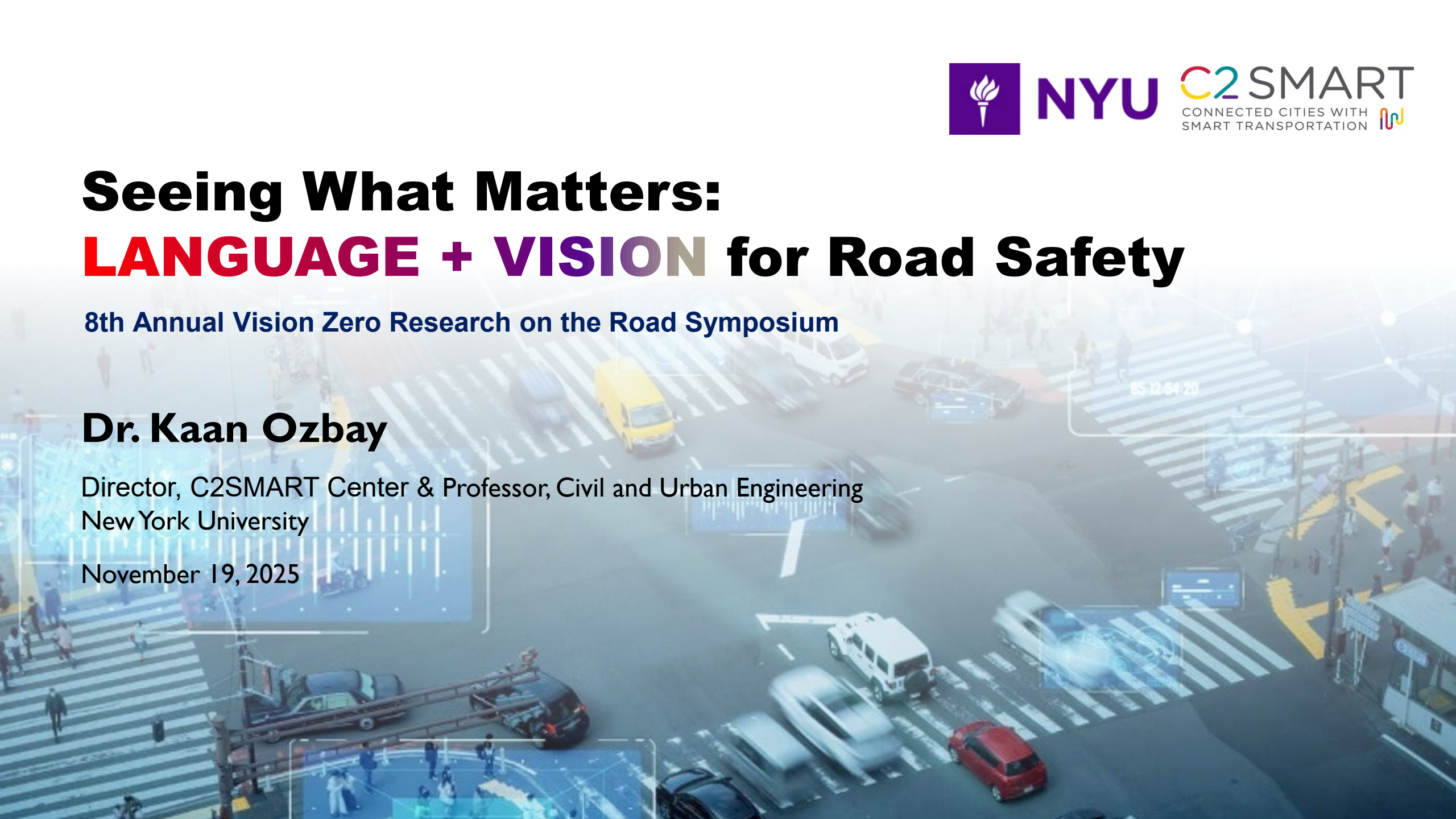
Seeing What Matters:
LANGUAGE + VISION for Road Safety

8th Annual Vision Zero Research on the Road Symposium



Dr. Kaan Ozbay

Director, C2SMART Center & Professor, Civil and Urban Engineering
New York University

November 19, 2025

Our Team

Speaker

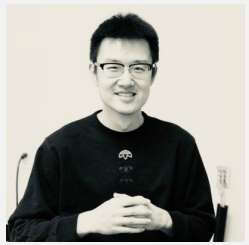
Dr. **Kaan Ozbay**
Director & Professor
C2SMART Center & PI@UrbanMITS Lab
New York University
kaan.ozbay@nyu.edu

Contributors

Ruixuan Zhang
Ph.D. Candidate
New York University

Beichen Wang
M.S. Student
New York University

Juexiao Zhang
Ph.D. Candidate
New York University



Dr. **Zilin Bian**
Assistant Professor
Rochester Institute
of Technology

Dr. **Chen Feng**
Institute Associate Professor
PI@AI4CE Lab
New York University

A Data-Rich Era

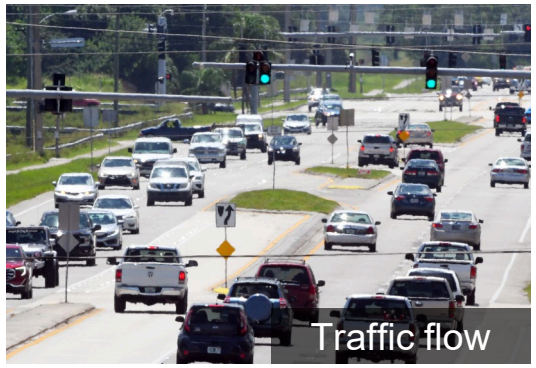
We are in a data-rich era. Beyond traditional sources, vast potential can be mined from our existing camera infrastructure.

Traffic Sensing Technology Evolution

1933	1960s	1970s	2000s	2010s	Now
					
Greenshields taking measurements	Inductive Loop Detectors	Microwave Radar	Traffic Cameras	Connected and Automated Vehicles	Edge devices w camera/LiDAR

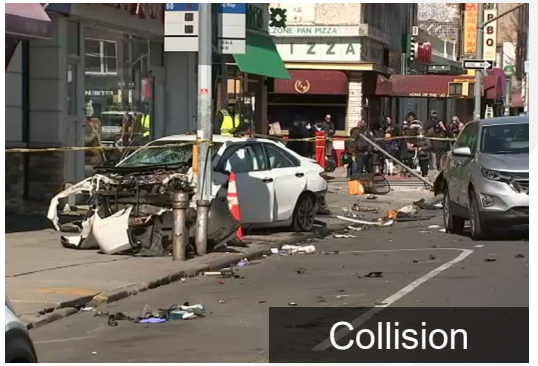
The Paradox: More Data ≠ More Insight

- The growing number of sensors is generating more data than ever before.
- Human operators have limited processing bandwidth.
- Vision data is gaining attention for its rich context and ability to reveal deeper insights.


