Drive-VLM: Efficient Autonomous Driving Video Understanding via Hybrid-Granularity Alignment

IERG5050 Presentation

XIE Zifan

December 20 2025

MSc in AI, The Chinese University of Hong Kong

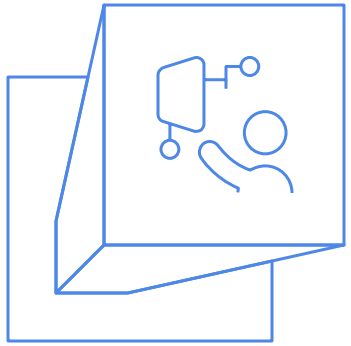
Table of content

Introduction: The Evolution of AD Perception

The Evolution of Autonomous Driving Perception

Trend

Autonomous driving is shifting from geometric perception (bounding boxes, lane lines) to high-level semantic understanding.

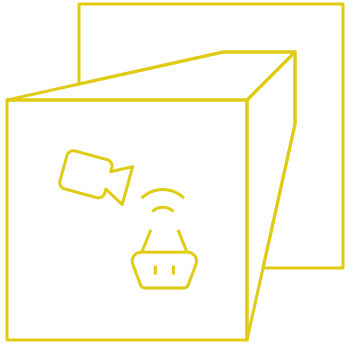


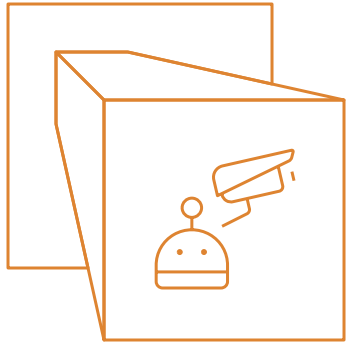
The Gap

Traditional stacks tell us "what" is there, but lack the reasoning to explain "why" or "how".

The Solution

Large Multimodal Models (LMMs) bridge this gap by enabling natural language interaction.





Goal

Create an agent that understands complex driving scenarios like a human.

Problem Statement: The "Granularity Mismatch"

Challenge: Applying general LMMs to driving creates a trade-off.

- **Verbose Outputs:** Standard models (e.g., Video-LLaVA) tend to generate lengthy descriptions, leading to hallucinations and slow inference.
- **Oversimplified Outputs:** Lacking critical safety details.

General VLM Output (Verbose / Hallucinatory)	Drive-VLM
"The video presents a cinematic night-time journey through a dimly lit urban landscape. From a first-person perspective, we observe the play of light and shadow as streetlamps reflect off the vehicle's hood. The scene opens with a moment of stillness at a red light, creating a calm atmosphere. As the signal turns green, the vehicle accelerates gently, passing a pedestrian who is crossing the street, adding a human element to the quiet city. The route is lined with glowing storefronts and parked cars, painting a vivid picture of typical late-night city life. The vehicle continues to navigate the rhythm of the traffic signals, eventually pausing again at another intersection, emphasizing the peaceful yet active nature of the urban environment."	Vehicle moves through a night urban area, stopping and starting at red traffic lights. Weather: clear, night. Ego-vehicle navigates intersections, occasionally slows; sparse traffic, street lights illuminate scene. No significant incidents or hazards.

