Multi-Agent Cooperation for Traffic Safety Description and Analysis

Ridham Kachhadiya, Dhanishtha Patil, and David C. Anastasiu
Computer Science and Engineering
Santa Clara University
{rkachhadiya, dpatil, danastasiu}@scu.edu
https://github.com/davidanastasiu/tsda

Abstract

Understanding complex traffic scenes from video remains a core challenge in building intelligent transportation systems, especially under varied viewpoints and semantic demands. To address this problem, we introduce a modular multi-agent framework targeting the Traffic Safety Description and Analysis track of the 2025 AI City Challenge. Our approach dynamically assigns specialized vision-language agents to individual sub-tasks, both for fine-grained safety question answering and for structured pedestrian and vehicle captioning, based on validation set performance. Each agent is trained using distinct supervision strategies, such as role-based data partitions, fact-augmented inputs, and semantically grouped QA subsets, allowing them to excel in specific roles. During inference, we route each input segment or question to its best-performing agent, yielding robust and context-aware outputs. Compared to unified models, our framework delivers consistent improvements across standard metrics like BLEU, METEOR, ROUGE, CIDEr, and QA accuracy. The proposed solution ranked 2nd on the official leaderboard, demonstrating the strength of targeted model specialization, task decomposition and cooperative inference in multimodal video understanding.

1. Introduction

As urban environments deploy increasingly dense networks of street-level and vehicle-mounted cameras, the ability to understand and describe traffic scenes from video has become crucial for intelligent transportation systems. Recent research on explainable anticipation systems [3] has advocated for traffic safety models that go beyond post hoc event detection and instead focus on proactive and explainable understanding of pedestrian-vehicle dynamics, capturing both visual and contextual cues before incidents occur.

Building on this vision, the AI City Challenge Traffic Safety Description and Analysis Track [1] was introduced in 2024 as a benchmark that pushes the boundaries of multimodal video understanding. The track was reprised in 2025 with two sub-tasks. Sub-task 1 requires phase-wise generation of structured captions and fine-grained safety question answering across multi-camera traffic event segments. For each incident, models must generate separate pedestrian and vehicle descriptions across two distinct datasets. The "internal" dataset, which re-created pedestrian-vehicle accidents using stunt people, was filmed by Woven By Toyota in a smart city concept in Japan and includes both overhead- and vehicle-view camera viewpoints. The "external" dataset contains a subset of ego vehicle videos from the BDD-100K dataset [17] denoting vehicles driving in a variety of US large cities. These diverse perspectives lead to variations in scene framing, lighting, and motion patterns. Sub-task 2 introduces significant complexity by adding a question answering (QA) component which spans 41 distinct question types covering appearance, trajectory, awareness, environment, and interaction cues for all scenes, a subset of which are asked both from the vehicle and pedestrian perspectives.

Challenges arising from varied viewpoints, occlusions, temporal complexity, and semantic diversity often render single-model systems inadequate [16], motivating the development of modular, adaptive reasoning pipelines tailored to specific sub-tasks. To address these challenges, we propose a multi-agent framework that dynamically selects the best performing model for each sub-task based on validation set behavior and output quality. Instead of relying on a single vision language model or static ensemble, we trained a diverse pool of models (*agents*) using variations in frame sampling strategies, epoch checkpoints, and QA-to-fact reformulations. For sub-task 1, we assigned separate agents to generate pedestrian and vehicle captions for each dataset split, enabling finer control over viewpoint and subject specialization. For sub-task 2, we evaluated per-question accuracy across all agents and routed each of the 41 safety questions to the model that performed best for that question. This validation-driven assignment strategy allowed each agent to specialize on the sub-task it handled best, leading to improved performance across both captioning and QA objectives. Our results demonstrate that targeted agent cooperation outperforms uniform modeling approaches while preserving consistency across outputs.

Our key contributions are threefold:

- We propose a validation-guided agent assignment framework, wherein different VLM agents are selected for specific QA types and caption segments based on their empirical strengths, highlighting how architectural diversity and training variation could be harnessed through selective routing.
- We show that this approach enables modular cooperation between large-scale models (e.g., InternVL3-14B), resulting in outputs that are both factually grounded and viewpoint-aware.
- Our solution is competitive against state-of-the-art approaches from some of the best teams in the world, achieving 2nd place on the 2025 AI City Challenge Track 2 leaderboard.

