Advancing Video Question Answering (VideoQA) through Attribution, Reasoning, and Counterfactual Inference Tasks

Nagasiri Poluri, Computer Science

Mentor: Bharatesh Chakravarthi, Assistant Teaching Professor, SCAI

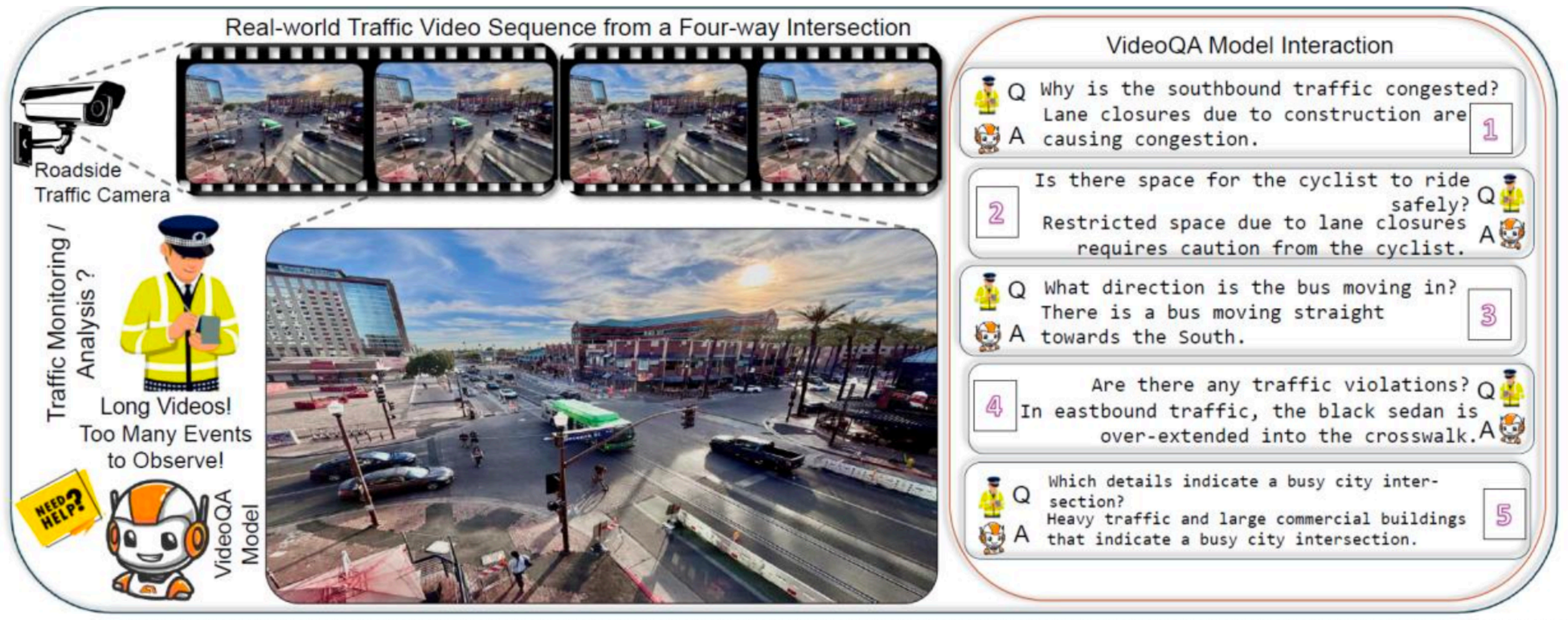
Research Question

How can advanced reasoning capabilities such as attribution, counting, event reasoning, reverse reasoning, and counterfactual inference be effectively modeled and evaluated in VideoQA systems?

Background

State-of-the-art VideoQA models mostly handle surface-level Q&A and struggle with deeper temporal/causal reasoning (order, cause, “what-if” scenarios), limiting reliability in real-world use.

This work targets the gap by defining and evaluating tasks for attribution, counting, event/reverse reasoning, and counterfactual inference on real-world videos. The goal is to make models both more accurate and explainable.


