NEURAL INFORMATION
PROCESSING SYSTEMS

# *i*Finder: Structured Zero-Shot Vision-Based LLM Grounding for Dash-Cam Video Reasoning

Manyi Yao[†]

Bingbing Zhuang[‡]

Sparsh Garg[‡]

Amit Roy-Chowdhury[†]

Christian Shelton[†]

Manmohan Chandraker[‡,⋆]

Abhishek Aich[‡]

[‡]NEC Laboratories America [†]University of California, Riverside [⋆]University of California, San Diego

# Background

- Modern cars record vast amounts of driving video data.
- These videos can be leveraged to improve Advanced Driver Assistance Systems (ADAS) safety by **analyzing patterns**, **detecting anomalies**, and **identifying risk factors**.

# Key Challenges

However, meaningful analysis at scale remains challenging.

- **Data Overload**: Analyzing thousands of hours of driving video data is time-consuming and labor-intensive.
- **Limited Annotations**: Fine-grained driving event labels are scarce and expensive to collect.

# Why Existing VLMs Fall Short

Limitations of Vision-Language Models (VLM):

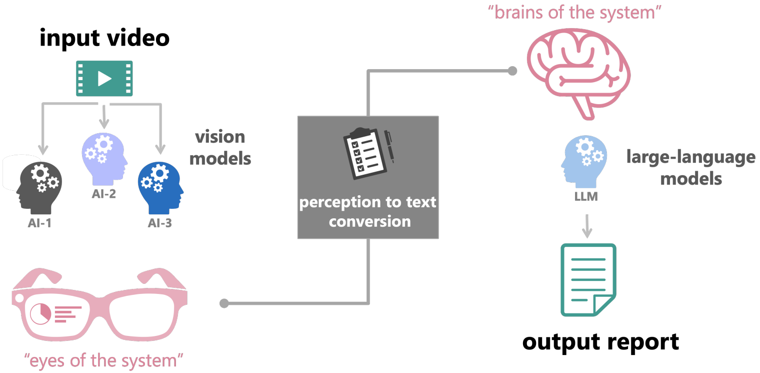
- **Critical Detail Gaps**: Struggles with driving-critical details.
- **Edge Case Failure**: Fails in corner cases due to limited training on long-tail events.
- **Unstructured Reasoning**: Lacks systematic analysis of spatial-temporal relationships, leading to vague or inconsistent conclusions.
- **Data Dependency**: Requires large-scale, high-quality training data.

# Desired Solution

Key Features:

- **Targeted Focus**: Prioritizes uncommon and significant elements.
- **Clear Descriptions**: Provides concise and accurate issue descriptions.
- **Insightful Analysis**: Explains the nature and impact of detected issues.
- **No Training Required**: Works out-of-the-box without model fine-tuning.

# Inspiration



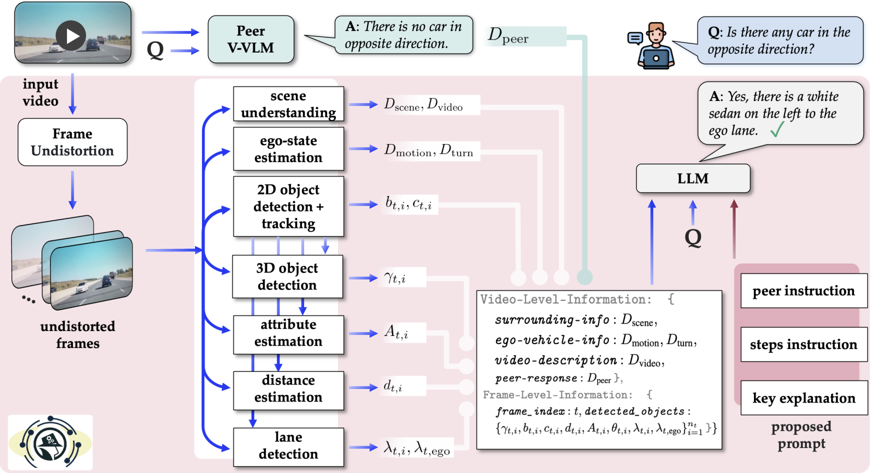
# Proposed Method – $\imath$Finder

A modular, training-free framework that integrates visual perception with language-based reasoning.