Traffic flow

Work zone

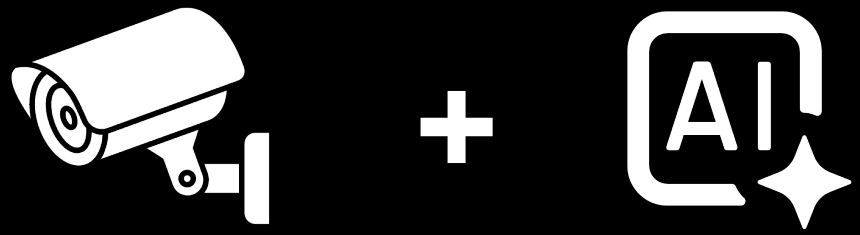

Flooded road


Collision

Beyond Traditional Data: The Power of Video & AI for Traffic Safety

AI turns video from watching traffic to understanding it

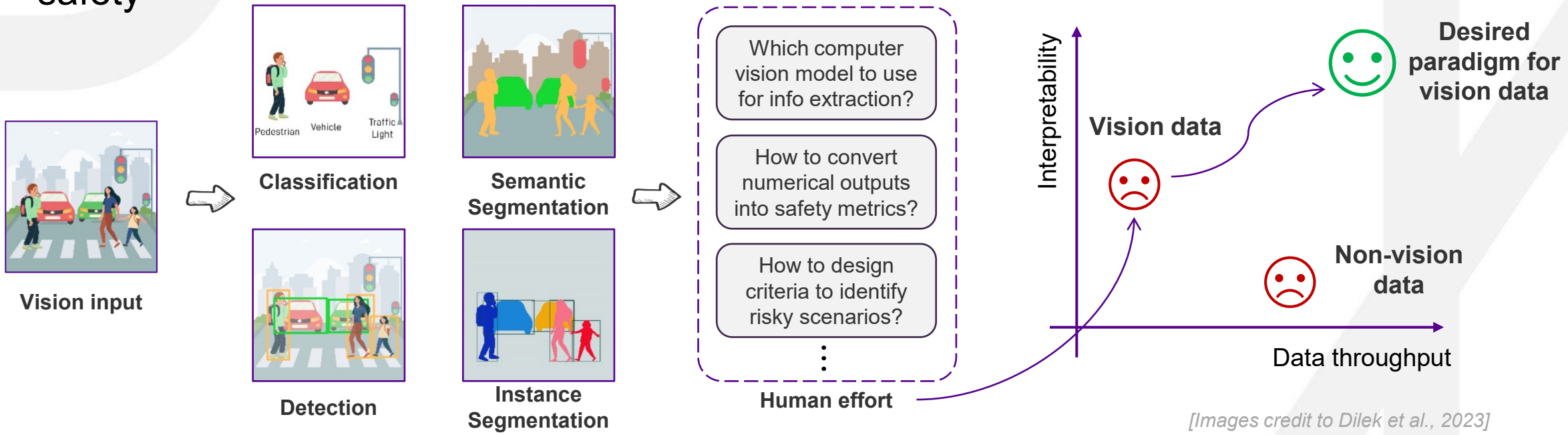
- Proactive Safety - See the “Near Misses”
- Understand Complex Behaviors
- Real-time Insights & Dynamic Response



Source: Google Gemini

The Power of Video & AI for Traffic Safety

- AI-powered image processing can “see” fine details and patterns in every frame
- Yet most systems still focus on single tasks, not the full context of traffic behavior
- **The challenge:** scaling from isolated algorithms to integrated intelligence for traffic safety