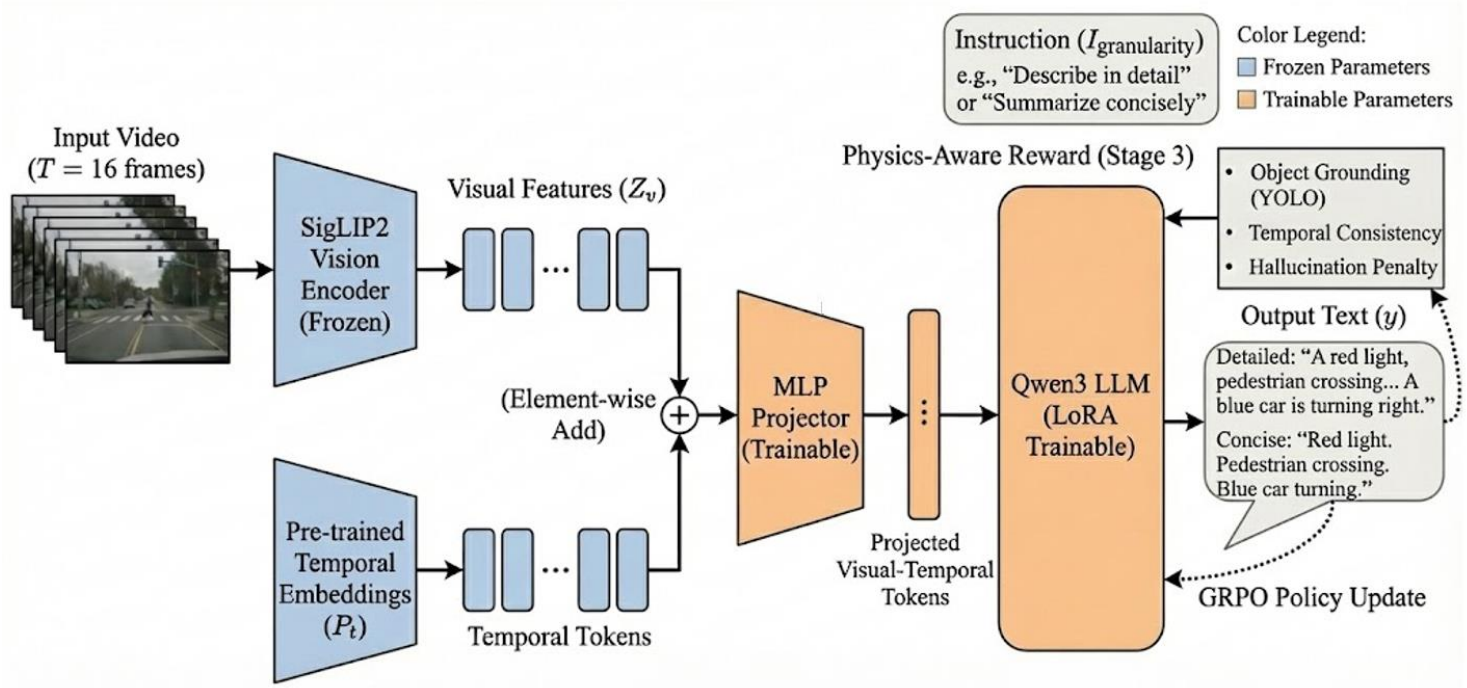
Two Needs in Driving:

- **"Microscope" Mode:** Exhaustive details for safety auditing.
- **"Copilot" Mode:** Concise, actionable summaries for real-time planning (e.g., "Red light. Stop.").

Constraint: Existing methods require massive datasets/compute, hindering academic research

Drive-VLM Framework Overview

Core Philosophy: An efficient framework built on SigLIP2 + Qwen3.



Three-Stage Pipeline:

1. **Stage 1:** Projector Pre-training (Feature Alignment).
2. **Stage 2:** Split-Entry Instruction Tuning (Granularity Control).
3. **Stage 3:** Physics-Aware Policy Optimization (GRPO) for hallucination suppression.

Methodology I: Hybrid-Granularity Data Pipeline

Innovation: Instead of using a single Ground Truth, we decouple data using a "Teacher LLM"

Comparison of Data Granularity Modes (Real Training Sample)

Mode	Example Content (Truncated)
Raw Caption	"The video provides a continuous observation from within a vehicle during a transition from twilight to a deep evening... surrounded by an urban landscape comprised of multi-lane roads... [Original GT]"
Detailed (Ours)	"The footage unfolds as the vehicle moves through an urban environment transitioning from the hush of twilight into a deeper evening... Static elements—guardrails, trees—line the route... The urban scene breathes with life... [Atmospheric Rewrite]"
Concise (Ours)	"Vehicle drives through moderate urban traffic at night, maintaining steady speed... Clear visibility in rainy conditions... Commercial area with lit signs and pedestrians... [Tactical Summary]"

Dual-Objective Dataset:

- **Detailed Mode ("Microscope"):** Narrative-driven, focuses on atmospheric elements and complex dynamics.
- **Concise Mode ("Copilot"):** Information distillation, retains only safety-critical states and ego-actions.

