TUMTraf VideoQA: Dataset and Benchmark for Unified Spatio-Temporal Video Understanding in Traffic Scenes

Technical University of Munich

Xingcheng Zhou†, Konstantinos Larintzakis, Hao Guo, Walter Zimmer, Mingyu Liu, Hu Cao†, Jiajie Zhang, Venkatnarayanan Lakshminarasimhan, Leah Strand, Alois C. Knoll

Contact: Xingcheng.zhou@tum.de

ICML International Conference On Machine Learning

Introduction

- First large-scale video-language dataset for roadside traffic scenes understanding under diverse real-world conditions, including weather variations and accidents.
- Provide unified annotations for multi-choice QA, spatio-temporal object grounding, and referred object captioning, enabling fine-grained traffic video reasoning.

Weather
Q: How is the weather in the video?
A: The weather in the video is clear.

Class
Q: Which category is [(c5, 0.0, 0.75, 0.34)]?
A: c5 is a pedestrian.

Spatial Temporal Grounding
Q: Could you identify [(c4, 0.66, 0.21, 0.36)] in the traffic footage and give the standardized spatiotemporal localization for both the first and last frames it is present in?
A: [(c4, 0.0, 0.13, 0.25), (c4, 1.0, 0.34, 0.45)].

Counting
Q: How many motorcycles are in the footage?
A: Zero.

Motion
Q: What is the motion status of [(c12, 0.0, 0.83, 0.34)]?
A: c12 is moving.

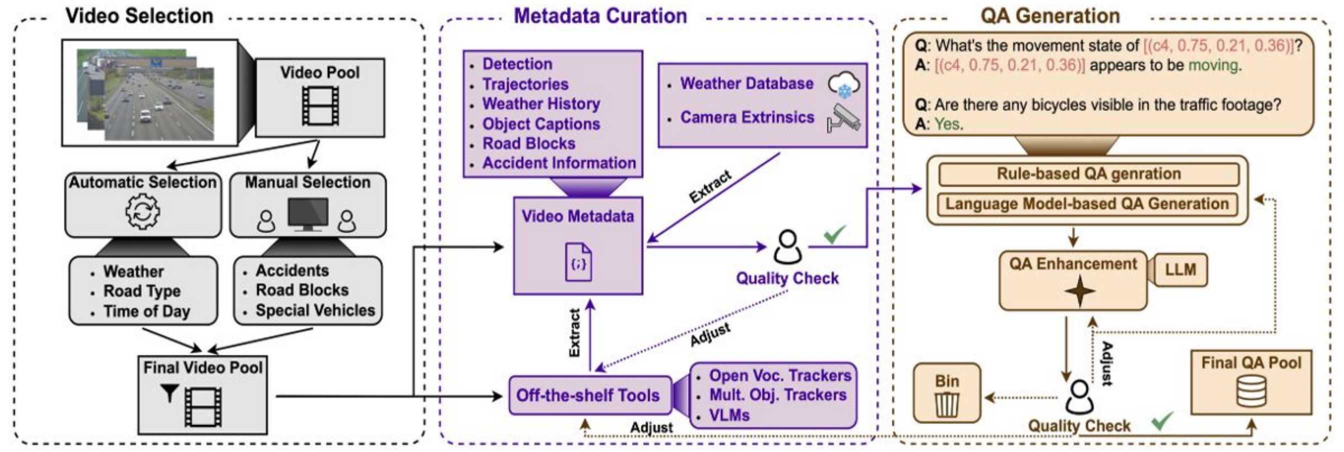
Referred Object Captioning
Q: What are the key details of [(c30, 1.0, 0.81, 0.38)]?
A: [(c30, 1.0, 0.81, 0.38)] is a bus featuring a green roof, white body, and blue lower section. It has large windows and a distinctive design that emphasizes its urban utility.

