From Segments to Scenes: Temporal Understanding for Agentic Autonomous Driving via Vision-Language Models

Kevin Cannons [*1] **Saeed Ranjbar Alvar** [*1] **Mohammad Asiful Hossain** [1] **Ahmad Rezaei** [1] **Mohsen Gholami** [1]
Alireza Heidarikhazaei [1] **Zhou Weimin** [2] **Yong Zhang** [†1] **Mohammad Akbari** [†1]

Code Dataset

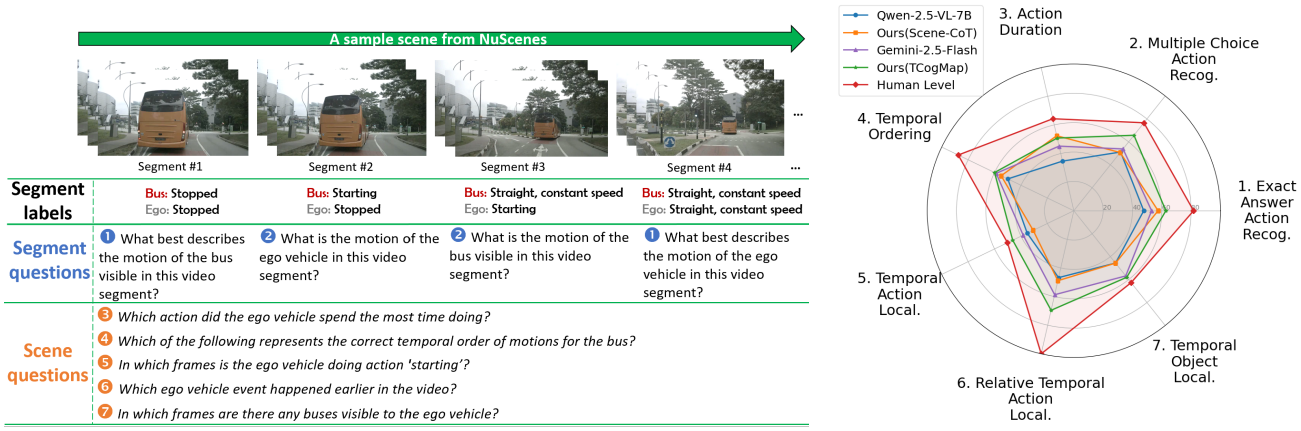

Figure 1. Overview of the proposed **TAD** benchmark. **Left**: An example of labels and question types. Each segment covers approximately 5 seconds of a video. Two question types pertain to actions within video segments, while the remaining five require information from the entire video. **Right:** Performance of open-source VLMs, the proposed training-free solutions (**Scene-CoT** and **TCogMap**), a closed-source VLM, and human performance across the 7 benchmark tasks in **TAD**.

Abstract

Vision-Language Models (VLMs) are increasingly deployed as the perception and reasoning backbone of autonomous agents acting in the wild, with autonomous driving (AD) being one of the most safety-critical instances. Reliable temporal understanding is essential for such agents to anticipate events, attribute causes, and act safely in dynamic environments, yet this remains a significant challenge even for state-of-the-art (SoTA) VLMs. Prior video benchmarks have emphasized other content (sports, cooking, etc.), yet no existing benchmark focuses *exclusively* on temporal understanding for *both short- and long-form* AD footage. To fill this gap, we present the Temporal Understanding in Autonomous Driving (TAD) benchmark, comprising nearly 6,000 question-answer (QA) pairs across 7 tasks, and evaluate 9 closed- and open-source generalist as well as AD-specialist models. Current SoTA models perform substantially below human accuracy on TAD. To improve the temporal reasoning of VLM-based driving agents, we propose two novel *training-free* solutions: **Scene-CoT**, which uses Chain-of-Thought (CoT) reasoning, and **TCogMap**, which incorporates an ego-centric temporal cognitive map produced by a trajectory-analysis module that operates as an agentic tool around the VLM. Integrated with existing VLMs, our methods improve average accuracy on TAD by up to 17.72% and by up to 10.35% on STSBench. By introducing TAD, benchmarking SoTA models, and proposing effective enhancements, this work aims to catalyze further progress on temporal under-

[*] Equal contribution. [†] Co-leads. [1]Huawei Technologies Canada Co., Ltd. [2]Huawei Cloud. Correspondence to: Kevin Cannons <kevin.cannons@huawei.com>, Saeed Ranjbar Alvar <saeed.ranjbar.alvar1@huawei.com>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

standing for agentic AD systems operating in the wild. The benchmark and evaluation code are available at Hugging Face and GitHub, respectively.

