# TUMTraffic-VideoQA: A Benchmark for Unified Spatio-Temporal Video Understanding in Traffic Scenes

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

We present TUMTraffic-VideoQA, a novel dataset and benchmark designed for spatio-temporal video understanding in complex roadside traffic scenarios. The dataset comprises 1,000 videos, featuring 85,000 multiple-choice QA pairs, 2,300 object captioning, and 5,700 object grounding annotations, encompassing diverse real-world conditions such as adverse weather and traffic anomalies. By incorporating tuple-based spatio-temporal object expressions, TUMTraffic-VideoQA unifies three essential tasks-multiple-choice video question answering, referred object captioning, and spatio-temporal object grounding-within a cohesive evaluation framework. We further introduce the TUMTraffic-Qwen baseline model, enhanced with visual token sampling strategies, providing valuable insights into the challenges of fine-grained spatio-temporal reasoning. Extensive experiments demonstrate the dataset's complexity, highlight the limitations of existing models, and position TUMTraffic-VideoQA as a robust foundation for advancing research in intelligent transportation systems. The dataset and benchmark are publicly available to facilitate further exploration.

Full document at https://doi.org/10.48550/arXiv.2502.02449