# Multi-Agent Cooperation for Traffic Safety Description and Analysis

Ridham Kachhadiya, Dhanishtha Patil, David C. Anastasiu

## Background

Understanding complex traffic interactions from video is crucial for intelligent transportation systems and smart-city safety. However, multi-camera scenes vary dramatically in viewpoint, illumination, and semantics, making unified vision-language models (VLMs) unreliable.

## Problem Statement

**How can we design a system that:**

- Describes pedestrian and vehicle interactions (dense captions)
- Answers safety questions (fine-grained VQA)
- Adapts to different viewpoints and environments
- Remains explainable for downstream decision-making

**Datasets**

- Internal (WTS): staged accidents filmed in Woven city, dual views (overhead + vehicle)
- External (BDD): natural urban driving from U.S. cities (vehicle view only)
- Combined Dataset: 6K video segments x 5 event phases *(Prerecognition, Recognition, Action, Avoidance, Judgement)*

*Fig 1. Vehicle view vs Overhead view*

## Motivation

- Traditional traffic video analysis models focus on detection and tracking, but they fail to explain why events occur or how risk develops.
- Traffic scenes are inherently multimodal, combining visual, spatial, and behavioral cues across multiple viewpoints.
- A single model cannot generalize across such variation.
- Our motivation is to build a modular, cooperative system that learns how different agents perceive and describe a scene, producing safety-aware and interpretable predictions.

## Proposed Framework

**Multi-Agent Framework**

- Backbone: InternVL3-14B for all agents.
- Specialized agents by role (ped. vs. veh.), domain (WTS vs. BDD), semantic group (hard QA subset), and prompt input style (facts vs key-value).
- Validation-guided selection chooses the best agent per QA type / caption segment.

**Frame Sampling & Grounding**

- Two samplers to capture temporal cues: Evenly-Spaced and Midpoint-Centric (k-spaced)
- Bounding box overlay: red for pedestrians and blue for vehicles, improving spatial grounding.

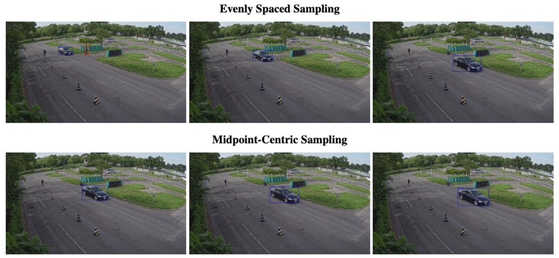


*Fig 2. Frames in both sampling strategies*

**Prompting & Fact Conditioning**

- Caption prompts: phase-aware, role-aware, bbox-aware.
- QA prompts: phase and viewpoint context; answer by letter.
- Fact augmentation: convert QA annotations into natural-language facts prepended to images, boosts semantic fluency.

## Pipeline Overview

- The proposed framework executes the two AI City Challenge Track 2 Sub-Tasks at inference (VQA and Captions).
- Each input video is divided into five event phases, and representative frames are extracted using evenly-spaced and midpoint-centric sampling.
- A multi-agent QA module first answers 43 questions from the frames, producing a structured dictionary of key-value pairs that captures scene semantics (e.g., pedestrian action, road layout, visibility).
- These outputs are transformed into natural-language fact prompts and passed together with the frames to a multi-agent captioning module, which generates phase-aware dense pedestrian and vehicle captions.

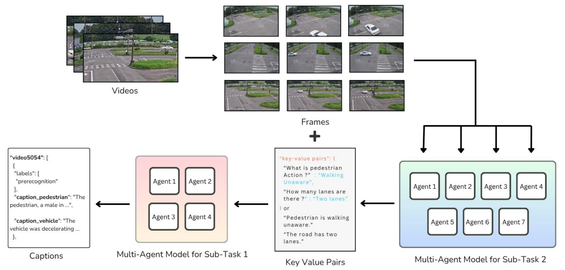


*Fig 3. Pipeline Diagram*

## Experiments and Ablation Study

**Training Setup**

- **InternVL3-14B**, full-parameter fine-tuning; DeepSpeed ZeRO-2 on HGX A100 (8×80GB).
- Metrics: **BLEU-4, METEOR, ROUGE-L, CIDEr** (captions) and **Accuracy** (VQA).