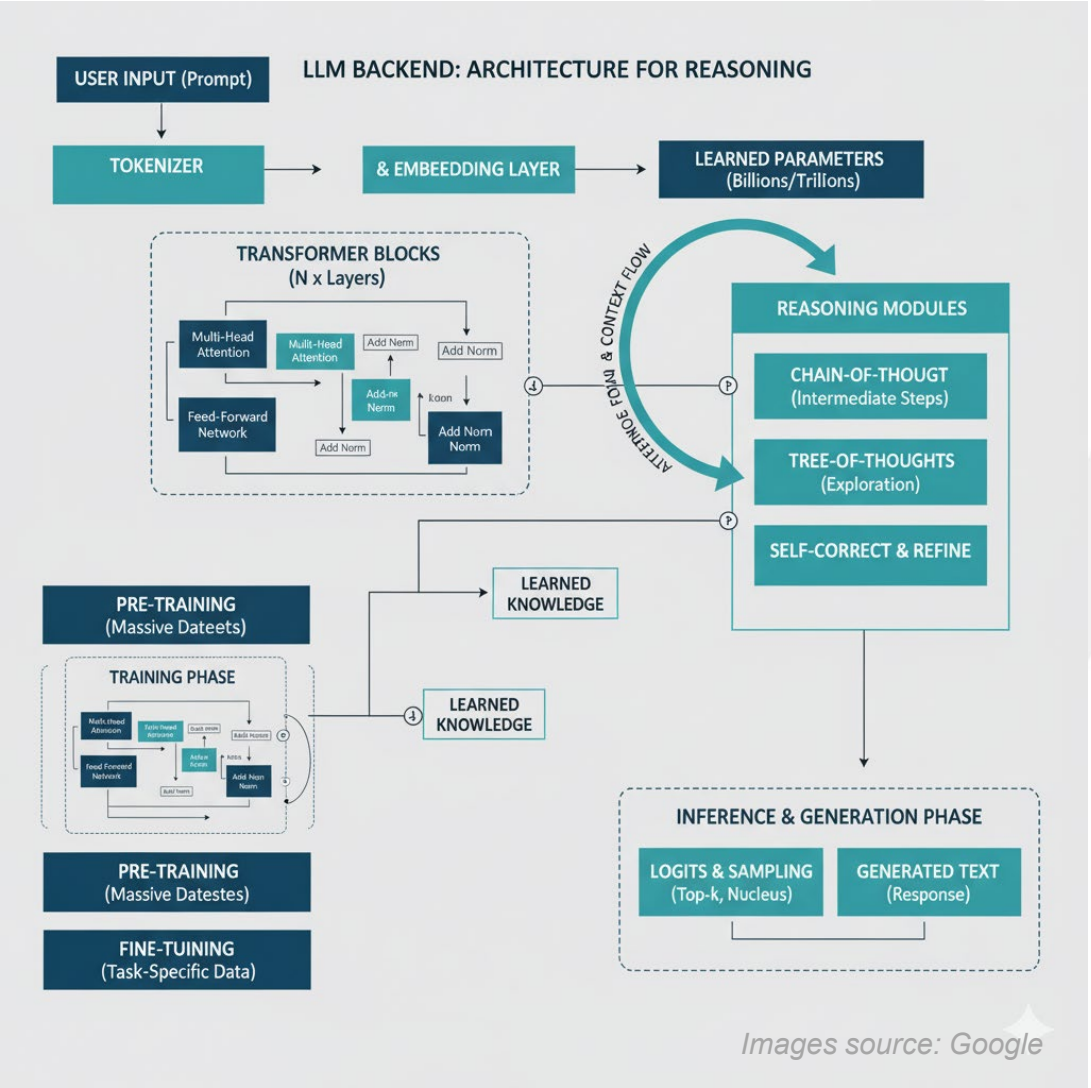


[Images credit to Dilek et al., 2023]

The Core Technology: What is Generative AI & Large Language Models?

- **Large Language Models** (LLMs) are powerful **AI** models that learn from **massive text** datasets to understand and produce ***human-like language***.

- **The Logic Behind LLMs:**
  - Uses TRANSFORMMER architecture (a complex neural network)
  - Process vast text data to identify patterns and relationships
  - Learn word sequence probabilities
  - Generate human-like text



Images source: Google

The Core Technology: What is Generative AI & Large Language Models?

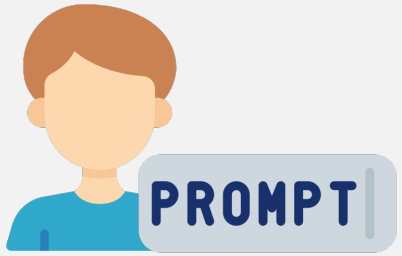
ChatGPT Gemini Claude
Command R LLaMA

Core Functions

- Text generation
- Text summarization
- Chatbots
- Code generation
- Sentiment analysis
- Translation

How it works?

- **Input:** You give the LLM a prompt.
- **Processing:** The model uses patterns learned to predict the next word/phrase.
- **Output:** The result is a coherent, often insightful answer that feels human-like.



Prompt: Generate a roadside message urging drivers to slow down in the rain.

Large Language Model: Analyzes traffic, weather, and driver behavior data to craft safety-focused messages.

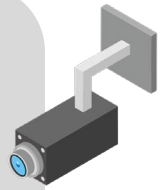
Output: *"Smart drivers slow down when it rains — thank you for being one!"*

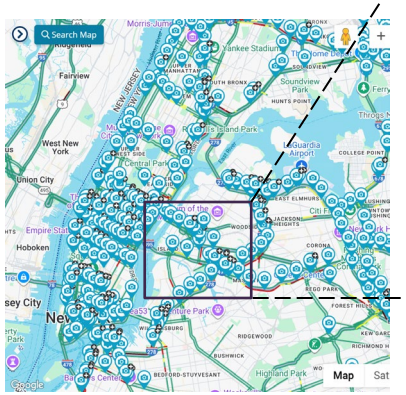
NYU | C2SMART

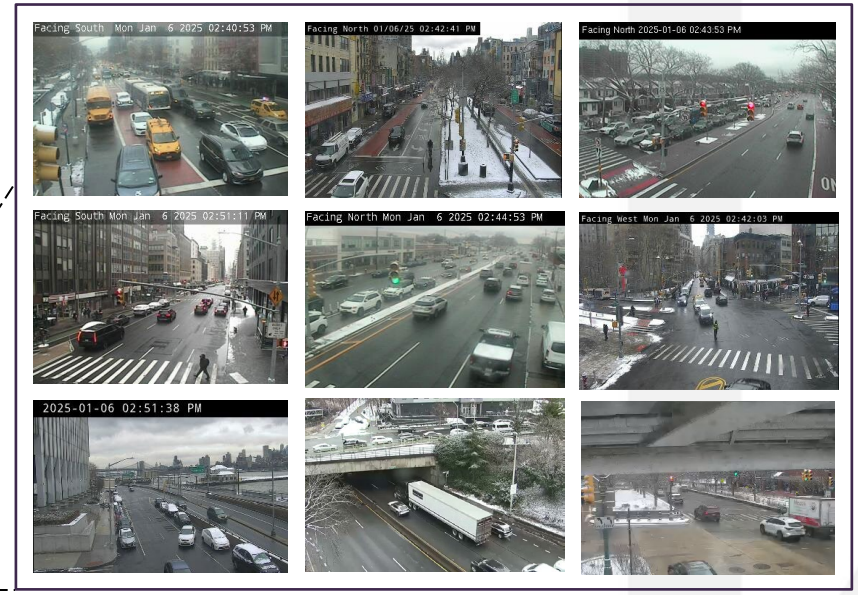
Gen-AI and LLM for Traffic Safety

Given existing thousands of traffic camera streaming 24/7:

1. How to find **rare but safety-critical** traffic incidents in large-scale video streams?
2. How to prioritize **human decision-making** over manual data processing?







User > *Please find videos of pedestrian-vehicle conflict given this region in the past month.*
System > *There are <NUMBER> qualified videos…this conflict is from camera <ID> at <TIME>. The weather was rainy… The white truck was too close to the pedestrian in black…*

…

User > *Please export to files…*
System > *…*

The Breakthrough: Language-Vision Models

- Vision and language are naturally favored by humans.
- A **Language-Vision Model,** one type of **multimodal LLM (MLLMs)**, combines the ***reasoning of an LLM*** with the ***'eyes' of a computer vision system***.

Language understanding
(text and speech)

+

Vision understanding
(images or videos)

Example: MLLM for explainable autonomous driving

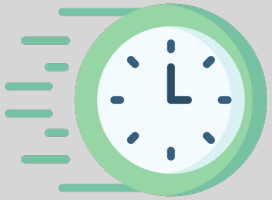
OmniDrive *[Wang et al., 2024]*

SeeUnsafe: An MLLM-empowered Framework for Road Safety

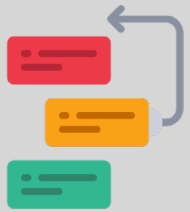
We propose An MLLM-empowered Framework **SeeUnsafe** to enhance traffic incident management and risk detection.