Together, these contributions offer a principled solution to the challenges of heterogeneous input-output mappings in multimodal traffic video analytics and provide a path for both explainable and accurate AI solutions to the problem.

2. Related Work

2.1. Video Understanding Models

Vision Language Models (VLMs) have made significant progress in multimodal reasoning, especially for video tasks that require temporal alignment and structured scene interpretation. Foundational models such as Flamingo [2], BLIP-2 [10], and InternVideo [15] integrate multi-frame visual inputs with large language models, enabling both video captioning and visual question answering (VQA).

Recent advancements like InternVL [5] and Qwen-VL [4] improved these capabilities by supporting high-resolution and context-rich visual inputs through dense supervision and image grouping mechanisms. However, applying these models to driving scenes remains challenging due to abrupt viewpoint shifts (e.g., vehicle to overhead), sparse temporal sampling, and complex inter-agent dynamics. In our experiments, LLaMA [13] and Qwen LLMs performed well in language generation but struggled with visual grounding without a strong encoder. InternVL-14B, equipped with an EVA-CLIP vision backbone and instruction-tuned language decoder, consistently outperformed alternatives on both captioning and QA tasks.

2.2. Traffic Scene Understanding with Vision–Language Models

Understanding traffic scenes from video has been central to intelligent transportation systems. The AI City Challenge Track 2 serves as a benchmark for structured captioning and safety-focused VQA. Recent top-performing methods demonstrated the utility of structured prompts and specialized input strategies.

CityLLaVA [8] introduced bounding-box prompts and dual-view encodings to improve spatial alignment. AIO-ISC [16] used semantic templates and post-processing rules, while TrafficVLM [7] applied phase-specific encoding to align language with event structure. These techniques highlight the benefits of adapting VLM behavior to different views or roles. Our approach builds on these insights but replaces handcrafted logic with agent-based specialization trained on partitioned data.

2.3. Specialization and Task Decomposition in Multimodal Learning

Recent work emphasizes decomposing multimodal reasoning into task-specific or role-specific components. This includes using separate models or attention heads for different question types, roles, or objects, which helps mitigate label noise and improve generalization. In the AI City context, teams applied template-driven prompts and segment-wise breakdowns to specialize model behavior [7, 8, 16]. While these systems did not formalize agent architectures, their successes validated the use of data-driven specialization. Our framework introduces semantically distinct agents, e.g., internal vs. external views, or pedestrian vs. vehicle descriptions, trained on curated supervision partitions. These agents are dynamically routed at inference time based on the validation performance of the agents, yielding modular performance improvements without architectural branching.

3. Methodology

3.1. Dataset, Input Structure, and Framework

All experiments were conducted on the AI City Challenge 2025 Track 2 dataset [9], which includes annotated traffic incidents split across two sources: the internal dataset (later referred to as WTS) contains both overhead and vehicle camera views of pedestrian-vehicle interactions in a smart-city setting in Japan, while the external dataset (later referred to as BDD) consists of ego vehicle views only from general US-city driving. These two domains differ significantly in camera angles and scene composition, view-point availability, and even the vehicle road driving side, motivating the need for specialized processing.

In sub-task 1 (Structured Captioning), the goal was to generate structured natural language captions describing either the pedestrian or the vehicle involved in the event. Each input included a video segment (frames) and optionally accompanying key-value derived content describing the agent-derived facts in the scene. In sub-task 2 (Question Answering), each training sample consisted of a sequence

