InterAct VideoQA: Dataset Description

Version: Beta (last submitted on March 07, 2025)

InterAct VideoQA: Download here

InterAct VideoQA is a benchmark dataset designed to facilitate Video Question Answering (VideoQA) research, specifically in traffic intersection monitoring. The dataset consists of real-world traffic footage captured from both stationary surveillance cameras and mobile devices. Each video is segmented into short clips (e.g., 10 seconds each) to capture specific events or interactions in detail. To further aid training and evaluation, each clip is accompanied by carefully curated question-answer (QA) pairs covering multiple reasoning types (e.g., spatio-temporal, counting, attribution, and event inference).

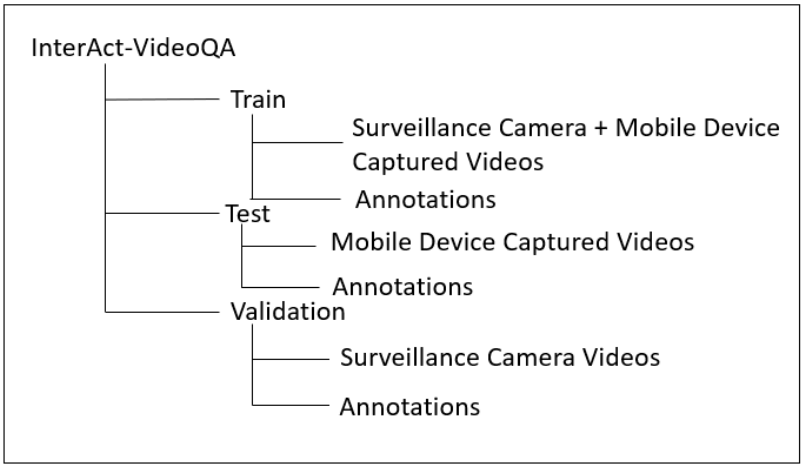
Data was gathered from multiple urban intersections—spanning high-density rush hours, moderate midday flows, and low-density nighttime traffic—using both stationary surveillance cameras and mobile devices. The surveillance cameras, mounted at intersections, captured top-down or mid-range views, while handheld or vehicle-mounted smartphones/tablets provided ground-level or moving perspectives, adding variety in angles and potential motion blur. To ensure comprehensive real-world coverage, the recording took place under diverse lighting (daytime, dawn, dusk, and nighttime) and weather conditions (clear, rainy, cloudy). Below are the location/s of the gathered data.

Location	Address	Complexity
Mill Avenue DownTown Tempe	**Intersection between South Mill Avenue and 7th Street, Tempe**	**A busy urban area with diverse data on vehicle dynamics, pedestrian flow, and complex interactions among traffic participants**

Key statistics of the **InterAct VideoQA** dataset include:

- **Total Duration: 8 hours of traffic footage.**
- **Clip Duration: 10 seconds per clip.**
- **Total Clips: 2880 clips.**
- **Total QA Pairs: 28800 question-answer pairs.**
- **Reasoning Categories: Attribution, Counting, Event Reasoning, Reverse Reasoning and CounterFactual Inference.**

The dataset is organized into three main splits**: Train, Test,** and **Validation.** Each split contains video clips and their corresponding annotation files. A simplified view of the directory layout is shown below:



Train: The training set combines videos from both fixed-position surveillance cameras and handheld/mobile devices, offering coverage of angles, resolutions, and lighting conditions. Annotations (in CSV) provide question-answer pairs and timestamps, ensuring a comprehensive resource for diverse VideoQA training needs.

Test: The test set primarily contains mobile-device-sourced video clips, which introduce different perspectives from surveillance feeds. It includes question-answer annotations generally unseen during training, enabling a thorough assessment of the model's generalization capabilities.

Validation: The validation set features exclusively surveillance camera footage, providing consistent viewpoints for a more controlled evaluation. It contains question-answer annotations and metadata used for interim performance checks, hyperparameter tuning, and early stopping decisions.

Note: Question Types are mixed across training,testing and validation.

Video clips are typically stored in standard formats (MP4 and MOV) at a consistent frame rate, with filenames that often include timestamps or location IDs for reference. Annotations, provided in structured formats such as CSV, include fields like a unique clip identifier (clip_id), timestamps indicating event start/end, textual questions about the scene (question), ground-truth responses (answer), and a reasoning category (reasoning_type).