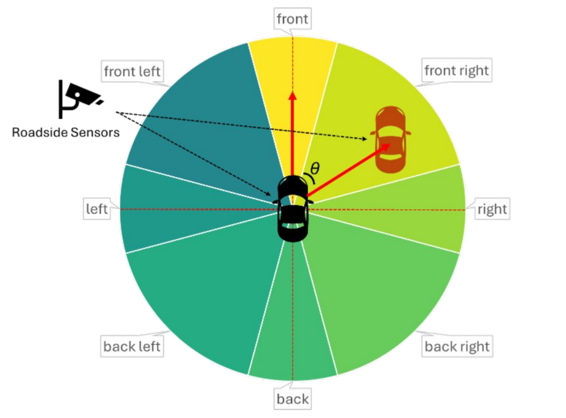
Existence
Q: Are there any trucks visible in the video?
A: Yes.

3D Positioning
Q: How does [(c23, 0.33, 0.46, 0.53)] relate to [(c28, 0.33, 0.3, 0.36)]?
A: [(c23, 0.33, 0.46, 0.53)] appears to the front of [(c28, 0.33, 0.3, 0.36)].

Dataset Curation

- Semi-automatic data pipeline leveraring LLM- and template-based methods with human verification, featuring 85,000 multiple-choice QA pairs, 2,300 object captioning, and 5,700 object grounding annotations.