1. Introduction

Vision-Language Models (VLMs) are rapidly becoming the perception-and-reasoning engine of autonomous agents that must act in open-ended real-world environments. Among such agents, autonomous-driving (AD) vehicles are arguably the most safety-critical: they must continuously perceive, reason, and act in cluttered, dynamic scenes where errors can be catastrophic. Reliable *temporal* understanding of video is a prerequisite for any such agent — to anticipate events, attribute causes, plan over multiple seconds, and produce trustworthy explanations of its own behavior. Accordingly, evaluating VLMs on video understanding tasks has emerged as an important research direction (Li et al., 2024b; Fu et al., 2025; Li et al., 2024a; Liu et al., 2024; Li et al., 2025b; Cai et al., 2024). However, existing benchmarks have generally focused on specific types of video, such as sports (Caba Heilbron et al., 2015), cooking (Zhou et al., 2018a), and movies (Shu et al., 2025).

AD videos exhibit unique challenges compared to general video data: **1) Temporal Scale.** Large variation in the temporal scale of activities, requiring methods that localize events of varying durations within videos of arbitrary length. **2) Ego-Centric View.** The camera viewpoint makes understanding the ego-vehicle's actions challenging, since the ego is not observable in the frames. **3) Fine-grained Actions.** Vehicle actions can be subtle and may require additional cues for robust detection (e.g., a gradual lane change vs. driving straight). Currently, there is no *AD-exclusive benchmark* for assessing VLMs on these unique aspects of *temporal understanding* that includes *both short and long videos*.

In light of this gap, this paper proposes the **T**emporal Understanding in **A**utonomous **D**riving (TAD) benchmark, which evaluates a range of temporal-understanding tasks for VLM-based driving agents. TAD is built on the popular NuScenes dataset (Caesar et al., 2020) and consists of 5,861 QA pairs across 150 videos. Most tasks focus on **"scene-level"** (i.e., **video-level**) understanding, but two consider **"segment-level"** action recognition (short video segments $\leq$ 5 seconds). Inclusion of both scene and segment questions is a novel characteristic of TAD. Segment questions assess fine-grained vehicle actions (e.g., left-hand turns vs. changing lanes to the left); scene questions require understanding the temporal dynamics among atomic actions that together form a longer video. In addition, 4,481 segment-level vehicle action annotations were created. Figure 1 shows an example.

The paper presents an extensive evaluation on 9 unique VLMs across 30 system configurations, including open-source generalist, AD-specialist, and closed-source models. The results in Figure 1 show a significant gap between VLMs and humans — a gap that directly limits the reliability of any agentic driving system that uses such a VLM as a component. To address this gap, two novel solutions, **Scene-CoT** and **TCogMap**, are proposed. Scene-CoT leverages Chain-of-Thought (CoT) reasoning to answer AD questions by breaking down scene motions. TCogMap constructs an ego-centric temporal cognitive map from the ego vehicle's trajectory, used to enrich the VLM's input. Crucially, TCogMap operates as a lightweight *agentic* wrapper around the VLM: a deterministic trajectory-analysis tool that produces structured motion summaries that the VLM consumes alongside the raw frames.