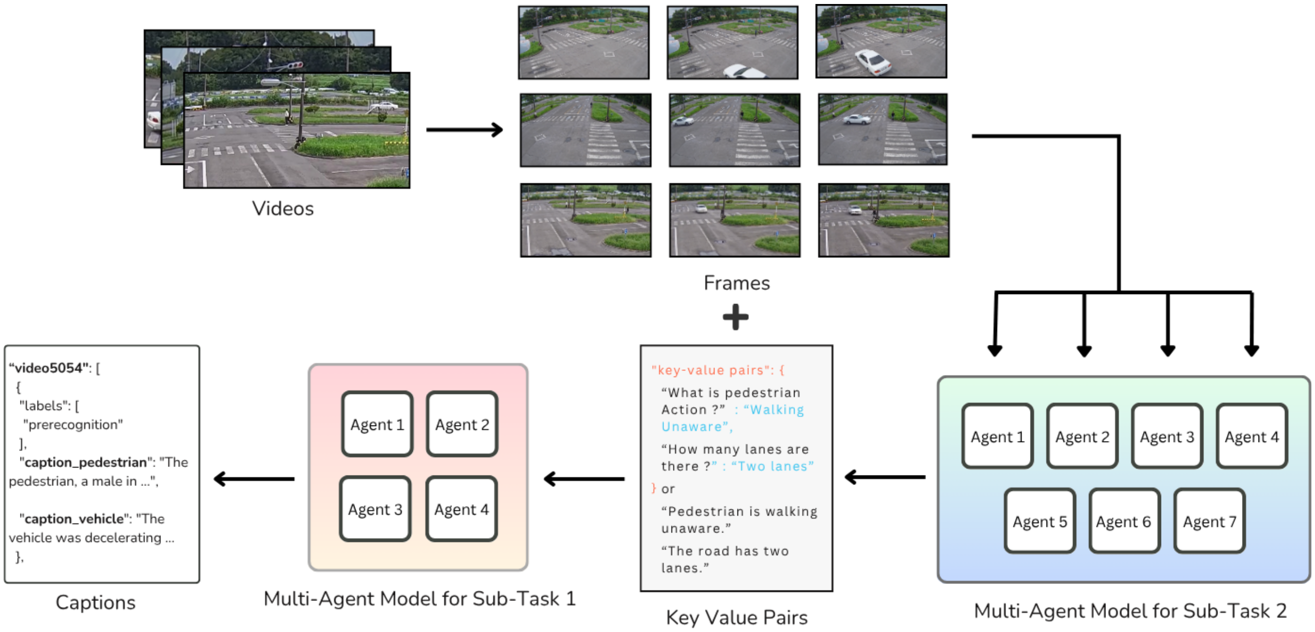

Videos
Frames
Agent 1
Agent 2
Agent 3
Agent 4
Agent 5
Agent 6
Agent 7
Multi-Agent Model for Sub-Task 2
"key-value pairs": {
"What is pedestrian Action ?" : "Walking Unaware",
"How many lanes are there ?" : "Two lanes"
} or
"Pedestrian is walking unaware."
"The road has two lanes."
Key Value Pairs
Agent 1
Agent 2
Agent 3
Agent 4
Multi-Agent Model for Sub-Task 1
"video5054": [
{
"labels": [
"prerecognition"
],
"caption_pedestrian": "The pedestrian, a male in ...",
"caption_vehicle": "The vehicle was decelerating ...
},
Captions

Figure 1. System overview. Video frames are first processed by a multi-agent QA model to answer 43 predefined questions. These outputs are combined with frames and passed to a multi-agent captioning model. The best agent per segment generates structured descriptions.

of extracted frames paired with one of 41 predefined challenge questions, covering pedestrian behavior, vehicle dynamics, environmental context, and spatial relations. For each question, an agent selected the correct answer from a fixed set of provided choices. The same question set was applied uniformly across both the internal and external datasets.

To support richer captioning supervision, based on descriptions we observed in the training and validation datasets, we introduced two additional question types that were not present in the original dataset: "What is the gender of the pedestrian?" and "What is the speed of the assailant vehicle?" These questions were added to both the training and validation QA sets, with answers inferred from the ground truth captions using rule-based extraction. For gender, we used keyword matching (e.g., "man", "woman") to assign binary labels. For vehicle speed, we parsed phrases like "25 km/h" and retained only values within a restricted, discrete set. During inference, these additional questions were included alongside the standard 41, expanding the QA task to 43 total questions.

Logically, our framework tackles the two sub-tasks of the challenge in reverse order, as noted in Figure 1. Using a variety of frames extracted from all available input videos for a scene, we employ a multi-agent approach to derive answers to all 43 questions, which we store as key-value pairs in a dictionary. These data are then provided to a second multi-agent model, along with the extracted frames, that produces the desired captions. In the following, we will detail our design of the two types of agents and their inputs.

3.2. Frame Extraction and Bounding Box Overlay

For both sub-tasks of the challenge, answers were requested for five distinct event phases in each video: *Prerecognition*, *Recognition*, *Action*, *Avoidance*, and *Judgement*. Start and end timestamps for each phase, along with bounding boxes for the person and vehicle of interest and other metadata, were provided by the challenge organizers.