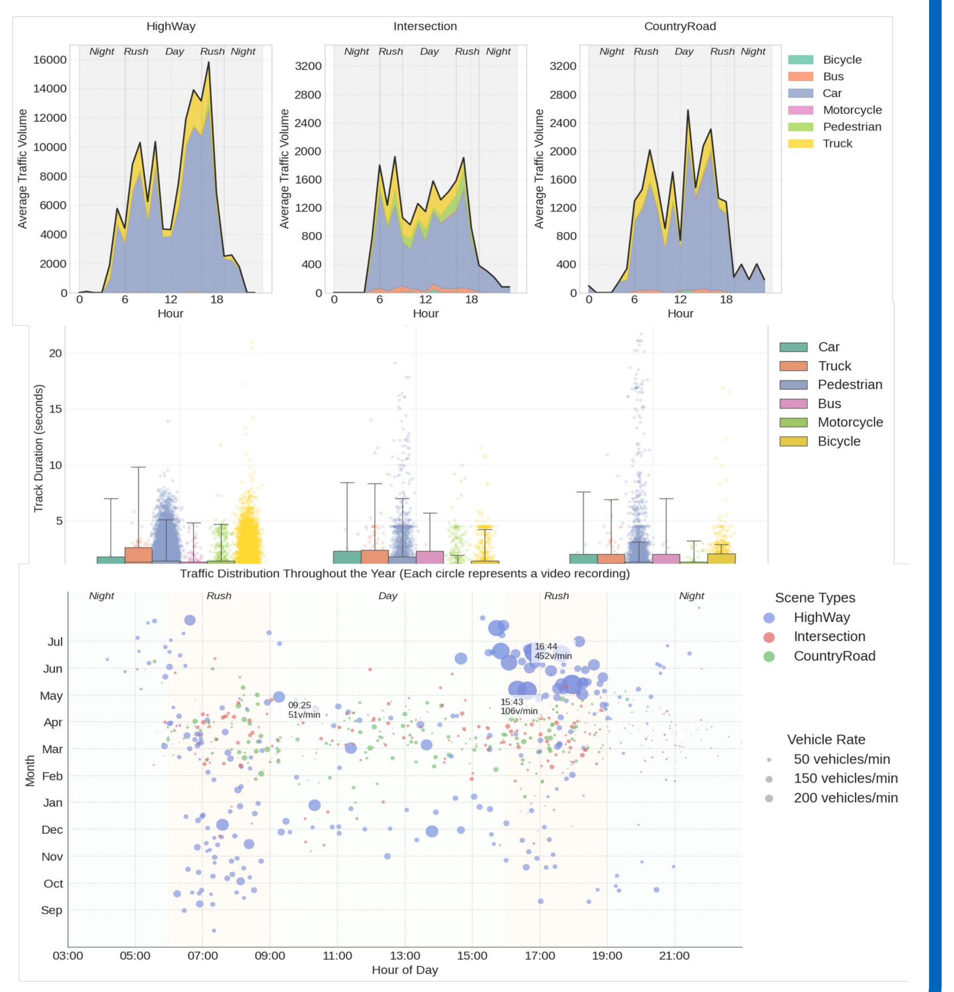


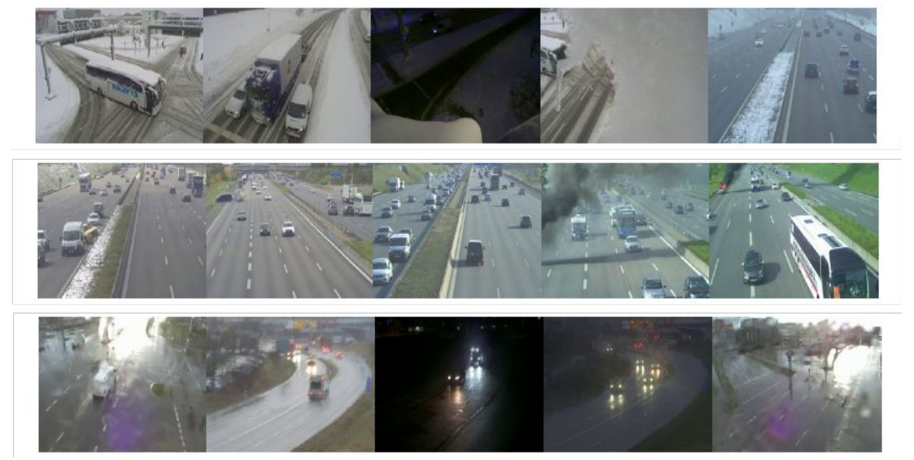
$$\text{relative position} = \begin{cases} \text{front} & \text{if } -15^\circ < \theta \le 15^\circ \\ \text{front left} & \text{if } 15^\circ < \theta \le 75^\circ \\ \text{left} & \text{if } 75^\circ < \theta \le 105^\circ \\ \text{front right} & \text{if } -75^\circ < \theta \le -15^\circ \\ \text{right} & \text{if } -105^\circ < \theta \le -75^\circ \\ \text{back left} & \text{if } 105^\circ < \theta \le 165^\circ \\ \text{back right} & \text{if } -165^\circ < \theta \le -105^\circ \\ \text{back} & \text{else.} \end{cases}$$

Dataset	Venue	Tasks	QA Gen.	# Videos/Scenes	# QAs/Captions	# Grounding	Domain
DRAMA (Malla et al., 2023)	WACV'23	Video QA	Manual	18k	102k	-	Driving
LingoQA (Marcu et al., 2024)	ECCV'24	Video QA	Manual	28k	419k	-	Driving
NuScenes-QA (Qian et al., 2024b)	AAAI'24	Image QA	Template	850	460k	-	Driving
DriveLM (Sima et al., 2024)	ECCV'24	Image QA	Temp. + Man.	188k	4.2M	-	Driving
City-3DQA (Sun et al., 2024)	ACM MM'24	Scene QA	Temp. + Man.	193	450k	-	City
HC-STVG (Tang et al., 2022)	ACM MM'22	Video Grounding	Manual	5.6k	-	5.6k	General
DVD-ST (Ji et al., 2024)	-	Video Grounding	Manual	2.7k	-	5.7k	General
UCA (Yuan et al., 2024)	CVPR'24	Video Ground. & Cap.	Manual	1.8k	-	23.5k	Surveillance
Refer-KITTI (Wu et al., 2023a)	CVPR'23	Referred-MOT	Manual	18	-	818	Driving
NuPrompt (Wu et al., 2023b)	AAAI'25	Referred-MOT	LLM	850	-	35k	Driving
TUMTraf VideoQA (Ours)	ICML'25	Video QA, ST Grounding	Temp. + LLM	1k	87.3k	5.7k	Roadside