The major contributions of this work are:

- Creation of TAD, the first QA benchmark *exclusively for AD* that evaluates *temporal understanding* for *both short- and long-form videos*, providing a testbed relevant to agentic driving systems operating in the wild.
- A new dataset of fine-grained, human-annotated action labels for ego and non-ego vehicles across 150 NuScenes videos.
- A comprehensive evaluation of SoTA open-source, closed-source and specialist VLMs on TAD.
- Introduction of Scene-CoT and TCogMap, training-free methods that improve average performance on TAD by up to 17.72% and on STSBench (Fruhwirth et al., 2026) by up to 10.35%.

2. Related Work

2.1. Temporal Understanding in General Videos

The temporal understanding of general-purpose videos has been explored through various tasks and benchmarks. Key tasks include *Temporal Localization* — specifying when an event occurs (Huang et al., 2024; Chen et al., 2024b; Wu et al., 2025b; Wang et al., 2024) — and *Dense Video Captioning*, which localizes and describes every event (Krishna et al., 2017; Zhou et al., 2018b; Deng et al., 2021; Wang et al., 2021; Yang et al., 2023; Jin et al., 2024; Kim et al., 2024; Zhou et al., 2024a). Some benchmarks, like MVBench (Li et al., 2024b) and Video-MME (Fu et al., 2025), assess a combination of spatial and temporal abilities. Others, such as TempCompass (Liu et al., 2024) and TemporalBench (Cai et al., 2024), focus on fine-grained temporal reasoning across daily life, sports, and documentaries. None

target the agentic driving setting, where temporal reasoning is a prerequisite for performing actions safely.

2.2. Benchmarks and Methods for VLM-Based AD

VLMs have emerged as a promising direction for enhancing AD systems, offering improved explainability and the ability to leverage multiple modalities (Jin et al., 2024; Marcu et al., 2024; Ma et al., 2024; Sima et al., 2024; Nie et al., 2024; Ding et al., 2024; Chen et al., 2024c; Pan et al., 2024; Shao et al., 2024; Tian et al., 2024; Wang et al., 2025a; Xu et al., 2025). Their applications range from conversational driving assistants (Ma et al., 2024) and VQA systems (Marcu et al., 2024; Sima et al., 2024) to integration within the driving pipeline itself (Xu et al., 2024; Pan et al., 2024; Shao et al., 2024; Chen et al., 2024c; Tian et al., 2024; Wang et al., 2025a; Xu et al., 2025; Wang et al., 2025b). In all such settings, the VLM acts as one component of a larger agentic system, and its temporal-reasoning failure modes propagate directly to that agent's behavior.

Several VLM-oriented benchmarks have been released in recent years to assess different AD-related tasks. LingoQA (Marcu et al., 2024), DriveLM (Sima et al., 2024), and VLADBench (Li et al., 2025a) offer broad QA datasets spanning perception, prediction, and planning. Other works focus on narrower aspects: BDD-X (Kim et al., 2018) provides ego-action descriptions and reasons; NuScenes-QA (Qian et al., 2024) and NuPrompt (Wu et al., 2025a) are perception-oriented; DRAMA (Malla et al., 2023) and Rank2Tell (Sachdeva et al., 2024) annotate critical scene objects with action suggestions; Ego3D-Bench (Gholami et al., 2026) considers 3D spatial understanding; SURDS (Guo et al., 2025a) targets fine-grained spatial understanding; CODA-LM (Chen et al., 2025) evaluates corner cases; and MetaVQA (Wang et al., 2025b) provides spatial and embodiment-related questions for closed-loop simulation.

The most closely-related works to TAD are ELM (Zhou et al., 2024b), STSBench (Fruhwirth et al., 2026), and STIBench (Li et al., 2025b). ELM is a training-oriented dataset; only two of its AD tasks require knowledge of previous frames, and its videos are 3.5 s long. STSBench contains only event/action recognition for short ($\leq$3 s) segments. STIBench uses 20 s videos for AD but focuses on general embodied scene understanding (e.g., camera-pose estimation), and most of its data is indoor.

In contrast, TAD is the first benchmark to exclusively assess temporal understanding in AD via both segment- and scene-level questions. Table 1 summarizes the most closely-related QA benchmarks; TAD is the second-largest video-only benchmark focused on temporal understanding. Note that ELM's #QAs is not directly comparable: it is training-oriented, mixes AD with other ego-centric imagery, and varies the input modality by question type.