To prepare input frames for vision-language model finetuning, we explored two sampling strategies tailored to the varying durations of event phases: evenly spaced sampling and midpoint-centric sampling. The evenly spaced approach (Algorithm 1) ensures broad temporal coverage by dividing each phase uniformly and extracting frames at regular intervals. In contrast, the midpoint-centric method (Algorithm 2) focuses on localized temporal cues by selecting frames centered around the midpoint with a fixed offset. The intuition behind the second strategy is that it may better capture movement dynamics than the first, as frames are sampled at a consistent cadence across all videos, leading to improved prediction of certain question categories such as the vehicle speed. While the second strategy proved more effective in short or dynamic segments, the first offered consistent coverage across variable-length phases. Figure 2 visually illustrates this strategy, showing how it captures distinct scene changes across the phase duration.

Evenly Spaced Sampling

In this approach, four frames are extracted at regular intervals across the duration of each phase. By computing fixed steps between the start and end timestamps, this method maintains consistency across samples while avoiding redundancy. It is especially effective for capturing gradual scene transitions or broad context in longer segments.

Algorithm 1 Evenly Spaced Frame Sampling

```
Require: Phase start $s$, end $e$
Ensure: Four evenly spaced frame indices
$T \leftarrow e - s$                    ▷ Total phase duration
$k \leftarrow \lfloor T/4 \rfloor$
for $i = 0$ to 3 do
    $f_i \leftarrow s + i \cdot k$
end for
return $\{f_0, f_1, f_2, f_3\}$
```

Midpoint-Centric Sampling

This approach emphasizes time correlation and better localized context capture. For each event phase, it selects the temporal midpoint and samples frames before and after using a step size $k = 10$. Fallback logic handles cases with insufficient length:

Algorithm 2 Midpoint-Centric Frame Sampling

```
Require: Phase start $s$, end $e$, step size $k$
Ensure: Three frame indices centered around the midpoint
$m \leftarrow s + \lfloor (e - s)/2 \rfloor$
if $m - k \geq s$ and $m + k < e$ then
    return $\{m - k, m, m + k\}$
else if $e - k \geq s$ then
    return $\{e - k - 1, e - 1\}$
else if $s + k < e$ then
    return $\{s, s + k\}$
else
    return $\{s, e - 1\}$
end if
```

This strategy more effectively captured localized context changes within a phase. However, smaller values of k (e.g., $k = 5$) often yielded visually redundant frames, while larger values reduced frame availability in shorter segments. Both strategies were used during model fine-tuning, and their comparative impact was evaluated on both sub-tasks.

Bounding Box Overlay

To improve visual grounding, all sampled frames were overlaid with bounding boxes marking the person or vehicle of interest (when available). Pedestrians were highlighted in red, and vehicles in blue. Bounding box availability varied by dataset and camera view:

- **WTS overhead views:** both pedestrian and vehicle boxes available.
- **WTS vehicle views and BDD vehicle views:** only pedestrian boxes available.

3.3. Prompt Engineering

We designed structured prompts tailored to the specific objectives of the two core tasks: *caption generation* and *multiple-choice question answering (QA)*. For captioning, the model was instructed to focus on a specific entity using visual cues and phase context. A representative prompt is:

> `<image> <image> ...`
> *You are given multiple images for the recognition phase of an accident scenario.*
> *If red and blue boxes are present, the red box highlights the pedestrian and the blue box highlights the vehicle.*
> *Generate only the vehicle description using all the available visual cues and associated facts.*
> {*Facts*}