Methodology II: Stage 2 Split-Entry Instruction Tuning

Strategy: Treat one video as multiple training samples conditioned on different instructions ($I_{granularity}$).

video "0028cbbf-92f30408.mov"
q "Summarize the driving scene concisely."
▼ captions [] 5 items
0 "A gray sedan and black SUV shift lanes; a white van and "Hercules Cleaning Service" truck overtake from the left. Traffic flows steadily with minor speed changes and lane adjust ments. Weather remains clear, with consistent daylight and no significant disruptions."
1 "A gray sedan and black SUV shift lanes; a white van and "Hercules Cleaning Service" truck overtake from the left. Traffic flows steadily with minor speed and positioning change s. Clear weather and consistent highway conditions dominate. No major incidents or stops."
2 "The ego-vehicle moves steadily on a congested urban highway under clear weather. Traffic flows continuously with minor lane changes and speed adjustments; a black SUV and white van shift positions, while a "City Wide" truck and sedan periodically align ahead. No major incidents or weather changes occur."
3 "A gray sedan and white SUV move with traffic on a congested urban highway under clear skies. The camera vehicle maintains steady speed, observing slight lane changes and o vertaking maneuvers, including a black SUV and a white van. No major incidents or weather changes occur; traffic remains stable with consistent flow and visibility."
4 "The ego-vehicle moves steadily on a congested urban highway under clear weather. Traffic flows continuously with minor lane changes and speed adjustments; a black SUV slow s, a white van overtakes left, and a "City Wide" truck and sedan periodically align with the camera. No major incidents or weather changes occur."

Key Technique: Freezing Temporal Embeddings.

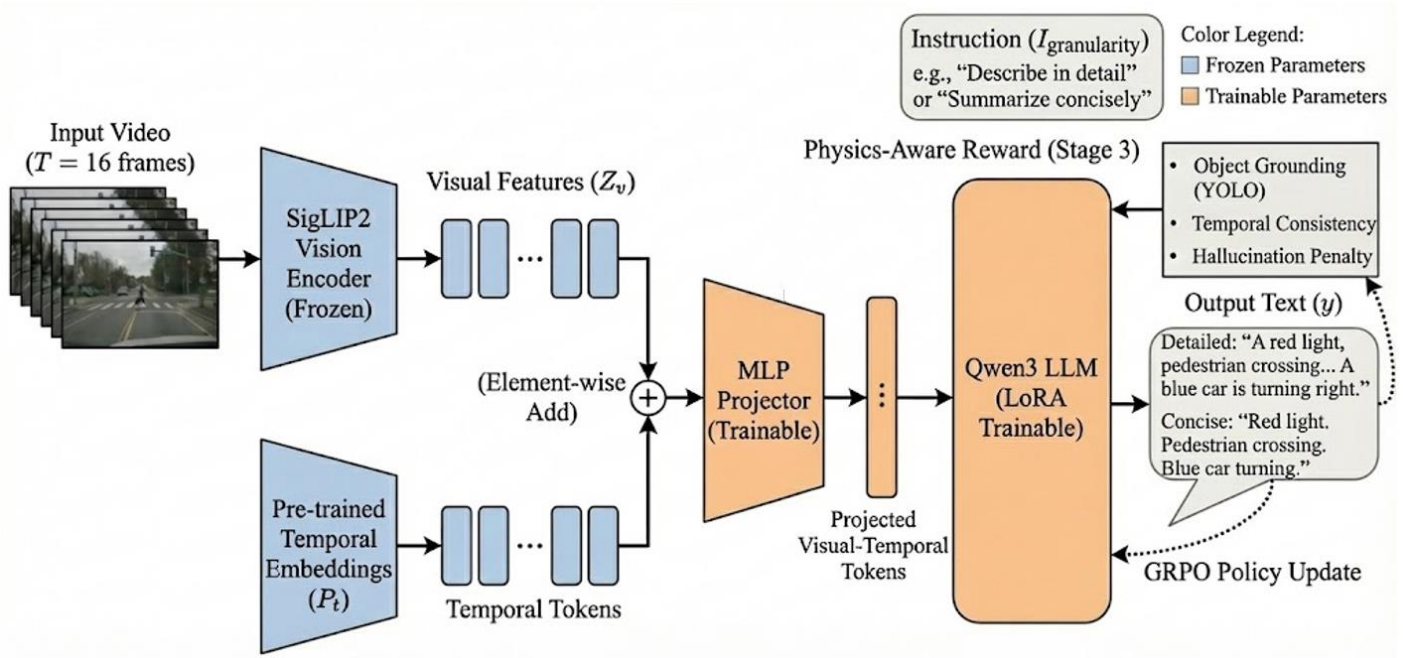
- **Reason:** Small dataset ($N = 446$).
- **Effect:** Prevents overfitting to specific video temporal patterns (memorizing "when" things happen vs. understanding "what" happens).
- **Focus:** Forces LoRA layers to learn generalized semantic reasoning.

