31.25	31.20	-	18.18	-	28.00	31.32	-	20.00
	Reverse Reasoning	2.65	23.74	37.33	130.88	22.25	7.39	27.23	37.61	178.02	23.81	0.77	14.95	37.94	130.25	5.88

Conclusion

The study highlights the need for specialized VideoQA datasets like InterAct VideoQA to tackle multi-event traffic challenges. Fine-tuning models significantly improves accuracy in complex scenarios, aiding traffic monitoring and autonomous systems. As an open-source resource, it invites contributions to support long-term intelligent transportation research.

FURI

ASU Ira A. Fulton Schools of Engineering
Arizona State University