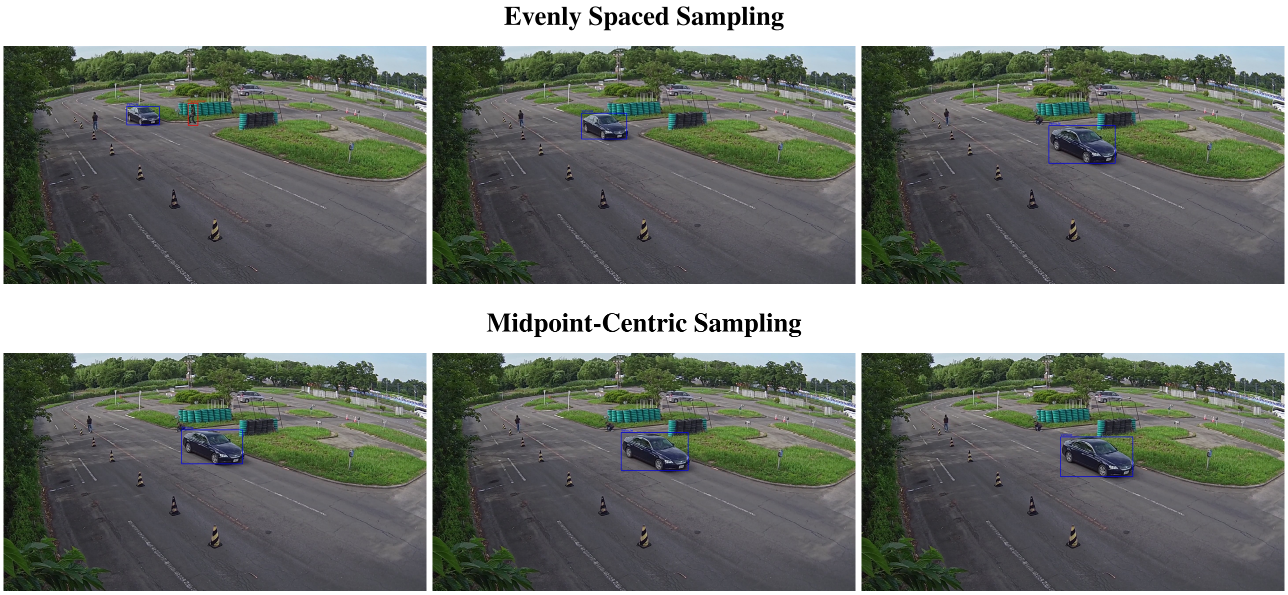
This prompt explicitly specifies the event phase (recognition) and relies on bounding box annotations for spatial grounding. For QA, prompts were designed to provide both the event phase and the viewpoint perspective to guide accurate answer selection. A representative prompt is:

> `<image> <image> ...`
> *This accident scenario is in the avoidance phase.*
> *Images are from the vehicle's front view (assailant's perspective).*
> *If red and blue boxes are present, the red box highlights the pedestrian and the blue box highlights the vehicle.*
> {*Question with multiple options*}
> *Answer with the option's letter from the given choices directly.*

This structure grounds the model in both temporal context (e.g., avoidance phase) and spatial perspective (e.g., vehicle or overhead view). For the WTS dataset, both overhead and vehicle views were used, whereas for BDD, only vehicle views were available. By standardizing prompts across datasets and aligning them with the task objectives, we ensured consistent model behavior and improved cross-domain generalization.

3.4. Model Architecture

All agents in our system for both captioning and question answering were built on top of the InternVL-14B vision-language model [5]. This model follows the ViT-MLP-LLM paradigm and integrates a frozen vision encoder (EVA-CLIP) with an instruction-tuned language decoder, enabling multi-frame image processing and structured textual reasoning. InternVL-14B was selected as the unified backbone for all agents due to its strong empirical performance across both sub-tasks.



Evenly Spaced Sampling

Midpoint-Centric Sampling

Figure 2. Visual comparison of two frame sampling strategies for an avoidance phase in a WTS dataset sample.

The architecture supports dynamic resolution input: each input frame was resized to 896×1344 and divided into 448×448 pixel tiles. A pixel unshuffle operation reduced visual token density, improving efficiency without sacrificing resolution. These operations, which align with intended strategies in our project, were introduced in InternVL 1.5 and extended in InternVL3.

While our training used the InternVL-14B network, we borrowed architectural advances from InternVL3 such as Variable Visual Position Encoding (V2PE) and multi-image/video input support to contextualize the strengths of the underlying vision backbone. V2PE enables finer spatial encoding granularity, while multi-image processing supports richer spatial-temporal understanding, which is key for traffic scene comprehension.

All downstream agents in our system shared this InternVL-14B backbone but differed in training supervision: some were fine-tuned on internal (WTS) or external (BDD) data, others were grouped by semantic question type, and captioning agents varied by subject role (pedestrian vs. vehicle) and supervision format (key-value vs. facts). At inference time, validation-based routing dynamically assigned inputs to the most appropriate agent.

3.5. Role and Domain Specialization

We partitioned training configurations by entity role in captioning and by dataset domain in question answering, enabling agents to learn from semantically or visually coherent subspaces and reducing interference between distinct reasoning styles.

In the captioning task, we trained separate models for pedestrian and vehicle captions. This decision was informed by a semantic analysis of the training labels: pedestrian captions were more tightly aligned with environment-related and pedestrian-specific key-value pairs (e.g., position on crosswalk, gaze direction, attention), while vehicle captions reflected motion-related or avoidance behavior. The two caption types emphasized different visual features and descriptive styles. By training agents on a single role, we enabled role-specific modeling.