- **Specialized Vision Models**: Extract precise, structured visual attributes (e.g., distance estimation, lane occupancy.)
- **VLM**: Provides coarse semantic summaries of the visual scene.
- **LLM**: Conducts high-level reasoning over structured text inputs (e.g., event interpretation, accident causality.)

# *i*Finder Overview

NEURAL INFORMATION PROCESSING SYSTEMS



Q

Peer V-VLM

**A**: *There is no car in opposite direction.*

$D_{\text{peer}}$

**Q**: *Is there any car in the opposite direction?*

input video

Frame Undistortion

undistorted frames

scene understanding → $D_{\text{scene}}, D_{\text{video}}$

ego-state estimation → $D_{\text{motion}}, D_{\text{turn}}$

2D object detection + tracking → $b_{t,i}, c_{t,i}$

3D object detection → $\gamma_{t,i}$

attribute estimation → $A_{t,i}$

distance estimation → $d_{t,i}$

lane detection → $\lambda_{t,i}, \lambda_{t,\text{ego}}$

Video-Level-Information: {

*surrounding-info* : $D_{\text{scene}}$,

*ego-vehicle-info* : $D_{\text{motion}}, D_{\text{turn}}$,

*video-description* : $D_{\text{video}}$,

*peer-response* : $D_{\text{peer}}$ },

Frame-Level-Information: {

*frame_index* : $t$, *detected_objects* :

$\{\gamma_{t,i}, b_{t,i}, c_{t,i}, d_{t,i}, A_{t,i}, \theta_{t,i}, \lambda_{t,i}, \lambda_{t,\text{ego}}\}_{i=1}^{n_t}$ }}

LLM

Q

**A**: *Yes, there is a white sedan on the left to the ego lane.* ✓

peer instruction

steps instruction

key explanation

proposed prompt

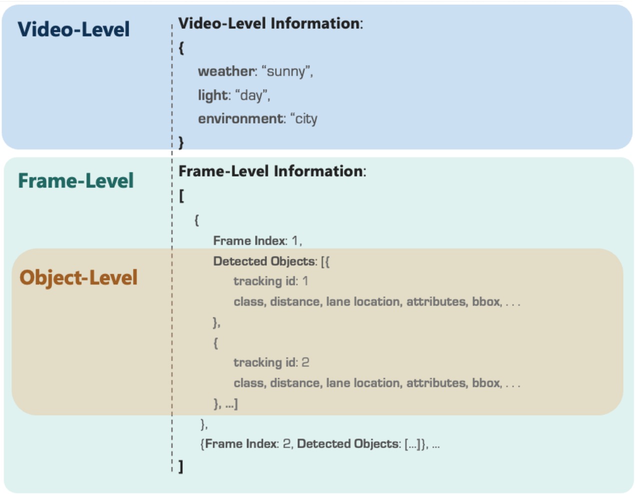
UC RIVERSIDE

NEC

NEC Laboratories America

# Perception-to-text Conversion

pre-defined data structure =



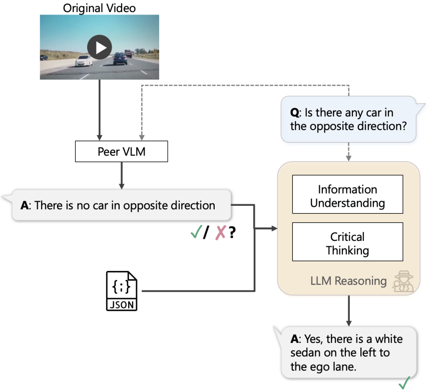
Video-Level

Frame-Level

Object-Level

```
Video-Level Information:
{
    weather: "sunny",
    light: "day",
    environment: "city
}
Frame-Level Information:
[
  {
      Frame Index: 1,
      Detected Objects: [{
          tracking id: 1
          class, distance, lane location, attributes, bbox, . . .
      },
      {
          tracking id: 2
          class, distance, lane location, attributes, bbox, . . .
      }, ...]
  },
  {Frame Index: 2, Detected Objects: [...]}, ...
]
```

# Peer VLM-assisted Understanding



Original Video
Peer VLM
Q: Is there any car in the opposite direction?
A: There is no car in opposite direction
✓/ ✗?
Information Understanding
Critical Thinking
LLM Reasoning
{;}
JSON
A: Yes, there is a white sedan on the left to the ego lane.
✓

- Intuition: Leverage strengths of video-VLMs while addressing their limitations.
- Steps:
  1. Use a peer video-VLM to generate an initial response based on the video and question.
  2. Combine this response with structured visual data.
  3. Feed the combined input into an LLM for final reasoning.