Methodology III: From SFT to Policy Optimization

Limitation of Stage 2 (SFT): Supervised Fine-Tuning alone cannot eliminate hallucinations; models overfit to language priors.

The Solution: GRPO (Group Relative Policy Optimization).

Shift: Unlike standard RLHF (Human Preference), we use a **Physics-Aware Reward Function** to anchor output to verifiable visual data.



Methodology III: Physics-Aware Reward Function

Reward Formulation: $R_{total} = w_1 R_{obj} + w_2 R_{temp} + w_3 R_{detail} + w_4 R_{hall}$

```
DRIVING_OBJECTS = {
    'vehicles': ['car', 'truck', 'bus', 'suv', 'vehicle', 'motorcycle'],
    'traffic_control': ['traffic light', 'stop sign', 'traffic signal', 'signal'],
    'pedestrians': ['pedestrian', 'person', 'people'],
    'road_elements': ['road', 'street', 'intersection', 'crosswalk', 'lane'],
    'landmarks': ['building', 'storefront', 'sign', 'fence']
}

ACTION_VERBS = [
    'moving', 'stopped', 'turning', 'crossing', 'passing',
    'approaching', 'waiting', 'driving', 'proceeding'
]
```

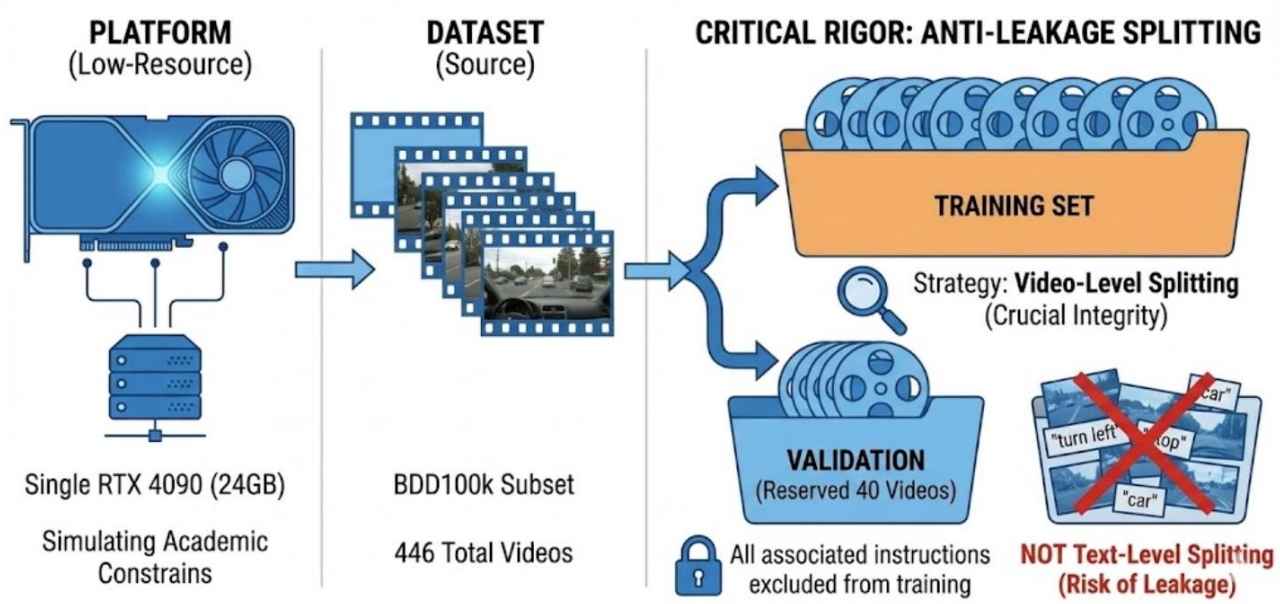
```
DESCRIPTIVE_ADJECTIVES = [
    'illuminated', 'lit', 'glowing', 'visible', 'marked',
    'nighttime', 'dark', 'bright', 'dimly', 'sparse'
]

TEMPORAL_KEYWORDS = [
    'initially', 'then', 'as', 'after', 'before', 'when',
    'during', 'subsequently', 'eventually', 'continues', 'throughout'
]
```

Key Components:

- **Object Grounding (R_{obj})**: F1-score against YOLOv8x detections. If the text says "car", YOLO must see a “car”.
- **Temporal Consistency (R_{temp}):** Checks for dynamic verbs and temporal connectives.
- **Detail Richness (R_{detail}):** ensures the model doesn't become too brief.
- **Hallucination Penalty (R_{hall}):** Strict penalty for inventing safety-critical objects (e.g., non-existent fire trucks).