In the QA task, we trained separate agents on the internal (WTS) and external (BDD) datasets due to their distinct recording configurations. The WTS dataset [9] included overhead and vehicle views, often structured and centered around key interactions in staged scenarios. In contrast, BDD contained only vehicle views from naturalistic driving, with high variability in framing and lighting. These domain differences impacted the visual context. Training separate agents per domain allowed specialization in interpreting their unique perspectives.

3.6. Semantic Grouping in Question Answering

To improve performance on visually entangled questions, we introduced a specialized QA agent trained on a grouped subset of 14 low-performing questions. These involved nuanced pedestrian and vehicle behaviors like gaze, line of sight, and relational motion requiring subtle, multi-frame reasoning.

We began by training a base InternVL-14B model on all 41 questions. After validation, we identified questions with low accuracy and grouped them based on shared semantic traits. A new InternVL-14B model was then retrained using only this subset, preserving architecture and input format.

By constraining supervision to a focused domain, this model developed sharper attention to cues required for answering difficult questions. The grouped QA agent was included in the inference-time selection pool (Section 3.8).

3.7. Fact Augmented Conditioning in Captioning

To enhance semantic grounding and improve scene understanding, we explored an input conditioning strategy that augments vision-language model inputs with structured factual statements. These statements were generated by converting question-answer (QA) annotations into natural language using fixed templates.

For example, the QA pair, "*Q: What is the age group of the pedestrian? A: 30s*" was converted into the statement "*The pedestrian is in the 30s age group.*".

This process was applied across all safety-relevant attributes, including road conditions, pedestrian orientation, visual awareness, obstacle geometry, and environmental features. Some additional fact-style statement examples include

```
"The pedestrian is squatting."
"The road surface is wet."
"There are sidewalks on both sides
of the road."
```

These fact-enhanced prompts were prepended to visual tokens and proved particularly useful for internal pedestrian captioning, where multi-view inputs and dense annotations enabled strong semantic alignment. Inspired by fact-based prompting strategies in recent VLMs [2, 10], our method guides the model toward more explicit and accurate descriptions.

3.8. Agent Selection

Given the diverse agent configurations explored across both captioning and QA tasks, we employed a performance-driven agent selection strategy to choose the best model per output unit.

For QA, selection was done per question. Each QA model was evaluated independently, and for each of the 43 questions, we selected the model with the highest validation accuracy for that question. We evaluated candidate agents on a held-out validation set using accuracy and chose the highest performing agent among all the trained agents for each question.

For captioning, agent selection was done independently for each of the four output segments: internal pedestrian, internal vehicle, external pedestrian, and external vehicle captions. We evaluated candidate agents on a held-out validation set using BLEU-4 [12], METEOR [6], ROUGE-L [11], and CIDEr [14]. The agent with the highest average score was selected per segment.

This selection framework allowed us to adaptively combine the strengths of specialized agent models and improved overall robustness across tasks.

4. Experiments

4.1. Fine-Tuning Setup

All individual agents for both sub-tasks were built on top of the InternVL3-14B vision-language model [5]. Each agent was trained independently using *full-parameter fine-tuning*, which included updating the weights of the LLM decoder, MLP layers, and the vision encoder backbone. This level of adaptation allowed agents to specialize deeply in their respective domains and reasoning tasks.

Training was performed using distributed infrastructure with DeepSpeed ZeRO Stage 2 optimization, enabling efficient memory management and scalability for large-scale vision-language models. All models were trained in mixed precision mode with automatic bf16 or fp16 casting on a Supermicro SYS-420GP-TNAR+ system equipped with NVIDIA HGX A100 8-way GPUs (80 GB RAM each) running Rocky Linux 9.4 (Blue Onyx).

4.2. Evaluation Setup

We evaluated each agent on the validation and test splits using the metrics specified by AI City Challenge Track 2: BLEU-4, METEOR, ROUGE-L, and CIDEr for captioning, and classification accuracy for question answering. Agent selection is performed based on validation performance.

4.3. Unified Baseline

While unified multitask learning is common in large vision-language models (VLMs), we found that a single model trained jointly on both captioning and QA struggles to balance the conflicting objectives of descriptive generation and categorical reasoning. The unified baseline learns broad representations but lacks task-specific specialization.