What it does: Identify risky events in traffic footage by combining language reasoning with visual intelligence

SeeUnsafe:
A New Framework for Proactive Safety, turning passive video feeds into an active, automated safety monitoring system.



Deeper insights over time
Recognizes risk that might be missed in short clips

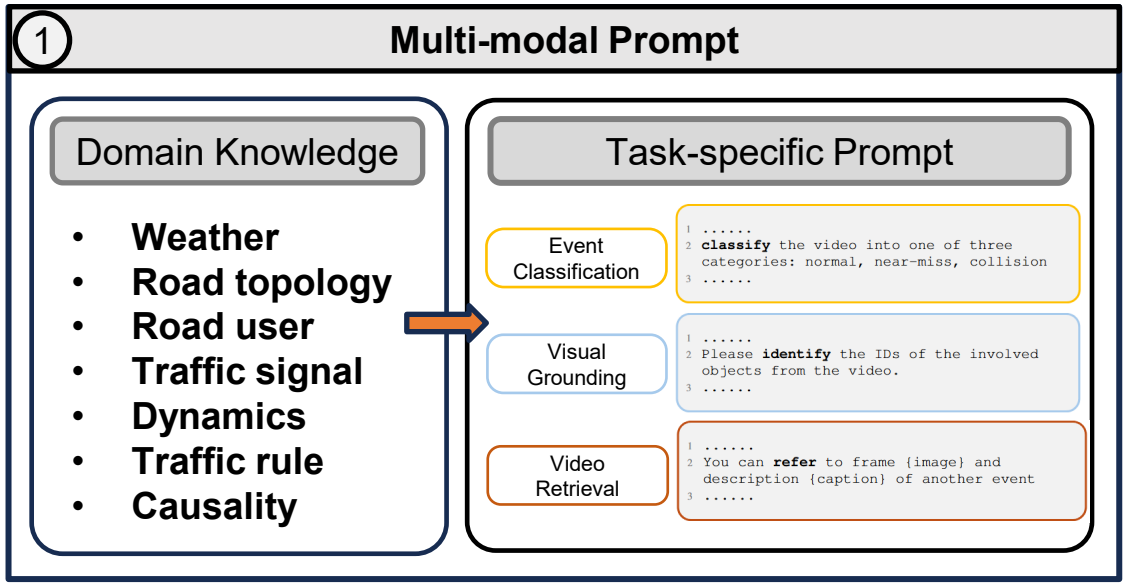


Smarter Prioritization
Groups risky moments by severity

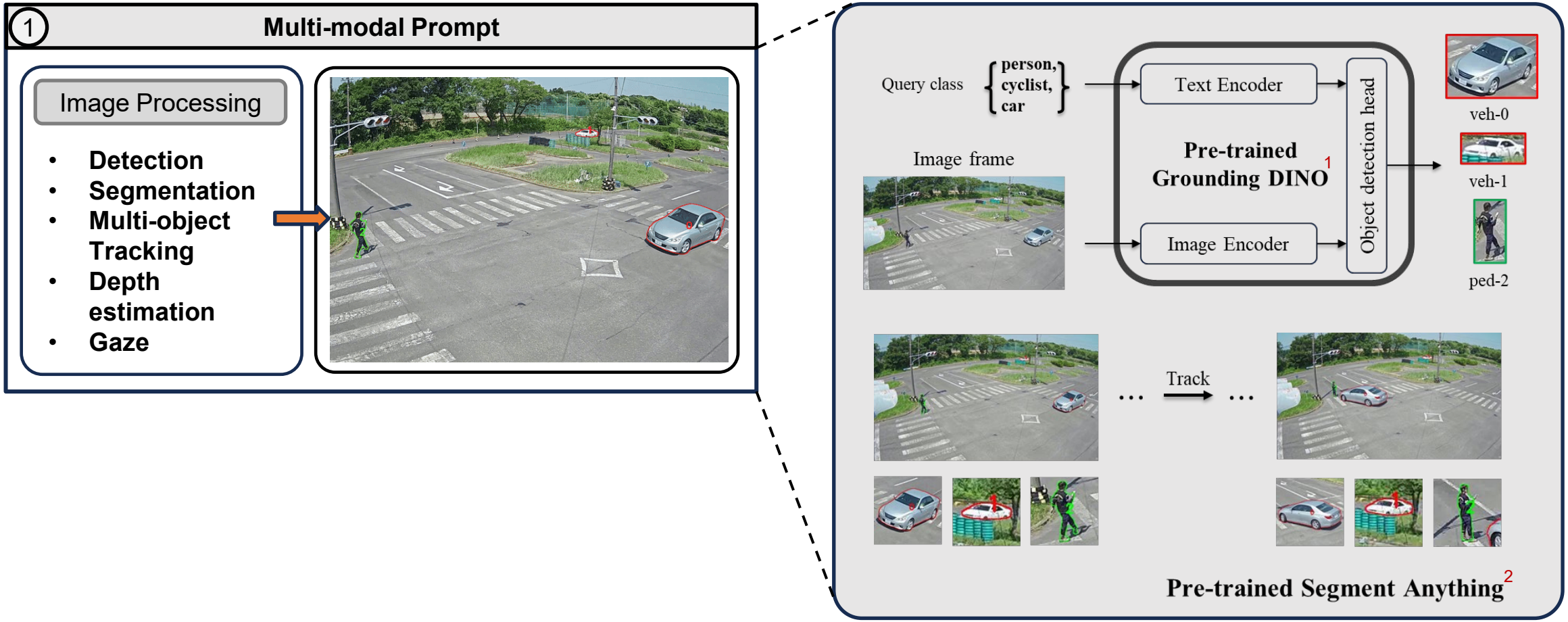
Beyond scene detection
Provide fine-grained analysis on objects and behaviors

Reference: *Zhang, R., Wang, B., Zhang, J., Bian, Z., Feng, C. and Ozbay, K., 2025. When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis. Accident Analysis & Prevention, 219, p.108077.*