Experimental Setup



Platform: Single NVIDIA **RTX 4090 (24GB)** – Simulating low-resource academic settings.

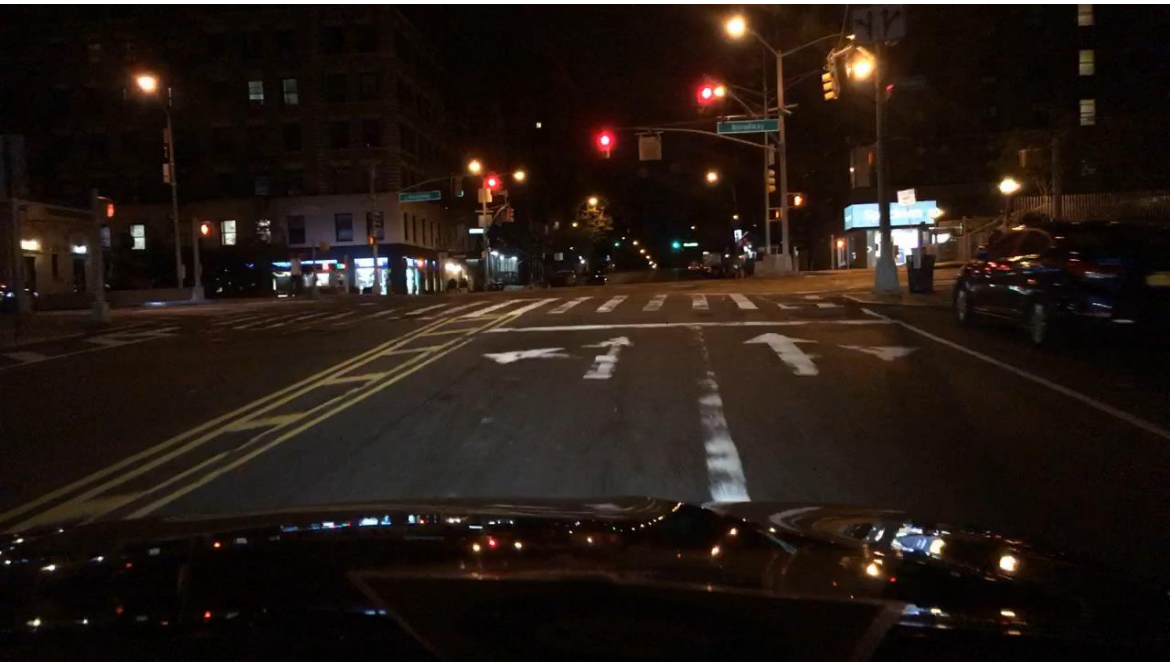
Dataset: BDD100k Subset (446 Videos).

Critical Rigor: Anti-Leakage Data Splitting.

- **Strategy:** Video-Level Splitting (not text-level).
- **Validation:** 40 videos reserved. All instructions associated with these videos are excluded from training to ensure true generalization.

Qualitative Analysis: "Microscope" Mode

Scenario: Nighttime urban intersection.



Instruction: “Describe in detail.”