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

Q Why is the southbound traffic congested?
A Lane closures due to construction are causing congestion. 1

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

Q What direction is the bus moving in?
A There is a bus moving straight towards the South. 3

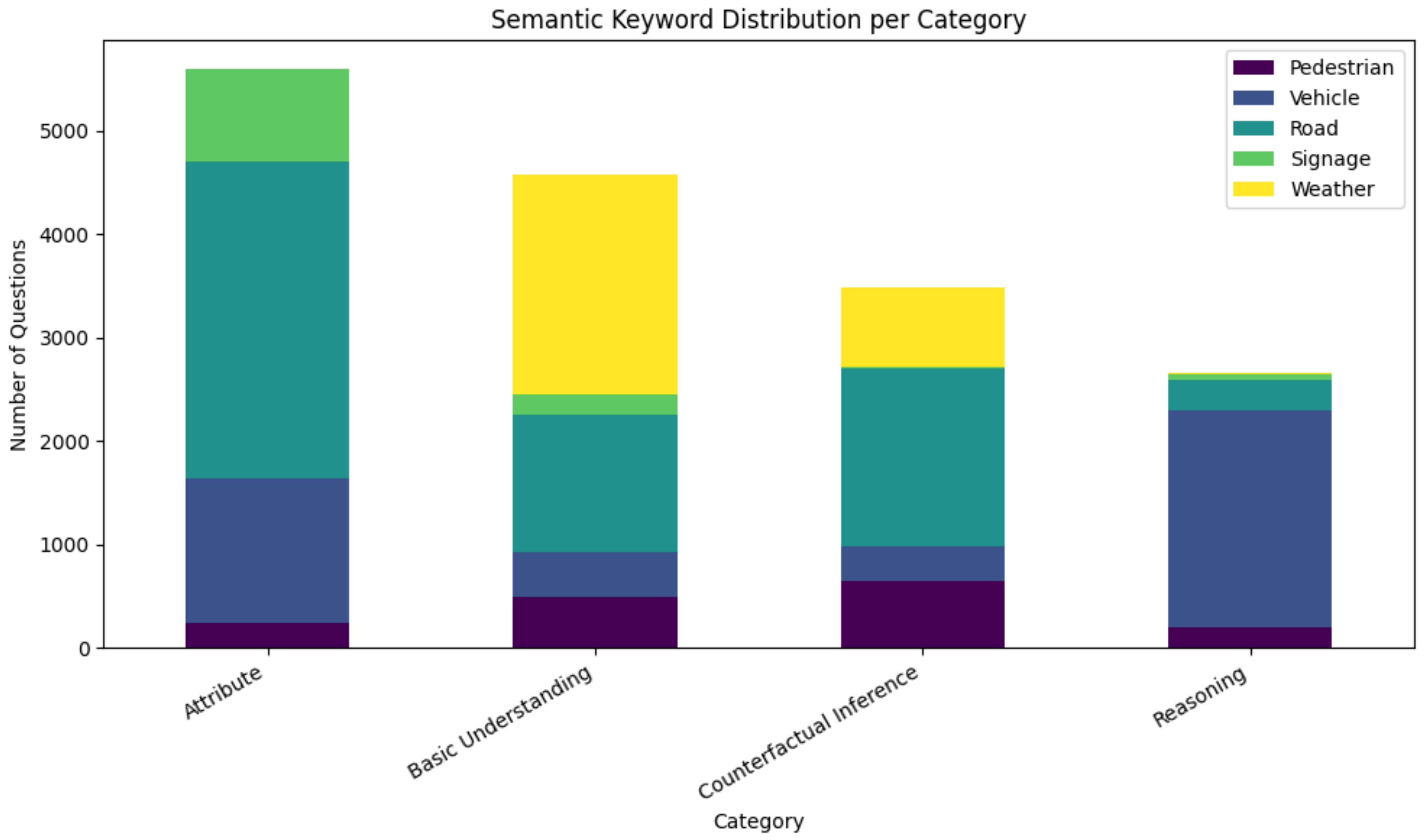
4 Are there any traffic violations? Q
In eastbound traffic, the black sedan is over-extended into the crosswalk. A

Q Which details indicate a busy city intersection?
A Heavy traffic and large commercial buildings that indicate a busy city intersection. 5