Table 1. AD benchmarks for spatiotemporal understanding. "#QAs": number of questions. "I/V": image or video. "Seg./Scene Qs": segment- and scene-level questions.

Benchmark	#QAs	I/V	Seg.	Scene
DriveLM (Sima et al., 2024)	15k	I	✗	✗
DriveBench (Xie et al., 2025)	20k	I	✗	✗
SURDS (Guo et al., 2025a)	9.2k	I	✗	✗
Ego3D (Gholami et al., 2026)	8.6k	I	✗	✗
DriveLLM-o1 (Ishaq et al., 2025)	4.6k	I	✗	✗
ELM (Zhou et al., 2024b)	1525k	I & V	✓	✗
LingoQA (Marcu et al., 2024)	1k	V	✓	✗
VLADBench (Li et al., 2025a)	12k	V	✗	✓
STSBench (Fruhwirth et al., 2026)	971	V	✓	✗
STIBench (Li et al., 2025b)	2k	V	✗	✓
TAD (Ours)	5.8k	V	✓	✓

3. TAD Benchmark

3.1. Benchmark Annotations

The TAD benchmark is built on top of the NuScenes (Caesar et al., 2020) validation split, which contains 150 videos, each roughly 20 seconds in length, with accompanying sensor data and 3D bounding-box annotations. TAD focuses on the following fine-grained vehicle classes defined in NuScenes: car, bus, bicycle, construction vehicle, motorcycle, trailer, and truck. All other object annotations were ignored.

NuScenes does not provide action labels for scene objects, yet action understanding is a critical aspect of understanding the temporal dynamics of a scene, and thus, a critical aspect of the situational awareness needed by an AD agent. Hence, the first annotation task to create TAD was to label vehicle actions in NuScenes, which required a pre-processing step that divides each video into segments, as shown in Figure 2. For vehicle action understanding, employing a short segment-level approach is appropriate, since an action typically cannot be discerned from a single frame, and in a long video, multiple actions may occur. For TAD, five-second segments were selected, since most vehicle atomic actions are completed within five seconds. Thus, each NuScenes video was divided into ten uniformly-distributed, overlapping segments.

To streamline annotation and focus on the most relevant vehicles, each segment was filtered to include only vehicles within 50 m of the ego. This threshold was empirically set to retain the most relevant vehicles while optimizing human annotation effort. Vehicles with negligible trajectory displacement were automatically labeled as *stopped* and excluded from manual labeling.

Video segments with the corresponding bird's-eye view (BEV) visualization, including the trajectories of the ego and nearby vehicles, were provided to annotators. Annotators were asked to label a vehicle's action corresponding to