To validate this, we trained a unified InternVL-14B model on a mix of both captioning and QA data using shared prompts. During training, the model was exposed to both types of inputs randomly. However, as shown in Table 1, this unified model underperformed significantly across both tasks: it achieved only 23.47 average score on captioning and 81.43% QA accuracy.

In contrast, our **multi-agent framework routes each sub-task to a specialized agent** trained using task-pure supervision and selected via validation resulting in a consistent performance gain across both modalities. These results confirm that modular specialization is critical for effectively handling complex multimodal reasoning tasks in traffic video understanding.

Table 1. Validation Performance of the Unified Baseline Compared to Our Modular Agent-Based Framework

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

Table 2. Validation Accuracy (%) for Individual QA Agents Across Sampling Strategies and Data Partitions vs. Our Multi-Agent Model

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**

Table 3. Agent Selection Results for Captioning

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**

We report the average score across BLEU-4, METEOR, ROUGE-L, and CIDEr on the validation set for each captioning segment. Each row represents an InternVL3-14B-based agent variant with different input configurations. Best-performing agents per segment (bolded) are selected for the final system described in §3.8.

4.4. Question Answering Effectiveness

We evaluated question answering (QA) performance using question-wise classification accuracy on the validation set. Each question was routed to the agent that performed best for that specific query type, allowing the system to leverage agent-specific strengths.

As shown in Table 2, this routing mechanism led to consistent improvements across both datasets: the multi-agent model achieved 84.31% on WTS and 87.46% on BDD, compared to the best single-agent scores of 81.75% and 86.37% respectively, for an overall gain from 85.90% to 87.14%. These improvements were observed even though the highest-scoring individual configurations differed between datasets and frame strategies, showing that validation-guided routing successfully captures complementary strengths.

Notably, questions involving pedestrian awareness, gaze direction, and relative motion benefited from the specialized QA agent trained on grouped low-performing queries as in §3.6. By narrowing supervision to a focused domain, this agent captured subtle visual cues that are often overlooked by generalist models.

These results affirm the importance of modular specialization in safety-oriented traffic video understanding. Rather than relying on a single model to generalize across all questions and domains, our validation-guided approach enables more accurate and context-aware reasoning through tailored agent assignment.

4.5. Captioning Effectiveness

Results from the QA model are then processed to be used as input to captioning models either in the form of key-value pairs or fact-based sentences. As exemplified in Table 3, we observed that role-specific models trained solely on pedestrian captions outperformed those trained jointly on both roles in some cases. For example, the model trained specifically for internal pedestrian captions achieved a score of 32.65 versus the model trained on both roles only obtained a score of 31.49, suggesting it was better able to focus on attributes like pose and gaze. Fact-based inputs derived from QA annotations also yielded superior CIDEr and METEOR scores compared to raw key-value supervision, for example +1.14 for external vehicles (44.45 vs. 43.31) and +0.96 for external pedestrians (31.33 vs. 30.37), indicating improved semantic fluency and alignment in generated captions.

We also examined the impact of input representation. Using the method described in §3.7, QA annotations were converted into natural language statements to form fact-based inputs. These inputs improved semantic grounding and fluency, particularly in external-view segments, where diverse visual content benefits from richer contextual language. The gains observed suggest that fact-based conditioning enhances alignment with ground-truth semantics under less constrained viewpoints.

When we analyzed the effect of role specific modeling, we found that models trained jointly on both pedestrian and vehicle captions consistently performed better than those trained on only one role in several configurations, indicating that combining supervision across roles helps the model learn more comprehensive visual-contextual relationships, such as pedestrian-vehicle interactions and scene-level cues, ultimately leading to stronger captioning performance.

Figure 3 shows a representative example of the generated captions. The outputs reflect strong grounding in role-specific semantics, capturing a range of attributes such as pedestrian posture and awareness, vehicle dynamics, environmental context, and safety indicators. Color highlighting is used to emphasize different semantic categories without any post-editing or handcrafted templates.

Finally, the best results are achieved through a multi-agent approach, where each agent is selected based on its performance on a specific segment. This paradigm leverages the strengths of different frame sampling strategies and

Pedestrian caption: The pedestrian, a male in his 30s with a height of 160 cm, was wearing a black T-shirt and black slacks. He was standing directly in front of a vehicle on a residential road with two-way traffic. The pedestrian's body was oriented in the opposite direction to the vehicle, and his line of sight was focused on a smartphone that he held in his hand. Although closely watching his surroundings, he seemed unaware of the vehicle. The pedestrian was moving slowly, going straight ahead, and traveling in a car lane, despite being far from the vehicle. The weather was clear, but the brightness was dim. The road surface was dry with level incline, made of asphalt. The traffic volume was light, and there were no street lights. Overall, the pedestrian's action and the environmental conditions suggested a potentially risky situation due to the pedestrian's lack of awareness of the vehicle and the absence of certain safety features.