Contributions

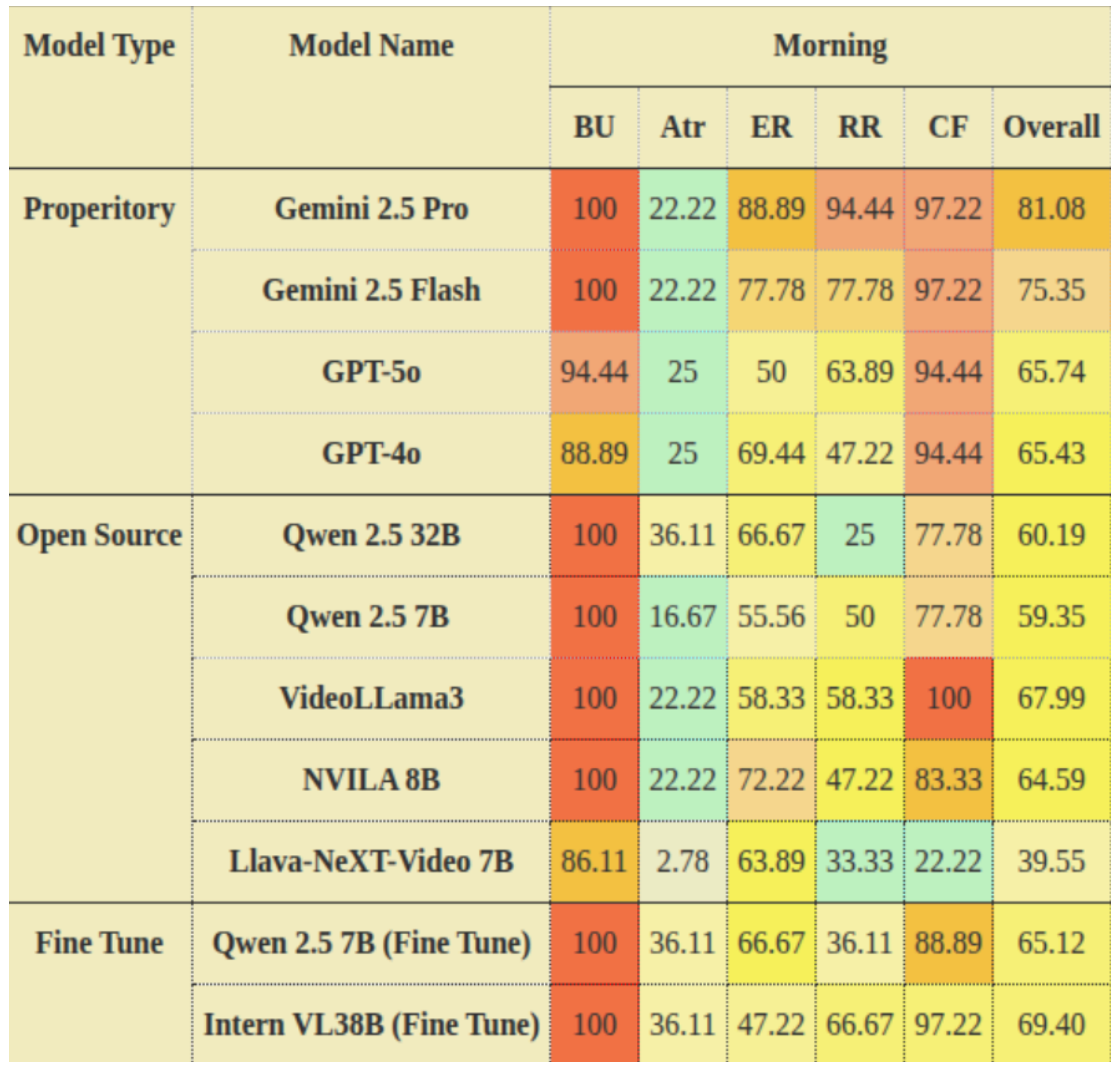
- **Defines five reasoning tasks:** attribution, counting, event, reverse, counterfactual.
- **Curates** targeted real-world QA splits for these tasks.
- **Fine-tune**s SOTA VideoQA models (Qwen 2.5 7b VL, Intern VL38b).
- **Designs metrics and ablations** to measure accuracy and generalization.

UDVideoQA Dataset Distribution



Evaluations

This study found that fine-tuned models achieved a **~5–10%** accuracy improvement over non-fine-tuned baselines, with the largest gains on attribution and event/reverse reasoning (+6–9 pts), moderate gains on counting (+4–7 pts), and smaller but consistent gains on counterfactuals (+2–4 pts). Performance was stronger on vehicular and daytime splits and dipped under occlusion and low-light conditions.



Model Type	Model Name	Morning					
		BU	Atr	ER	RR	CF	Overall
Properitory	Gemini 2.5 Pro	100	22.22	88.89	94.44	97.22	81.08
	Gemini 2.5 Flash	100	22.22	77.78	77.78	97.22	75.35
	GPT-5o	94.44	25	50	63.89	94.44	65.74
	GPT-4o	88.89	25	69.44	47.22	94.44	65.43
Open Source	Qwen 2.5 32B	100	36.11	66.67	25	77.78	60.19
	Qwen 2.5 7B	100	16.67	55.56	50	77.78	59.35
	VideoLLama3	100	22.22	58.33	58.33	100	67.99
	NVILA 8B	100	22.22	72.22	47.22	83.33	64.59
	Llava-NeXT-Video 7B	86.11	2.78	63.89	33.33	22.22	39.55
Fine Tune	Qwen 2.5 7B (Fine Tune)	100	36.11	66.67	36.11	88.89	65.12
	Intern VL38B (Fine Tune)	100	36.11	47.22	66.67	97.22	69.40