# Experiments: Tasks

- Multiple-choice VQA
- Open-ended VQA
- Accident Occurrence Prediction

# Experiments: Comparison Models

- **General Video-VLMs**: VideoLLaMA2[*], VideoChat2[†], and VideoLLaVA[‡]
- **Specialist AD Video-VLMs**: DriveMM[§] and WiseAD[¶]

[*] Zesen Cheng et al. "VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs". In: *arXiv preprint arXiv:2406.07476* (2024).

[†] KunChang Li et al. "Videochat: Chat-centric video understanding". In: *arXiv preprint arXiv:2305.06355* (2023).

[‡] Bin Lin et al. "Video-llava: Learning united visual representation by alignment before projection". In: *arXiv preprint arXiv:2311.10122* (2023).

[§] Zhijian Huang et al. "Drivemm: All-in-one large multimodal model for autonomous driving". In: *arXiv preprint arXiv:2412.07689* (2024).

[¶] Songyan Zhang et al. "WiseAD: Knowledge Augmented End-to-End Autonomous Driving with Vision-Language Model". In: *arXiv preprint arXiv:2412.09951* (2024).

# Multiple-choice VQA

- Datasets
    - **MM-AU** (Multi-Modal Accident Video Understanding)[‖]
        - 1953 videos focused on accident cause understanding.
    - **SUTD-TrafficQA** (Traffic Question Answering)[**]
        - 6075 QA pairs across 4111 traffic videos.
- Metric: Accuracy (%)

[‖] Jianwu Fang et al. "Abductive ego-view accident video understanding for safe driving perception". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[**] Li Xu, He Huang, and Jun Liu. "SUTD-TrafficQA: A Question Answering Benchmark and an Efficient Network for Video Reasoning Over Traffic Events". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

# Results on Multiple-choice VQA

$i$**Finder** surpasses both general-purpose and driving-specialist VLMs without fine-tuning, demonstrating that fine-grained details enhance factual reasoning.

| Method | MM-AU | SUTD |
|---|---|---|
| Generalist Models | | |
| VideoLLaMA2 | 50.95 | 47.51 |
| VideoLLaMA2 (w/ pt) | 52.89 | - |
| VideoChat2 | 49.56 | 42.17 |
| Video-LLaVA | 43.63 | 38.35 |
| AD-specialized Methods | | |
| DriveMM | 24.22 | 43.90 |
| $i$**Finder** (Ours) | **63.39** | **50.93** |

# Result breakdown on SUTD categories

*i***Finder** achieves the best accuracy in all 6 categories.

| Method | U | F | R | C | I | A |
|---|---|---|---|---|---|---|
| Generalist Models | | | | | | |
| VideoLLaMA2 | 49.2 | 39.0 | 48.5 | 53.5 | 35.8 | 45.2 |
| VideoChat2 | 42.5 | 38.1 | 43.8 | 49.2 | 30.4 | 42.8 |
| Video-LLaVA | 39.7 | 37.2 | 35.8 | 40.5 | 31.1 | 36.4 |
| AD-specialized Methods | | | | | | |
| DriveMM | 47.6 | 38.6 | 40.1 | 43.2 | 38.5 | 37.7 |
| *i***Finder** (Ours) | **52.2** | **43.5** | **50.2** | **56.8** | **39.2** | **49.6** |

Basic Understanding (U), Event Forecasting (F), Reverse Reasoning (R), Counterfactual Inference (C), Introspection (I), and Attribution (A).

# Open-ended VQA

- Dataset: **LingoQA**[††]
    - 500 QA pairs across 100 videos.
- Metrics:
    - **Lingo-J** (Lingo-Judge Score)
    - **BLEU** (Bilingual Evaluation Understudy)
    - **METEOR** (Metric for Evaluation of Translation with Explicit ORdering)
    - **CIDEr** (Consensus-based Image Description Evaluation)

[††] Ana-Maria Marcu et al. “Lingoqa: Video question answering for autonomous driving”. In: *European Conference on Computer Vision*, pp. 252–269.

