Multi-Agent Cooperation for Traffic Safety Description and Analysis

Ridham Kachhadiya • Dhanishtha Patil • David C. Anastasiu

Santa Clara University, ICCV Workshop 2025, AI City Challenge Track 2

Motivation

Understanding traffic videos is critical for smart-city safety.

Diverse viewpoints and lighting make scene analysis hard.

Existing single-model systems lack robustness and explainability.

Goal: Design an **explainable multi-perspective framework** for traffic safety analysis.

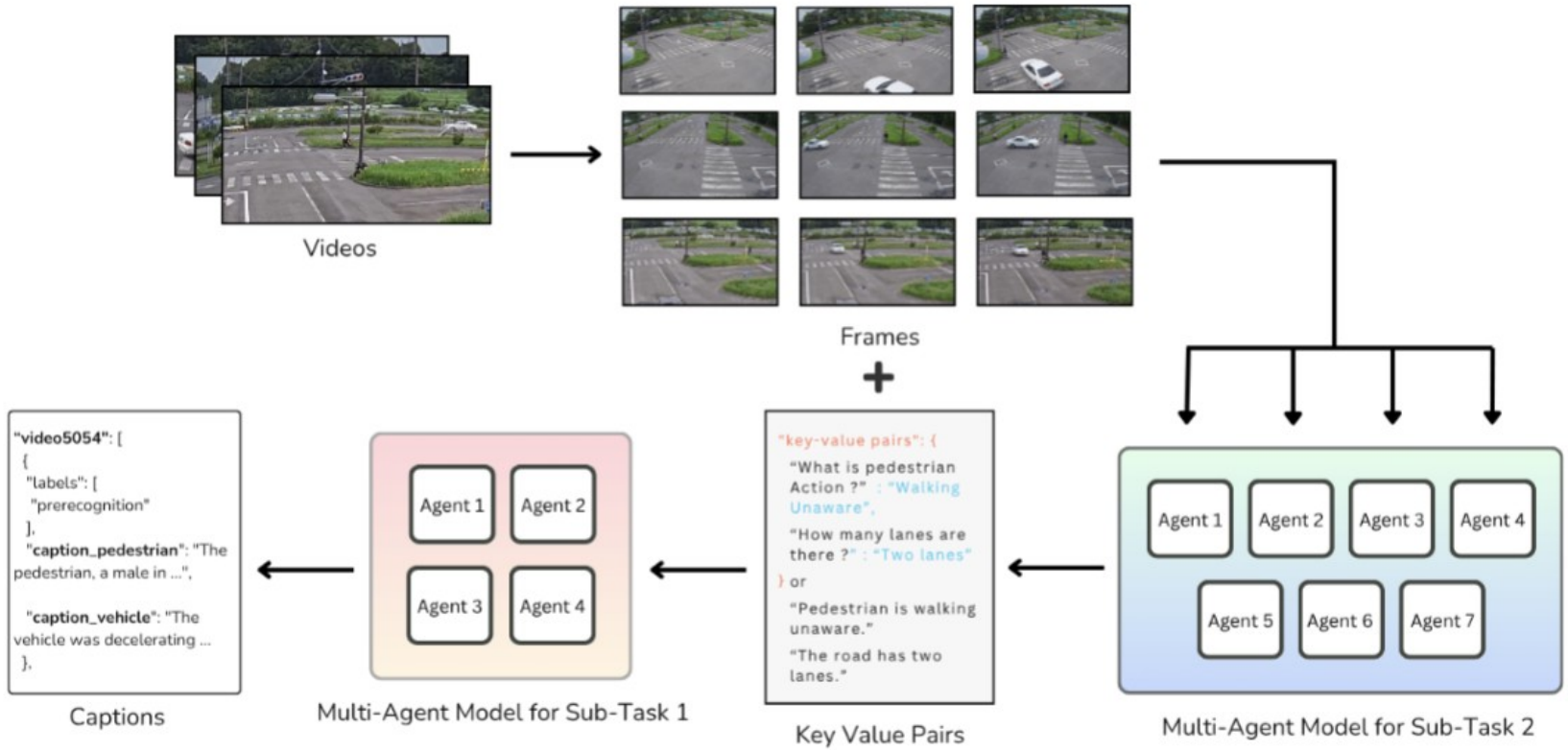
Problem & Challenge

- **Input** : multi-camera accident videos (overhead + vehicle views).
- **Tasks:**
  • Structured captions (vehicle & pedestrian)
  • Fine-Grained QA (trajectory, awareness, environment etc.)
- **Challenges:** viewpoint variance, occlusion, semantic diversity.
- Single VLM fails → need modular reasoning

Proposed Multi-Agent Framework

- Modular multi-agent setup with separate VLAs for pedestrian, vehicle, and QA reasoning.
- Two-stage workflow:
  → 43 Safety-QA extracts motion, awareness, and environment cues
  → Fact/QA - guided captioning generates detailed role-based descriptions.
- Validation-guided routing picks the most accurate agent per sub-task for consistent results.
- Outputs from all agents are combined into a unified scene summary with higher accuracy and clearer reasoning.
- Framework scales easily to new domains and camera views.

Video frames are analyzed by multi-agent QA models to extract key facts, then passed to captioning agents that generate structured scene descriptions using the best-performing model per segment.