SeeUnsafe : An MLLM -empowered Framework for Road Safety



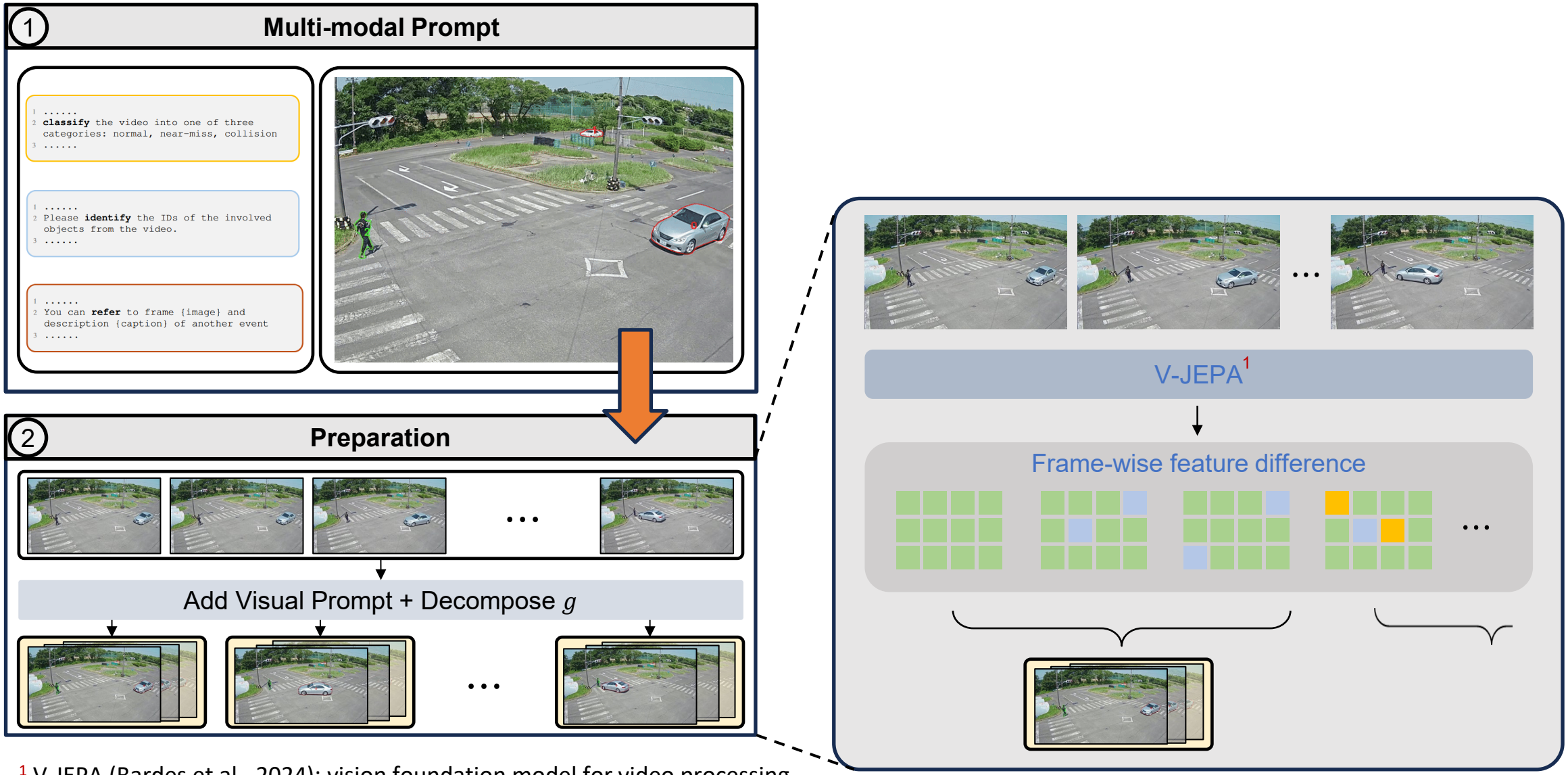
SeeUnsafe : An MLLM -empowered Framework for Road Safety



[1] Grounding DINO (Liu et al., 2024): open-set object detection

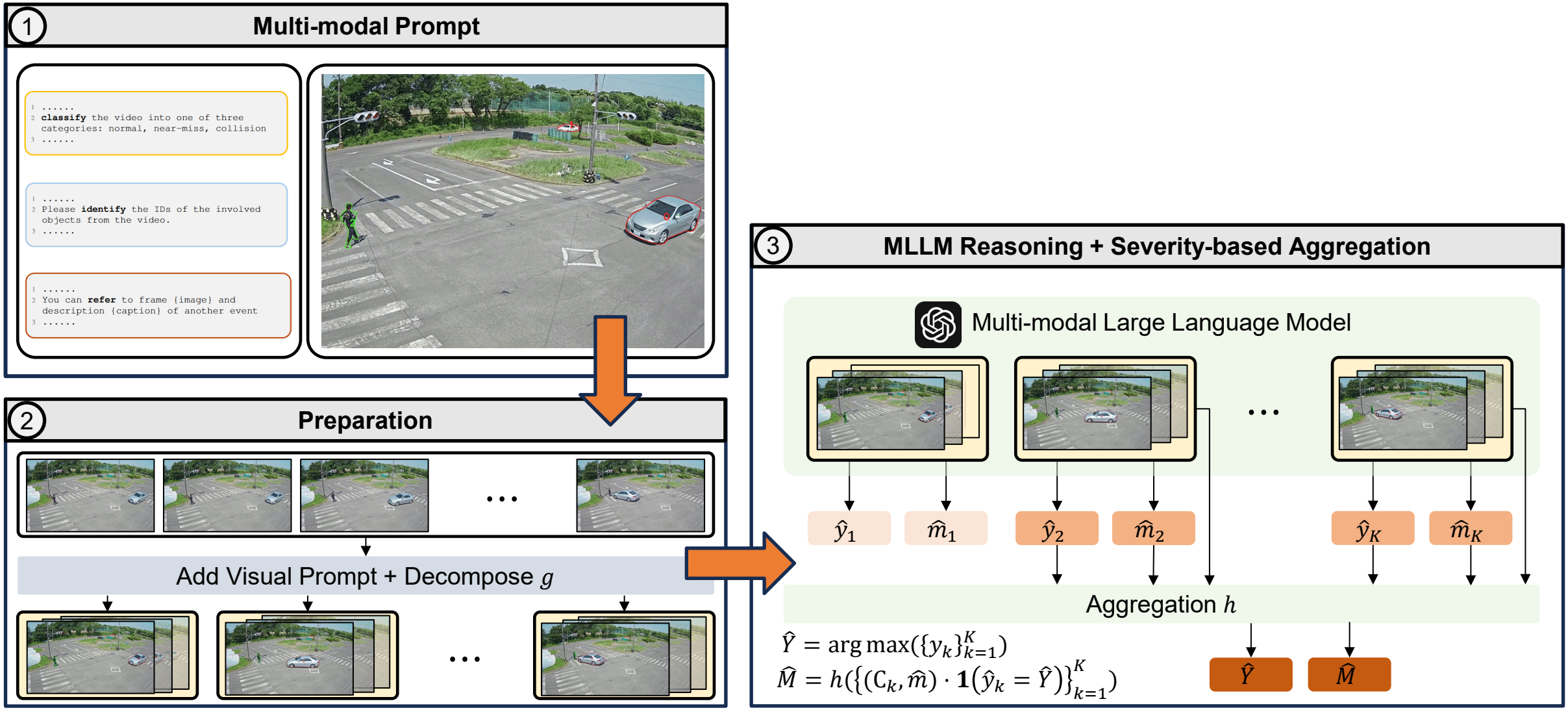
[2] Segment Anything (Kirillov et al., 2023): vision foundation model for segmentation