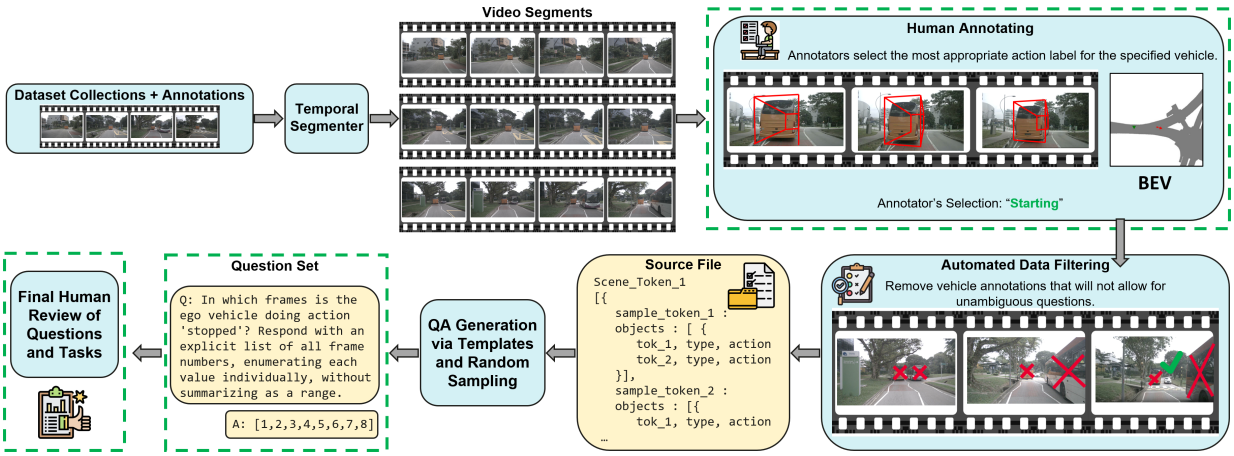

Figure 2. TAD annotation and question generation pipeline. Boxes with green dash: steps with human verification and quality assessment.

the dominant behavior during the video segment. Accordingly, only one action label was assigned to a vehicle per segment. The eight action categories used for TAD are: 1) traveling straight at constant speed, 2) stopping, 3) stopped, 4) starting, 5) turning left, 6) turning right, 7) changing lanes to the left, and 8) changing lanes to the right. In addition to the TAD benchmark, the vehicle-action annotations are released as a standalone resource that can be used to evaluate vehicle action recognition.

The final step of the annotation pipeline generates a JSON-formatted source file that stores the object and action labels per segment. These files constituted the raw materials used to construct the TAD QA set.

3.2. QA Creation

As shown in Figure 2, QA generation followed a template-driven approach. Seven QA templates were applied to the NuScenes videos. Logical checks were also included to verify that each question was unambiguous, e.g., ensuring that the objects referenced in each question are uniquely identifiable in the imagery. A final human-verification stage was included to further improve the quality of the QA pairs and task definitions.

After running the QA generation pipeline, 5,861 QAs were created for TAD, distributed across seven tasks. Several tasks — *Temporal Ordering*, *Temporal Action Localization*, *Relative Temporal Action Localization*, and *Temporal Object Localization* — are designed to challenge VLMs on **when** specific actions or events occur. *Temporal Ordering* and *Relative Temporal Action Localization* require the VLMs to understand the **temporal ordering** of events. The *Action Duration* task does not focus on when events happen, but **how long** they last. The final two tasks, *Exact Answer Action Recognition* and *Multiple Choice Action Recognition*, consider the VLM's ability to understand **fine-grained actions** in temporally cropped video segments. Sample questions are shown in Figure 1; detailed descriptions of each question type are provided in the supplementary materials.

3.3. TAD Benchmark Statistics

Question Distribution. The question distribution is shown in Table 2. In addition to the "Total" column, a breakdown is provided showing the number of ego vs. non-ego questions. All question types are well-represented, with a minimum of 80 per category. Two question categories, *Action Duration* and *Relative Temporal Action Localization*, only consider the ego vehicle. Conversely, *Temporal Object Localization* only includes questions about non-ego car objects, since this question considers when those objects are visible to the ego. The overall split between ego and non-ego vehicle questions is almost even (53.7% and 46.3%, respectively), so there is no significant benchmark bias favouring ego or non-ego objects.

Object Distribution. A breakdown of the vehicle types considered in TAD, as well as their counts, is shown in Table 3. The benchmark considers all vehicle categories relatively evenly, the lone exception being "Emergency Vehicle", which is rare in the original data itself. Note that some questions refer to "Non-Ego" vehicles, which corresponds to any vehicle other than the ego. Further stats, including instance count per action category, are in the supplementary material.

4. Proposed Methods

We propose two training-free, plug-in methods that wrap any VLM to improve its temporal understanding for AD agents: **Scene-CoT** and **TCogMap**. Both adopt the now-standard *tool-augmented* pattern of agentic systems, a structured preprocessing step that decomposes the video and produces text that the VLM consumes alongside the raw frames.

Table 2. Question counts per task in TAD. **AR**: Action Recognition. **AL**: Action Localization. **OL**: Object Localization.