Statistics

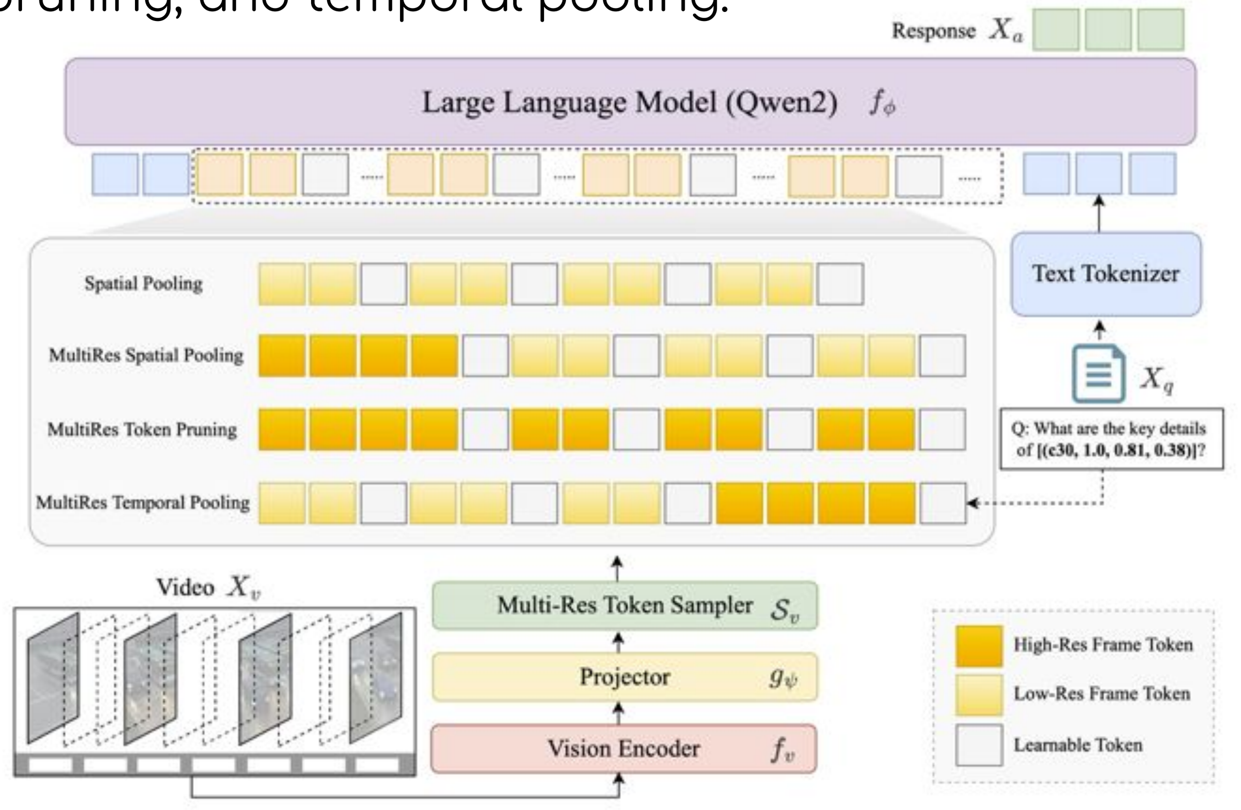
- Covers 1,000 roadside videos with diverse weather, lighting, and incidents across highway, urban, and rural scenes.
- Balanced coverage across times of day and seasons under varied traffic conditions, reflecting real-world traffic dynamics.





TraffiX-Qwen

- Video-LLM for unified traffic video understanding, across multi-choice QA, spatio-temporal grounding, and referred object captioning.
- Leveraging multi-resolution visual sampling strategies, i.e., multi-res spatial pooling, token pruning, and temporal pooling.