SeeUnsafe : An MLLM -empowered Framework for Road Safety

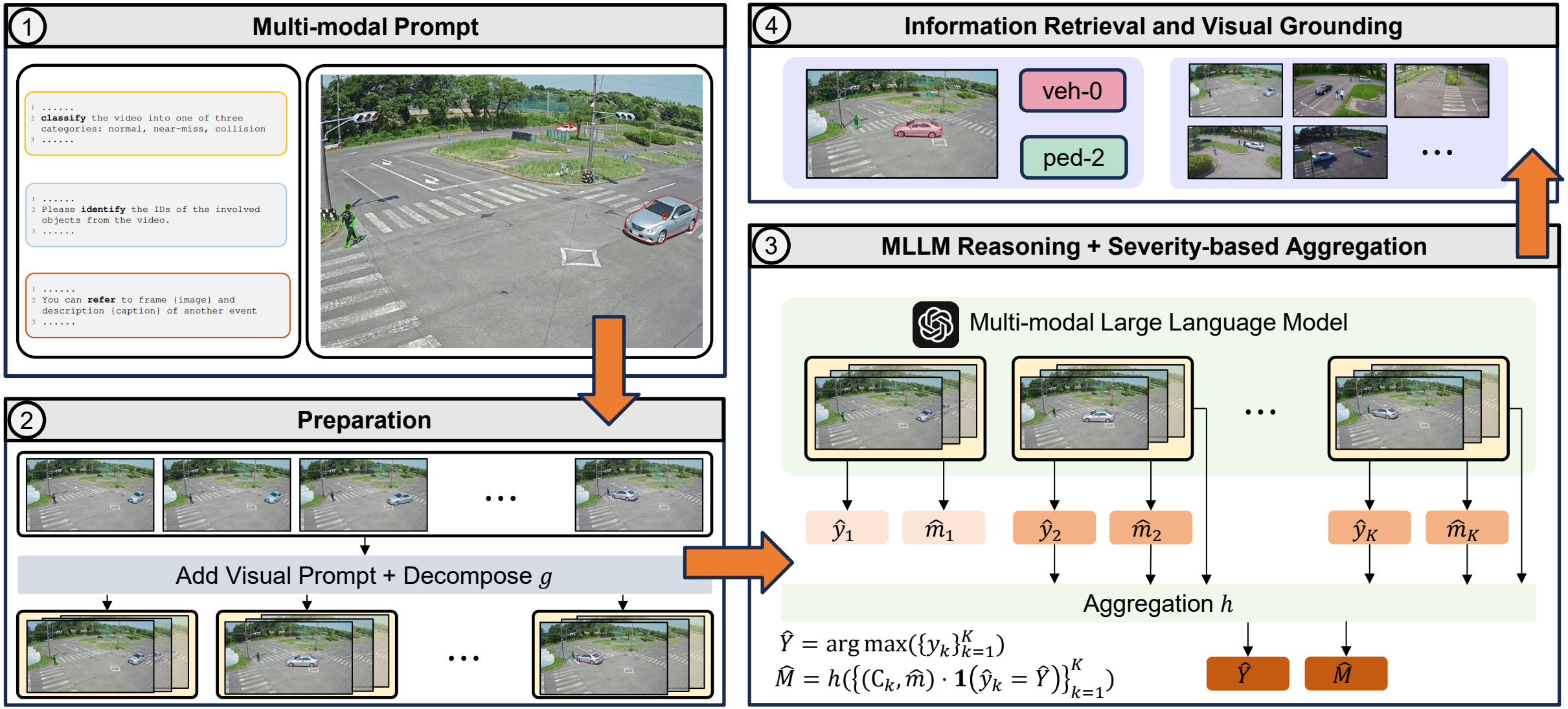


[1] V-JEPA (Bardes et al., 2024): vision foundation model for video processing

SeeUnsafe : An MLLM -empowered Framework for Road Safety



SeeUnsafe : An MLLM -empowered Framework for Road Safety



SeeUnsafe : An MLLM -empowered Framework for Road Safety

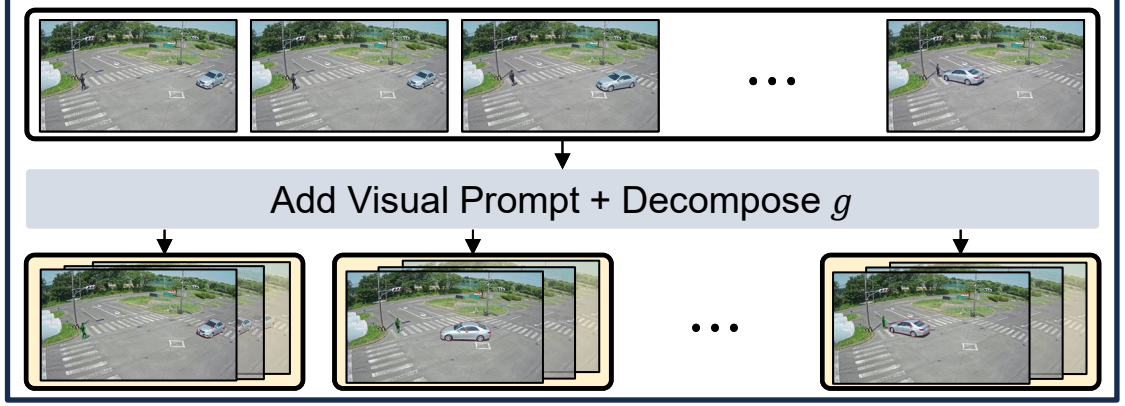
1 Multi-modal Prompt

```
.....
classify the video into one of three categories: normal, near-miss, collision
.....
```

```
.....
Please identify the IDs of the involved objects from the video.
.....
```

```
.....
You can refer to frame {image} and description {caption} of another event
.....
```

2 Preparation



5 Structured Output

Event class:
Near-miss

Scene context:
Sunny day with dry road surface at an intersection, under daytime lighting with light traffic.