Category	Ego	Non-Ego	Total
Exact Answer AR	1500	1104	2604
Multiple Choice AR	1030	756	1786
Action Duration	124	0	124
Temporal Ordering	62	18	80
Temporal AL	338	432	770
Relative Temporal AL	92	0	92
Temporal OL	0	405	405
Total Dataset	**3146**	**2715**	**5861**

Table 3. Statistics of vehicle types in the TAD questions.

Type	Count	Type	Count
Bicycle	153	Motorcycle	257
Bus	286	Non-Ego	432
Car	749	Trailer	108
Construction Vehicle	123	Truck	594
Emergency Vehicle	13	**Total**	**2715**

4.1. Scene-CoT: CoT for Temporal Understanding

The proposed Scene-CoT framework introduces a CoT-based (Wei et al., 2022) approach for structured scene reasoning in dynamic driving environments, consisting of three modules: 1) Video partitioning, 2) CoT reasoning, and 3) Large Language Model (LLM) QA. The system modules are described next, while a figure showing the Scene-CoT architecture is provided in the supplementary materials.

Video Partitioning. Scene-CoT first partitions the input video into smaller segments. Let $\mathcal{V}$ denote the original video and $S = \{s_1, s_2, \ldots, s_l\}$ be the set of l segments. The segment length and overlap are configurable to meet design and resource constraints. In the current implementation, segments are uniformly distributed across $\mathcal{V}$, each lasting 5 seconds with approximately 50% overlap between consecutive segments. This choice is suitable for driving scenarios, as a five-second window with overlap provides sufficient temporal context to capture common maneuvers (e.g., lane changes, turns) while maintaining responsiveness to rapidly changing scene dynamics. Exploration of more sophisticated segmentation strategies is left for future work.

To reduce pixel-wise redundancy from frame to frame and to limit the number of visual tokens input to the VLM, only a subset of frames is used to form the CoT descriptions. Specifically, four frames are uniformly sampled across each segment, i.e., $\{f_j^1, f_j^2, f_j^3, f_j^4\}$, where f denotes a frame and j corresponds to a segment number.

CoT Reasoning. For each segment, s_j, the VLM is prompted to generate a reasoning trace sufficient to answer the questions in TAD. To that end, a four-step CoT reasoning procedure is designed: **1) Scene description.** The VLM provides a high-level description of the scene. **2) Ego vehicle motion.** The VLM focuses on the ego car motion and estimates its action. **3) Motion of nearby vehicles.** The VLM isolates other vehicles close to the ego and describes their motion. **4) Summary formatting.** The VLM generates a final JSON-style summary containing fields for motions of all vehicles and nearby-vehicle identifiers. The final output for segment s_j, denoted c_j, is the concatenation of the scene description (Step 1) with the summary (Step 4).

LLM-Based QA. The LLM takes as input the benchmark question, q, together with the temporally ordered CoT reasoning outputs for segments $\mathcal{C} = (c_1, c_2, c_3, \ldots, c_l)$. Since the first step of Scene-CoT divides the video into short, temporally ordered segments, the resulting CoT outputs support questions that require temporal reasoning and ordering. Formally, the final answer is

$$a_{\text{Scene-CoT}} = \mathcal{L}(\mathcal{C}, q), \tag{1}$$

where $\mathcal{L}$ denotes the LLM. Additional details regarding Scene-CoT are provided in the supplementary materials.

4.2. TCogMap: Ego-Centric Temporal Cognitive Maps

The second proposed method, TCogMap, is a solution tailored to temporal understanding in the AD domain, as it leverages a vehicle trajectory-analysis module that operates as an agentic tool around the VLM. Figure 3 shows the three steps in TCogMap, described next.

Video Partitioning. TCogMap uses the same video partitioning module as Scene-CoT, resulting in a set of video segments, $S = \{s_1, s_2, s_3, \ldots, s_l\}$.

Ego-Vehicle Temporal Cognitive Map. Recent works have shown that generating a so-called "cognitive map" representation assists a VLM in question-answering (Gholami et al., 2026; Yang et al., 2025b). Building on this insight, the proposed approach extends cognitive maps to the video domain for improved temporal understanding in outdoor and dynamic AD scenes. A full extension that includes information about all objects is appealing, but may overload the VLM with extraneous information. TCogMap therefore considers only the most important object in a driving video: the ego vehicle.