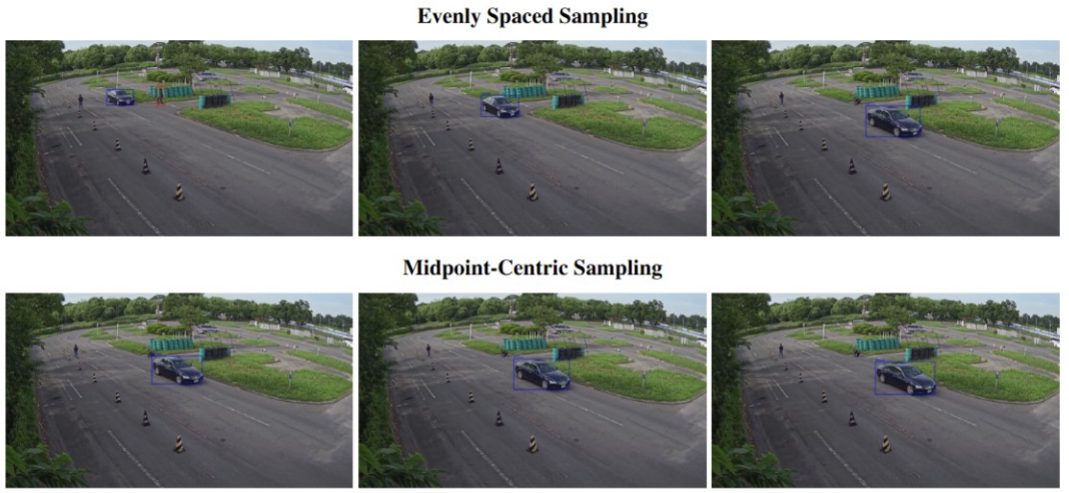
Frame Sampling Strategies

- **Evenly Spaced Sampling**

  Extracts four frames at regular intervals across the phase duration. Effective for capturing gradual scene transitions and broad context.

- **Midpoint-Centric Sampling**

  Focuses on localized temporal cues by selecting frames centered around the midpoint with a fixed offset (k=10). Better for capturing movement dynamics.



All sampled frames were overlaid with bounding boxes (red for pedestrian, blue for vehicle) to improve visual grounding.

Agent Specialization and Training

All agents share the **InternVL3-14B backbone**, but differ in their training supervision to focus on distinct reasoning tasks.

- **Role Specialization:**
  Separate models are trained for pedestrian and vehicle captions, enabling focus on role-specific cues such as gaze, posture, and motion.

- **Domain Specialization:**
  Independent agents are trained on internal (WTS) and external (BDD) datasets to handle different camera angles, viewpoints, and scene compositions.

- **Semantic Grouping (QA):**
  A specialized QA agent is trained on a subset of 14 challenging questions involving complex visual cues like gaze and line of sight, improving fine-grained reasoning.

Fact-Augmented Captioning

- To strengthen semantic grounding, we enrich VLM inputs with structured factual statements derived from QA annotations.
- Example:
  Q: What is the age group of the pedestrian? => A: 30s
  Fact statement: *"The pedestrian is in the 30s age group."*
- These fact statements are added before the visual tokens, giving the model explicit, human-like context.
- The approach helps the captioning agents generate more accurate and detailed descriptions, especially for internal pedestrian scenes where semantic cues are dense.




Pedestrian caption: The pedestrian, a man in his 20s, stood diagonally to the right in front of the vehicle on a residential road. He was unaware of the vehicle approaching him, which was traveling in the opposite direction. The pedestrian was wearing a brown jacket and navy blue slacks, and the weather was rainy with dim lighting. The road surface was dry and level, comprised of asphalt. As the vehicle approached, the pedestrian suddenly rushed out and began moving slowly in the direction of straight ahead. There were no street lights present, but the road had usual traffic volume. The event took place on a two-way traffic road with no sidewalks on both sides, roadside strips, or street lights. The pedestrian's action of rushing out put him in close proximity to the vehicle, posing a potential danger given his unawareness of its presence.

Vehicle caption: The vehicle is positioned on the right side of the pedestrian and is close to them. The vehicle can see the pedestrian as it is in its field of view. The vehicle is currently turning right at a speed of 10 km/h. The environment conditions indicate that the pedestrian is a male in his 20s with a height of 170 cm. He is wearing a brown jacket and navy blue slacks. The weather is rainy and the brightness is dim. The road surface conditions are dry and the road is not inclined. The road surface is made of asphalt and the traffic volume is usual. The road is classified as a residential road with two-way traffic and there is no sidewalk on either side.

Performance Gains: Modular vs. Unified

Our multi-agent framework consistently outperformed a single unified model trained jointly on both captioning and QA tasks.

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

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

The multi-agent QA model achieved 87.14% QA accuracy, a gain from the best single-agent QA-only model score of 85.90%, confirming that validation-guided routing successfully captures complementary strengths.

Conclusion: The Power of Agentic Modeling

The proposed modular, validation-driven multi-agent system achieved superior performance in the 2025 AI City Challenge.

Targeted Specialization

Role-aware training and domain-specific agents adapt better to visual and semantic characteristics.

Validation-Driven Routing

Dynamically selects the best model per output unit, improving overall robustness.

Future Traffic Intelligence

Validates the agentic approach for modular, context-sensitive VLM in real-world traffic analysis.

This work opens pathways for future research into explainable and accurate AI solutions for traffic safety.

Questions?

Contact: David C. Anastasiu
(danastasiu@scu.edu)

Dhanishtha Patil

Ridham Kachhadiya