Object description:
A male pedestrian and a silver sedan

Justification:
The pedestrian is in the crosswalk and the sedan is turning. The pedestrian and the vehicle come extremely close to each other, with the vehicle making a sudden stop to avoid a collision

Visual grounding:
car id: {0}, person id: {2}

Event class (severity)
Normal: 0
Near-miss: 1
Collision: 2

Scene context
Description for static environment, such as weather, road condition, traffic condition, etc.

Object description
Description for object type and appearance

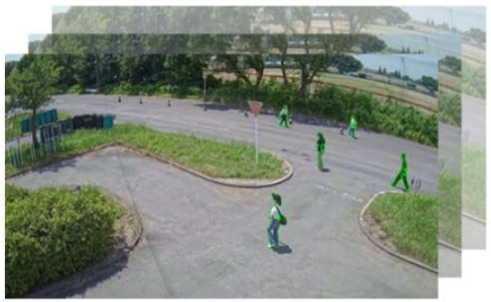
Justification
Description for object dynamics and possible reason for classified event class.

Visual grounding
object class: {ID #}

Case Example

- We evaluated SeeUnsafe on Toyota Woven Traffic Safety (WTS) Dataset.
- In this example, we successfully reasoned the collision with consistent scene context and object descriptions using SeeUnsafe. The justifications correctly describe the object dynamics.

Ground truth: Collision.



Event class: Normal
Scene context: Sunny day with dry road surface in a curbside area near an intersection, under daytime lighting with moderate mixed traffic.
Object description: Multiple pedestrians.
Justification: The pedestrians maintain a consistent walking pace and direction without any sudden changes or interactions that suggest a near-miss or collision.

Event class: Collision
Scene context: Sunny day with dry road surface in a curbside area near an intersection, under daytime lighting with moderate mixed traffic.
Object description: A white sedan and a male pedestrian.
Justification: The pedestrian's path intersects with the vehicle's trajectory, and the vehicle does not deviate or stop, resulting in a collision with the pedestrian.

Event class: Collision
Scene context: Sunny day with dry road surface in a curbside area near an intersection, under daytime lighting with moderate mixed traffic.
Object description: A pedestrian and a white sedan
Justification: The pedestrian is in direct contact with the sedan, indicating a collision. The vehicle's position suggests it failed to stop in time, resulting in the impact.

Visual grounding: car id: {5}, person id: {3}

The Big Picture: A New Framework for Analyzing Road Safety

A New Framework for Proactive Safety, turning unstructured data such as video feeds into an active, automated safety monitoring system.

Unstructured Data
- Image Understanding
- Visual Question Answering
- Image Generation
- Audio Processing
- …

→ **Multimodal LLMs** →

Actionable Insights
- Detect early warning signs/near misses
- Reason why incidents happen
- Evaluate past video data automatically
- …

+ Crash reports

+ Road User feedback

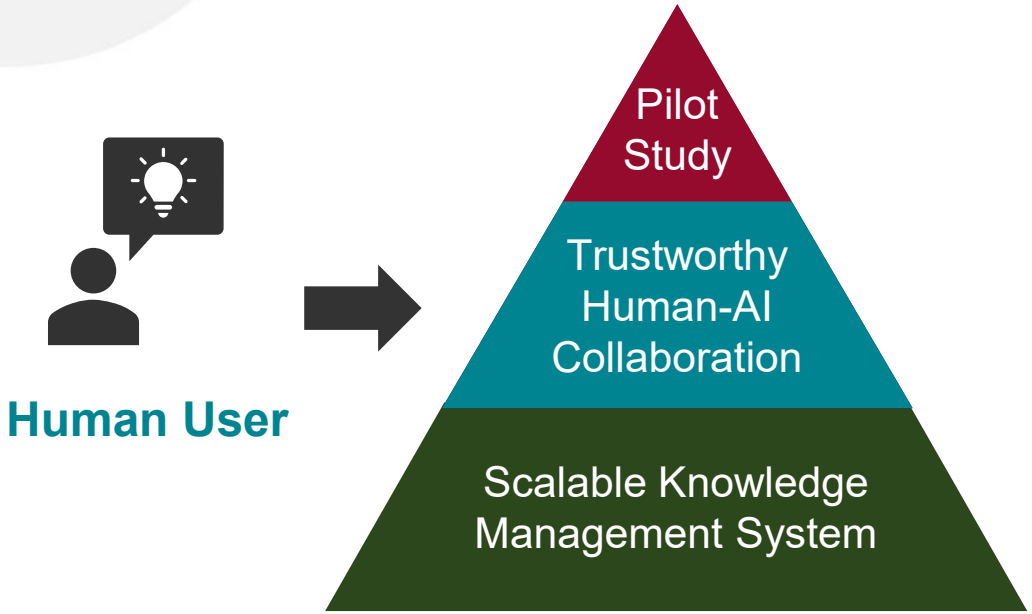
+ Work zone videos

+ Automated Vehicle trajectories

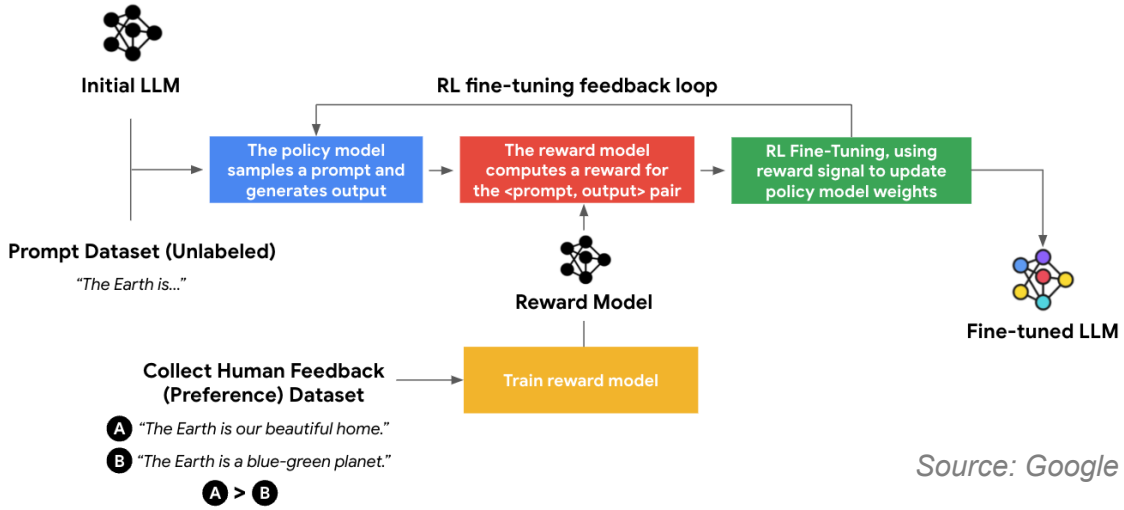
Benefits
- Enable automation- Perform continuous analysis once limited by human bandwidth - *Just like a tireless human!*
- Reduce burden of summarization
- Lower barrier between users and outcomes

Are We All Set?