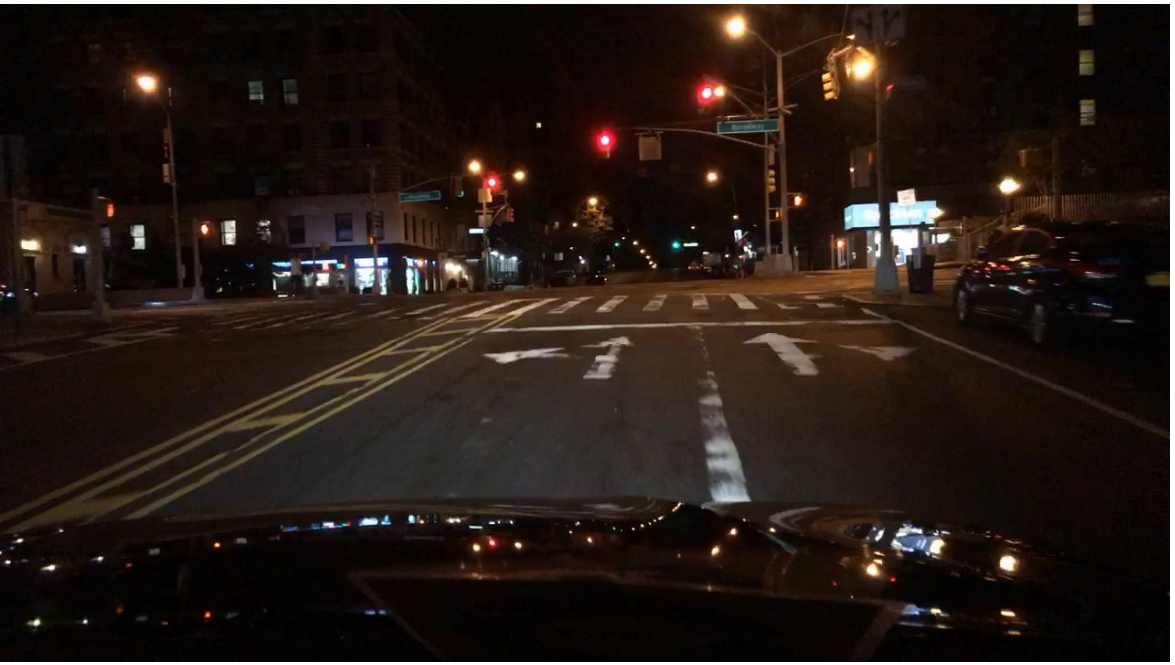
Comparison:

- **Stage 2 (SFT):** The video captures a nighttime journey through a city street. The traffic light changes from red to green, signaling the start of movement. The vehicle begins to move, passing by a bus stop and a pedestrian crossing marked by a yellow line. The scene includes a pedestrian visible on the sidewalk, adding a human element to the urban setting.”
- **Stage 3 (GRPO):** “Nighttimeurban street. The ego-vehicle is stationary at a red traffic light. A pedestrian crosses the street on the crosswalk. As the signal turns green, the vehicle accelerates forward, passing parked cars on the right. Streetlights illuminate the wet road surface.”

Key Takeaway: Hallucination Penalty forces factual precision.

Qualitative Analysis: "Copilot" Mode

Scenario: Nighttime urban intersection.



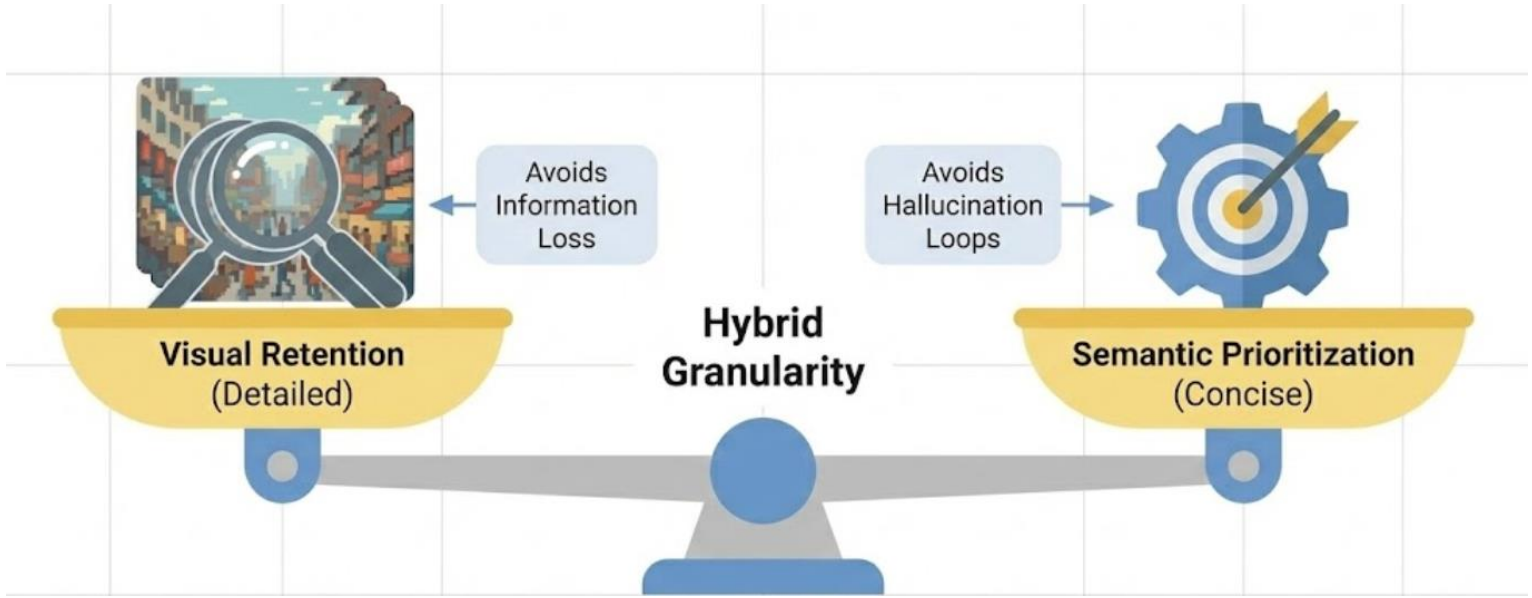
Instruction: “Summarize concisely.”