Vehicle caption: The vehicle is positioned in front of the pedestrian, and the relative distance between them is far. The vehicle's field of view allows it to see the pedestrian clearly. The vehicle is going straight ahead at a speed of 20 km/h. The environment condition includes a male pedestrian in his 30s with a height of 160 cm. He is wearing a black T-shirt and black slacks. The weather is clear, with dim brightness. The road surface is dry and level, made of asphalt. There are no roadside strips, but there are street lights along the road.

Figure 3. Sample pedestrian and vehicle captions generated by our multi-agent framework.

input types, resulting in the highest overall captioning score. These findings support the effectiveness of agentic modeling, where modular specialization enables better adaptation to the visual and semantic characteristics of each sub-task.

4.6. State-of-the-Art Comparison

The performance trends observed in Table 2 and Table 3 show that our multi-agent system consistently outperforms a unified baseline trained jointly across all data. This performance gap, seen across both QA accuracy and captioning metrics, supports our design choice to use specialized agents for different segments and questions.

As a way to compare our model against other state-of-the-art approaches, we submitted test set inference results using our multi-agent framework to Track 2 of the 2025 AI City Challenge. As shown in Table 4, our method was able to achieve 2nd place in the challenge, confirming the effectiveness of our agentic framework in a competitive setting.

Table 4. Test Set Effectiveness Comparison

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

We show 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.

5. Conclusion

In this work, we presented a modular, validation-driven multi-agent system for structured captioning and fine-grained safety question answering in multimodal traffic videos. By decomposing the problem into role- and question-specific subtasks and leveraging specialized vision-language agents trained on targeted supervision, our framework achieves superior performance compared to monolithic models. Key design strategies such as fact-based input conditioning, role-aware training, and domain-specific agent specialization proved effective in capturing the complex visual and contextual cues inherent in real-world traffic footage. The competitive performance of our system at the 2025 AI City Challenge validates the agentic approach and opens pathways for future research into modular, context-sensitive VLM architectures for real-world traffic intelligence.

Acknowledgments

This work was made possible in part by the National Science Foundation under Grant #2442814. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. GPU hardware for our research was provided by NVIDIA and Supermicro.

References

[1] AI City Challenge Organizers. Ai city challenge 2025 track 2. https://www.aicitychallenge.org/2025-track2/, 2025. 1

[2] Jean-Baptiste et al. Alayrac. Flamingo: A visual language model for few-shot learning. *NeurIPS*, 2022. 2, 6

[3] David C. Anastasiu. Explainable ai for real-time video anomaly anticipation. *SIAM International Conference on Data Mining (SDM)*, 2025. 1

[4] Yuxiao et al. Bai. Qwen-vl: A versatile vision–language model. *arXiv preprint arXiv:2308.12966*, 2023. 2

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2, 4, 6

[6] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, 2014. 6

[7] Quang Minh Dinh, Khoa Nguyen, Phong Nguyen, and Minh-Triet Nguyen. Trafficvlm: A controllable visual language model for traffic video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[8] Zhizhao Duan, Yujia Zheng, Ying Fan, Yuxuan Li, Menglin Xu, Xiaofei Zhang, Yixuan Li, Xinlong Wang, Liang Wang, and Han Hu. Cityllava: Efficient fine-tuning for vlms in city scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[9] Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, David C Anastasiu, Yoichi Sato, and Norimasa Kobori. Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding, 2024. 2, 5

[10] Junnan et al. Li. Blip-2: Bootstrapping language-image pre-training. *ICML*, 2023. 2, 6

[11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81, 2004. 6

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[13] Hugo et al. Touvron. Llama: Open and efficient language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[14] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 6

[15] Zhaoyang et al. Wang. Internvideo: Scaling video-language pretraining. *arXiv preprint arXiv:2306.00988*, 2023. 2

[16] Khai Trinh Xuan, Ngoc-Trung Nguyen, Thanh-Dat Nguyen, Binh-Son Hua, and Minh-Triet Nguyen. Divide and conquer boosting for enhanced traffic safety description and analysis with large vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[17] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1