- Even advanced AI systems can **misinterpret complex scenes** — especially when visibility is limited, image quality is poor, or the model "fills in" missing details that aren't truly.
- **Human expert knowledge** is crucial to enforce common sense and cause-effect relationships.
- **Solution:** Using **trustworthy Human-AI collaboration**, such as Human-Feedback Reinforcement Learning, can enhance current MLLM by refining their decision-making, reducing hallucinations, and aligning responses with real-world logic.



Human-Feedback Reinforcement Learning



Source: Google

20

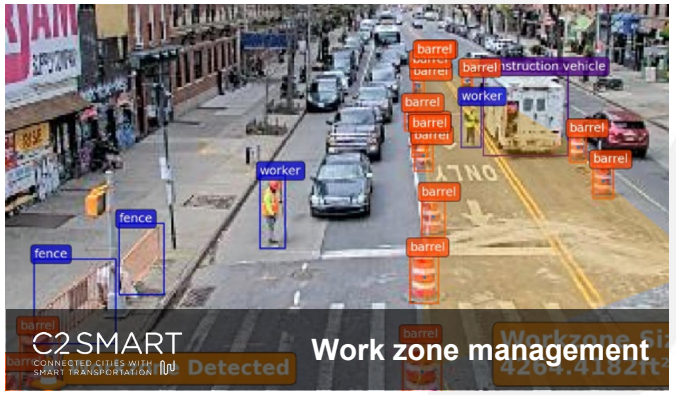
The Road Ahead

Advancing Vision Zero with MLLMs

- Analyze intersections to find out “why” accidents happen not just “that” they happen.
- Dynamic signage for work zones.
- Automatically detect near-misses and personal protective equipment (PPE) violations.
- Identify risky bottlenecks for large event foot traffic.
- Enable adaptive signal control using Gen AI + computer vision (e.g., recognizing pedestrians with mobility aids).

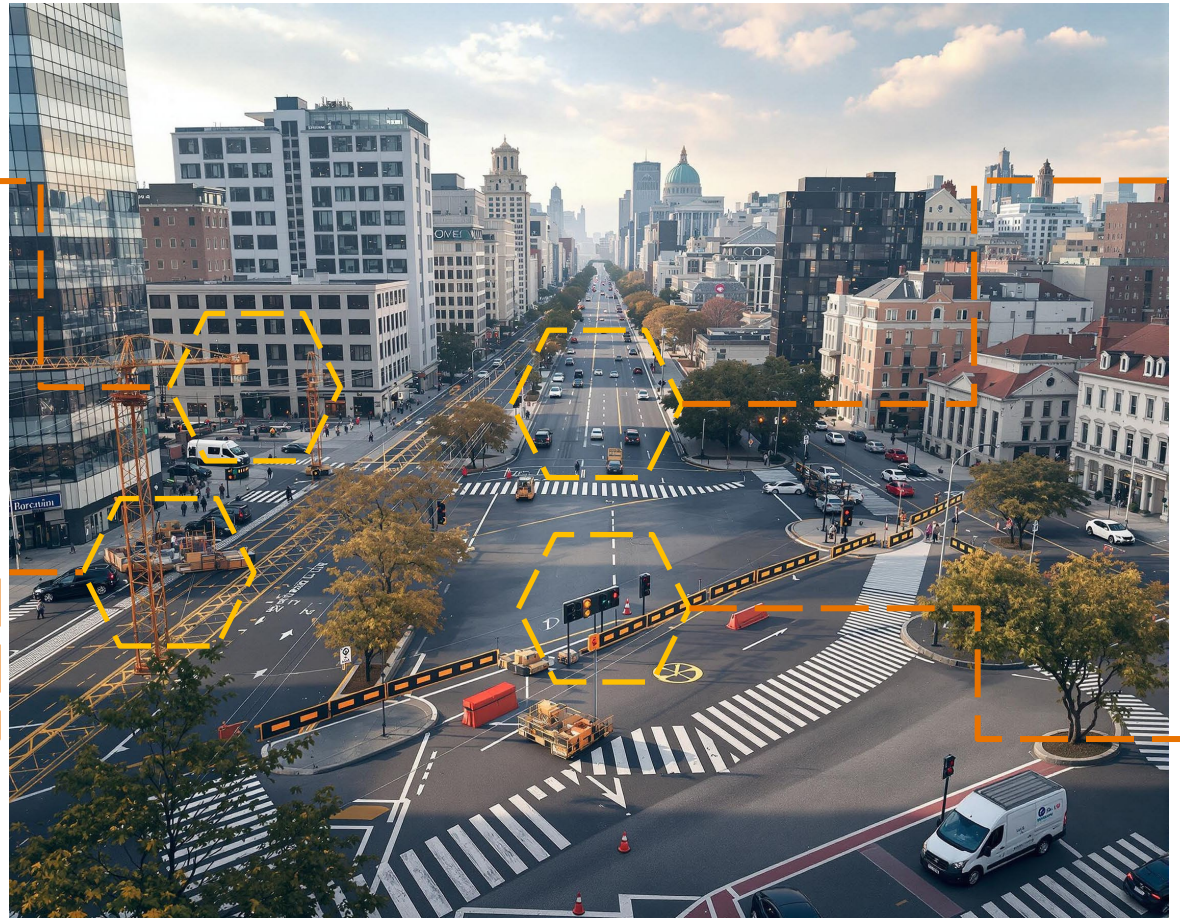
Develop evaluation criteria and testing protocols for the Gen AI tool that aims for cost-effective operations

Community engagement and agency-industry-academia collaboration



The Road Ahead

Generative AI: A Solution or A New Problem?



Reference:

- Zhang, R., Wang, B., Zhang, J., Bian, Z., Feng, C. and Ozbay, K., 2025. When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis. Accident Analysis & Prevention, 219, p.108077.

Explore More:

Scan the QR code to discover more about our research and ongoing projects.

TAHNK YOU

C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Contacts
c2smart.engineering.nyu.edu
c2smart@nyu.edu

Dr. **Kaan Ozbay**, kaan.ozbay@nyu.edu
C2SMART Center
New York University
Tandon School of Engineering
6 MetroTech Center, Brooklyn, NY 11201