TCogMap analyzes the sequence of ego-vehicle poses (translation, rotation, and timestamp, all provided with NuScenes) to classify the ego's motion per segment. The procedure, detailed in Algorithm 1, computes frame-wise global velocities and 2D speeds via finite differences (L4-7). A preliminary check classifies the segment as "Stopped" if instantaneous speeds fall below a stationary threshold, V_{stat} (L8-10). For

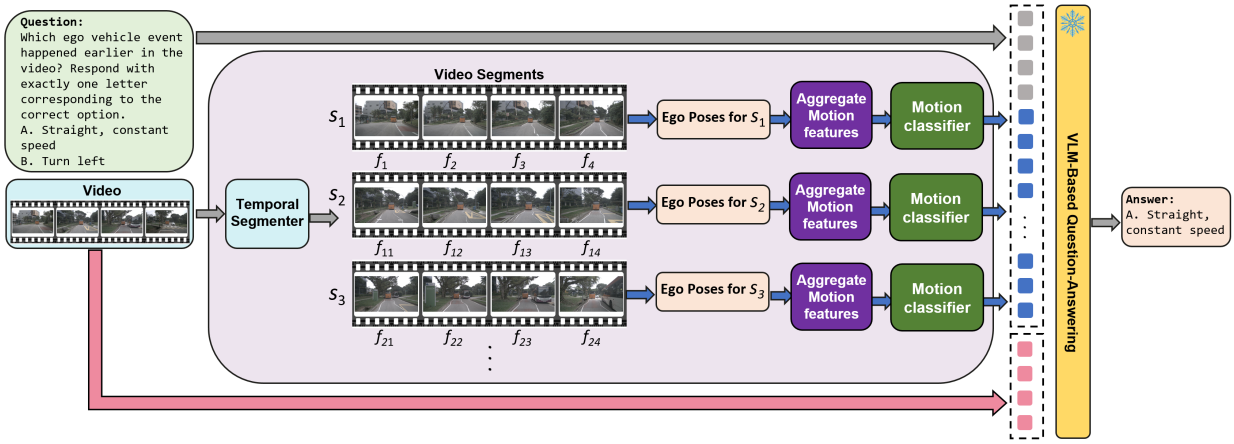

Figure 3. Overview of the proposed TCogMap method. A deterministic trajectory-analysis tool produces structured ego-motion summaries that augment the VLM's input — a lightweight agentic pattern around the VLM.

Algorithm 1 Ego-Vehicle Motion Classification

Input: Ego-poses $P = \{p_0, \ldots, p_{N-1}\}$, each p_i with translation $\mathbf{x}_i$, quaternion $\mathbf{q}_i$, timestamp t_i.
Input: Thresholds $V_{\text{stat}}, V_{\text{stopping}}, \Psi_{\text{turn}}, V_{y,\text{lc}}, V_{x,\text{lc}}$.
Output: Motion label (e.g., "Turn left").
for $i = 1$ **to** $N-1$ **do**
  $\mathbf{v}_i \leftarrow (\mathbf{x}_i - \mathbf{x}_{i-1})/(t_i - t_{i-1})$
  $\omega_i \leftarrow \|\mathbf{v}_i[0:2]\|$
end for
if $(\sum_{i=1}^{N-1}[\omega_i < V_{\text{stat}}])/(N-1) > 0.5$ **then**
  return "Stopped"
end if
$\Delta\psi \leftarrow \text{yaw}(\mathbf{q}_{N-1}) - \text{yaw}(\mathbf{q}_0)$
$\omega_{\text{start}} \leftarrow \omega_1,\ \omega_{\text{end}} \leftarrow \omega_{N-1}$
$\{\mathbf{v}_{\text{local},i}\} \leftarrow \text{ToLocalFrame}(\{\mathbf{v}_i\}, \{\mathbf{q}_i\})$
$\bar{v}_x, \bar{v}_y \leftarrow \text{mean}(\{\mathbf{v}_{\text{local},i}\})$
if $|\Delta\psi| > \Psi_{\text{turn}}$ **then**
  return "Turn left" if $\Delta\psi > 0$, else "Turn right"
else if $|\bar{v}_y| > V_{y,\text{lc}}$ **and** $|\bar{v}_x| > V_{x,\text{lc}}$ **then**
  return "Change lane left" if $\bar{v}_y > 0$, else "Change lane right"
else if $\omega_{\text{start}} < V_{\text{stopping}}$ **and** $\omega_{\text{end}} > 1.5V_{\text{stopping}}$ **then**
  return "Starting"
else if $\omega_{\text{start}} > 1.5V_{\text{stopping}}$ **and** $\omega_{\text{end}} < V_{\text{stopping}}$ **then**
  return "Stopping"
else
  return "Straight"
end if