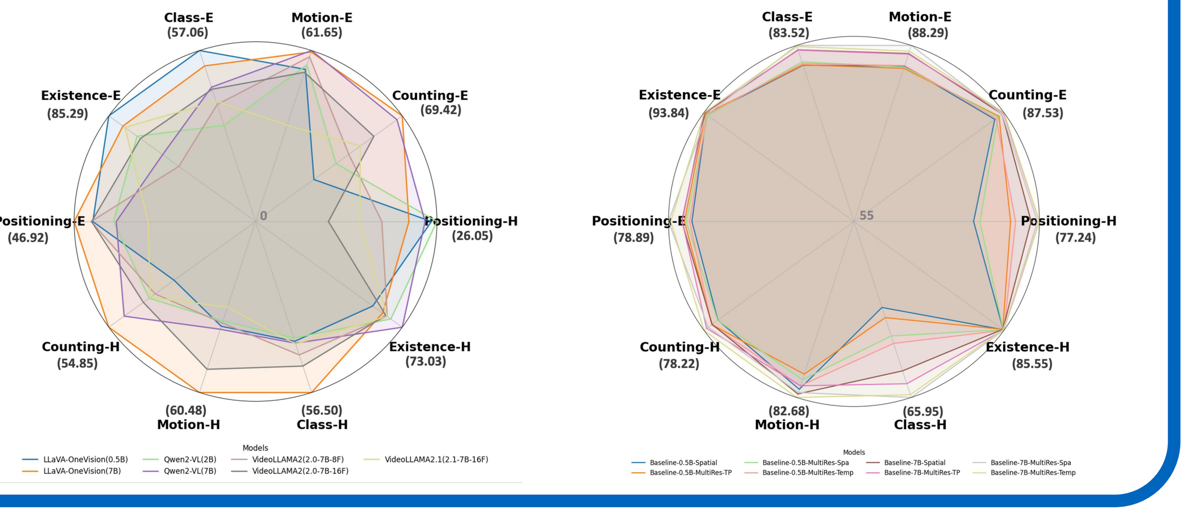
Benchmark

Models	Category	Positioning E	Positioning H	Counting E	Counting H	Motion E	Motion H	Class E	Class H	Existence E	Existence H	Overall
Open-Source Models												
LLAVA-OneVision	0.5B	42.10	25.26	27.62	30.45	54.87	37.04	**57.06**	39.57	**85.29**	58.35	45.82
	7B	**46.92**	22.03	**69.42**	**54.85**	61.14	**60.48**	51.92	**56.50**	77.08	63.25	**56.36**
Qwen2-VL	2B	36.73	**26.05**	38.10	39.78	56.46	35.19	32.10	38.49	68.87	67.32	43.91
	7B	36.03	24.35	66.91	49.11	**61.65**	38.10	44.83	40.20	54.00	**73.03**	48.82
VideoLLaMA2	2.0-7B-8F	42.54	18.14	44.13	37.56	59.37	35.87	39.05	44.07	44.56	65.56	43.09
	2.0-7B-16F	42.41	10.47	55.98	41.94	53.80	52.26	44.16	47.75	66.93	64.82	48.05
TraffiX-Qwen Baseline												
Baseline-0.5B (Ours)	Spatial Pooling	75.54	68.47	85.31	75.82	83.92	**81.26**	79.95	59.73	93.06	85.37	78.84
	MultiRes Spatial-Pooling	76.36	69.32	86.10	75.86	83.73	79.59	**80.57**	61.70	92.73	85.37	79.07
	MultiRes Token-Pruning	**76.61**	73.40	**86.33**	76.88	83.48	78.60	80.01	60.43	**93.34**	85.27	79.44
	MultiRes Temporal-Pooling	75.85	**74.07**	85.65	**76.92**	**84.05**	80.64	80.26	**62.21**	93.06	**85.55**	**79.83**
Baseline-7B (Ours)	Spatial Pooling	76.99	76.14	87.07	76.81	86.58	82.07	82.72	64.11	93.62	85.27	81.14
	MultiRes Spatial-Pooling	**78.89**	76.99	87.07	77.49	**88.29**	81.82	**83.52**	**65.95**	93.01	**85.51**	81.85
	MultiRes Token-Pruning	76.93	**77.24**	87.41	77.76	86.46	80.64	82.66	65.00	**93.84**	85.48	81.34
	MultiRes Temporal-Pooling	78.57	**77.24**	**87.53**	**78.22**	87.09	**82.68**	83.33	65.76	93.78	85.34	**81.95**

Model	Bleu_4	ROUGE_L	CIDEr	METEOR	SPICE
Open-Source Models					
LLAVA-OneVision (0.5B)	0.48	10.16	0.0102	-	-
LLAVA-OneVision (7B)	5.77	14.09	0.1326	-	-
Qwen2-VL (2B)	8.72	17.93	0.2086	-	-
Qwen2-VL (7B)	**10.47**	**20.14**	**0.4119**	-	-
VideoLLaMA2 (7B-8F)	6.25	19.94	0.2391	-	-
VideoLLaMA2 (7B-16F)	6.87	18.69	0.2111	-	-
TraffiX-Qwen Baseline 0.5B					
Spatial-Pooling	34.99	50.44	2.5195	35.24	46.35
MultiRes Spatial-Pooling	34.91	50.26	2.4306	35.20	45.75
MultiRes Token-Pruning	35.07	50.79	**2.5730**	35.30	46.48
MultiRes Temporal-Pooling	**35.63**	**51.00**	2.5464	**35.77**	**47.17**
TraffiX-Qwen Baseline 7B					
Spatial-Pooling	36.74	52.04	2.5613	36.42	47.32
MultiRes Spatial-Pooling	37.60	**53.26**	2.6113	**37.31**	**49.16**
MultiRes Token-Pruning	**37.83**	52.31	**2.6162**	36.56	47.80
MultiRes Temporal-Pooling	37.48	52.58	2.4236	36.85	48.53