# Results on Open-ended VQA

*i***Finder** outperforms others on the Lingo-Judge accuracy without fine-tuning.

| Method | Lingo-J | BLEU | METEOR | CIDEr |
|---|---|---|---|---|
| Generalist Models | | | | |
| VideoLLaMA2 | 36.00 | 4.15 | 33.45 | 26.28 |
| VideoChat2 | 41.20 | **6.58** | **36.81** | 40.98 |
| Video-LLaVA | 21.00 | 4.26 | 26.99 | 31.23 |
| AD-specialized Methods | | | | |
| WiseAD | 13.40 | 2.20 | - | 21.50 |
| *i***Finder** (Ours) | **44.20** | 6.07 | 35.80 | **42.01** |

# Accident Occurrence Prediction

- Dataset: Nexar[‡‡]
  - 100 videos randomly sampled from the training set as the evaluation set, maintaining a balanced distribution of 50 accident and 50 non-accident videos, consistent with the ratio in the full training set.
- Metrics: Accuracy, F1-Score, Precision, and Recall

[‡‡] Nexar Inc. *Nexar Dashcam Crash Prediction Challenge*. 2025.

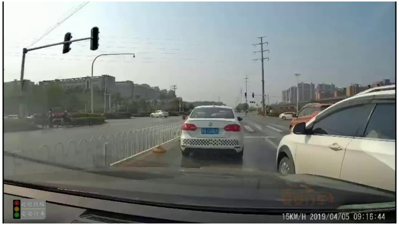
# Results on Accident Prediction

- VideoLLaMA2 and WiseAD exhibit a bias toward predicting accidents in all cases.
- We use VideoChat2 in peer-informed reasoning, further enhancing its performance.

| Method | Acc (%) | F1-Score | Precision | Recall |
|---|---|---|---|---|
| Generalist Models | | | | |
| VideoChat2 | 58.0 | 0.62 | 0.57 | 0.68 |
| VideoLLaMA2 | 50.0 | **0.67** | 0.50 | **1.00** |
| AD-specialized Methods | | | | |
| DriveMM | 49.0 | 0.32 | 0.48 | 0.24 |
| WiseAD | 50.0 | **0.67** | 0.50 | **1.00** |
| $i$**Finder** (Ours) | **62.0** | 0.59 | **0.64** | 0.54 |

# Qualitative Results

NEURAL INFORMATION PROCESSING SYSTEMS



**Q:** *What is the cause of the accident?*

**Ground Truth:** The car does not give way to normal driving vehicles when turning or changing lanes.

**VideoLLaMA2:** There is no accident in the video. However, there are several instances of cars stopping at red lights and waiting for them to turn green before continuing on their way.

**VideoLLaVA:** The cause of the accident is that the car in front of the white car suddenly stops, causing the white car to collide with the car in front of it.

**VideoChat2:** The cause of the accident is the white car's failure to stop at the red traffic light.

DriveMM

**DriveMM:** There is no accident in the video. The traffic is flowing smoothly with vehicles maintaining a safe distance from each other.

**iFinder:** A white car in the right lane attempted to cut into the ego lane, causing a near-collision, as seen in Object ID 13 moving from 9.78m (Frame 24) to 6.14m (Frame 40), with rot_y: -2.1, indicating a leftward turn.

UC RIVERSIDE

NEC

NEC Laboratories America

# Key Takeaway

**Integrating traditional vision modules provides more precise, task-relevant grounding, enabling LLMs to reason more accurately than standalone VLMs.**

# Conclusions

- Baseline models generate reasonable responses but exhibit limitations in **spatial reasoning, causal inference, and fine-grained scene understanding**.
- **Misinterpreting visual cues** can cause incorrect conclusions about hazards, traffic signals, or objects.
- By leveraging **structured scene representations**, $i$**Finder** mitigates these errors and produces more contextually accurate and reliable responses.

# Acknowledgments

This work was supported by NSF grant CNS-2312395 and NEC Laboratories America, Inc.

NEURAL INFORMATION
PROCESSING SYSTEMS

# Thank you!

Paper on arXiv

UC RIVERSIDE

NEC

NEC Laboratories **America**