| Frame Strategy | Dataset | Q-Type | Epochs | WTS (%) | BDD (%) | Overall (%) |
|---|---|---|---|---|---|---|
| Mid k-spaced | Both | all | 3 | 81.75 | 86.37 | 85.90 |
| Mid k-spaced | Both | all | 2.5 | 81.58 | 86.03 | 85.58 |
| Mid k-spaced | Both | all | 2 | 81.14 | 85.85 | 85.37 |
| Mid k-spaced | Both | subset | 2 | 73.57 | 76.81 | 76.37 |
| Mid k-spaced | BDD | all | 2 | n/a | 86.08 | n/a |
| Evenly spaced | WTS | all | 2 | 75.85 | n/a | n/a |
| Evenly spaced | BDD | all | 2 | n/a | 84.46 | n/a |
| **Multi-agent** | n/a | n/a | n/a | **84.31** | **87.46** | **87.14** |

| Strategy | Split | Input | Int Veh | Int Ped | Ext Veh | Ext Ped | Overall |
|---|---|---|---|---|---|---|---|
| Mid k-spaced | No | Facts | 43.12 | 31.49 | **44.45** | **31.33** | 37.59 |
| Evenly spaced | No | Key-Val | **45.68** | **32.65** | 43.31 | 30.37 | 38.00 |
| Mid k-spaced | Ped-Veh | Key-Val | 41.11 | 31.74 | 42.41 | 30.86 | 36.53 |
| **Multi-agent** | n/a | n/a | **45.68** | **32.65** | **44.45** | **31.33** | **38.53** |

*Fig 4. Scores for individual and multi-agent VQA and captioning agents*

## Results

**Why Modular Beats Unified**

- A single cross-task model underperforms on both captioning and QA versus specialized multi-agents.
- Comparison of Unified and Modular avg score for both the subtasks in Fig 5.

| Task | Unified Cross-Task Model | Ours (Modular) |
|---|---|---|
| Captioning | 23.47 | **38.53** |
| QA Accuracy (%) | 81.43 | **87.14** |

*Fig 5. Avg. score in both cases*

## Leaderboard

| Rank | Team ID | Team Name (Affiliation) | Score |
|---|---|---|---|
| 1 | 145 | CHTTLIOT (Chunghwa Telecom, Taiwan) | 60.0393 |
| **2** | **1** | **SCU_Anastasiu (SCU, USA)** | **59.1184** |
| 3 | 52 | Metropolis_Video_Intelligence (NVIDIA, USA) | 58.8483 |
| 4 | 137 | ARV (ARV, Thailand) | 57.9138 |
| 5 | 121 | Rutgers ECE MM (Rutgers University, USA) | 57.4658 |

*Fig 6. Avg. score in both cases*

**Fig. 6 shows the top 5 teams in the final 2025 AI City Challenge Track 2 leaderboard. Our team, SCU Anastasiu, secured second place using the proposed multi-agent framework.**

## Conclusion

- **Targeted Specialization:** Role-aware and domain-specific agents adapt more effectively to visual and semantic variations in traffic scenes.
- **Validation-Driven Routing:** Each output unit is dynamically assigned to the best-performing agent, leading to stronger consistency and robustness.
- **Future Traffic Intelligence:** Demonstrates the potential of agentic modeling for scalable, context-sensitive, and explainable vision-language systems in real-world traffic analysis.

This work opens pathways for future research on modular and interpretable AI for traffic safety and intelligent transportation systems.

## Acknowledgements

Research supported by a Supermicro GPU SuperServer SYS-420GP-TNAR+ node contributed by Supermicro and NVIDIA, integrated into the Santa Clara University HPC.

Contact Information: danastasiu@scu.edu