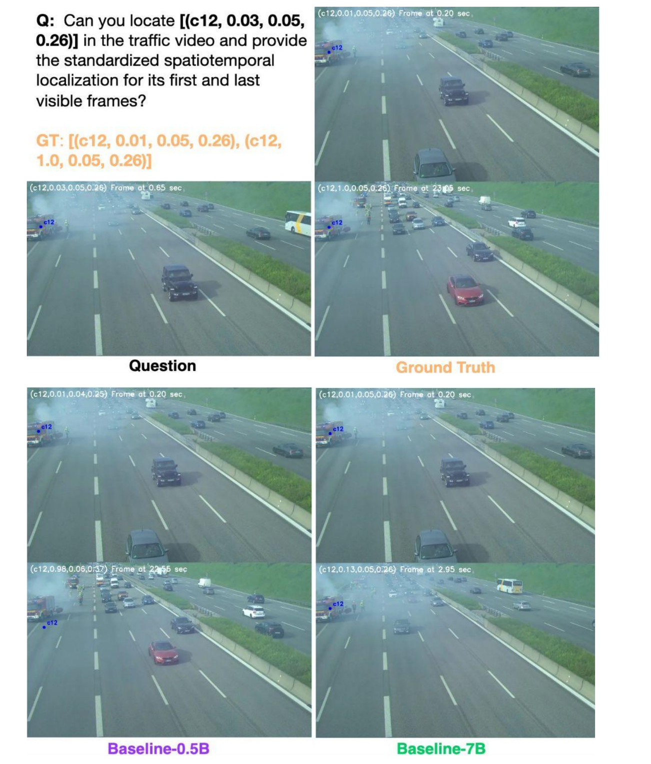
Model	Temporal E↓	Spatial E↓	ST E↓
Open-Source Models			
LLAVA-OneVision (0.5B)	0.7285	0.7212	0.8415
LLAVA-OneVision (7B)	0.7680	0.7750	0.8142
Qwen2-VL (2B)	0.7729	0.7793	0.8127
Qwen2-VL (7B)	0.7615	0.7647	0.8032
VideoLLaMA2 (7B-8F)	**0.6225**	**0.6360**	**0.6896**
VideoLLaMA2 (7B-16F)	0.7218	0.7383	0.7895
TraffiX-Qwen Baseline			
0.5B-Spatial-Pooling	0.1220	**0.1892**	**0.2600**
0.5B-MultiRes-Spatial-Pooling	**0.1211**	0.1894	0.2607
0.5B-MultiRes-Token-Pruning	0.1230	0.1934	0.2650
0.5B-MultiRes-Temporal-Pooling	0.1228	0.1912	0.2629
7B-Spatial-Pooling	**0.1083**	**0.1737**	**0.2382**
7B-MultiRes-Spatial-Pooling	0.1136	0.1822	0.2493
7B-MultiRes-Token-Pruning	0.1152	0.1748	0.2454
7B-MultiRes-Temporal-Pooling	0.1166	0.1790	0.2496

- Multi-resolution visual sampling strategies yield moderate improvements in multi-choice QA and referred object captioning, but introduce challenges for spatio-temporal grounding due to inter-frame association ambiguities.
- Tasks requiring fine-grained 3D spatial and temporal reasoning, including object positioning and motion analysis, remain challenging for current Visual LLMs.
- Accurate spatio-temporal object localization and cross-frame object association remain difficult for visual LLMs in dynamic traffic environments.



Qualitative

Spatio-Temporal Grounding



Referred Object Captioning