Comparison:

- **Stage 2 (SFT):** “The scene involves a vehicle approaching and then stopping at a red traffic light on a city street. The vehicle remains stationary, waiting for the light to turn green, with no significant movement or changes in the environment or other vehicles.” (38 words)
- **Stage 3 (GRPO):** “Stopped at red light. Pedestrian crossing. Light turns green. Resuming motion.” (12 words)

Key Takeaway: Ideal for low-latency decision making.

Discussion: The Necessity of Hybrid Granularity



Solving the Dilemma: Avoids "hallucination loops" (forcing long text from short data) and "information loss".

Dual-Objective Alignment:

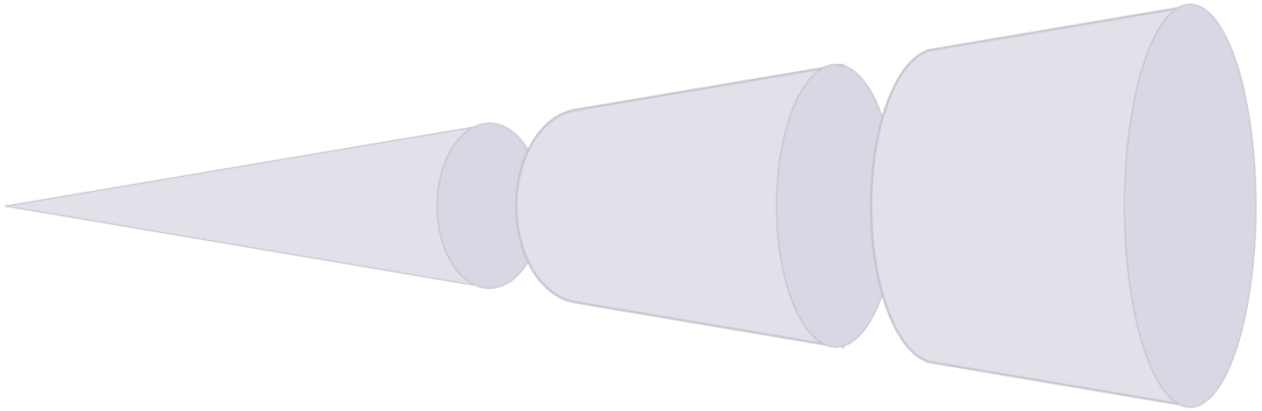
- **Visual Retention:** Detailed mode forces the projector to map all visual features.
- **Semantic Prioritization:** Concise mode teaches the LLM to identify high-value safety semantics.

Impact: A "best of both worlds" effect for embodied AI.

Conclusion & Future Work

Conclusion:

- Proposed **Drive-VLM**: A parameter-efficient framework (4B params).
- Introduced **Split-Entry Training** for granularity control.
- Validated **GRPO with YOLO** rewards to mitigate hallucinations.



Future Work:

- **End-to-End Control:** Connect to CARLA simulator.
- **Scaling Up:** Test on NuScenes dataset.
- **Multi-View Fusion:** Extend to BEV (Bird's Eye View) inputs.

Thank You