segments with significant motion, dynamic features are extracted: the change in yaw $\Delta\psi$, initial and final speeds, and the average lateral/forward velocities $\bar{v}_y$, $\bar{v}_x$. Each global velocity is transformed into the ego's local frame using its rotation $\mathbf{R}_i$:

$$\mathbf{v}_{\text{local},i} = \mathbf{R}_i^\top \mathbf{v}_i, \tag{2}$$

and averages are computed over the segment. A hierarchical rule-based classifier (L14–24) then assigns one of eight motion labels, translating complex kinematics into a concise summary that the VLM can consume directly — a deterministic "tool" that produces structured context for the agent's reasoning step.

VLM-Based QA. The final step is to answer the question. The VLM, ϕ, is provided with all K frames of the input video, $\mathcal{F} = \{F_k\}_{k=1}^{K}$, the question q, and the structured ego-motion summary $\mathcal{M} = \{m_i\}_{i=1}^{l}$. To provide clear temporal context, each m_i is formatted as a string specifying the corresponding frames (e.g., "Motion summary for Frame1 to Frame7: The ego-vehicle is stopped"). The final answer is then

$$a_{\text{TCogMap}} = \phi(\mathcal{F}, \mathcal{M}, q). \tag{3}$$

5. TAD Benchmarking

5.1. Experimental Setup

For most TAD benchmarking results, four variants of each model are presented: baseline, baseline + textual ego pose, Scene-CoT (baseline + CoT outputs), and TCogMap (baseline + Temporal Cognitive Map). For the baseline, the VLM receives only the scene frames (approximately 40) and the question. The two segment tasks, *Exact Answer Action Recognition* and *Multiple Choice Action Recognition*, are exceptions and instead utilize approximately ten frames from the relevant segment as input to the VLMs. The baseline + textual ego pose variant is included for fair comparison with TCogMap.

Most questions in TAD are multiple choice and are evaluated using accuracy. Performance for questions that output a frame list is measured via temporal mean intersection over union (mIoU), a standard metric for temporal localization (Huang et al., 2024). For questions expecting output text, a binary exact-match metric is used.

For Scene-CoT, Qwen2.5-14B-Instruct-1M (Yang et al., 2025a) was used as the LLM, unless specified otherwise. For TCogMap, the thresholds were empirically set to $V_{\text{stat}} =$

Table 4. Main results for open- and closed-source generalist VLMs as well as AD-specific VLMs. Bold and underline indicate best and second-best results when comparing the four system variants for each open-source VLM. +EP: raw textual ego pose information is included in the VLM's input context. Gen : generalist; Drv : driving specialist. **EA**: exact answer, **MC**: multiple choice.

Method	Type	EA Act. Recog.	MC Act. Recog.	Action Dur.	Temp. Ord.	Temp. Act. Local.	Rel. Temp. Local.	Temp. Obj. Local.	Avg
Human level		81.68	76.67	64.29	87.50	50.21	100	62.69	74.72
Chance level		12.17	24.24	25.81	30.00	41.70	47.83	58.85	34.37
GPT-5-mini (OpenAI, 2025)	Gen	54.57	52.86	45.97	70.00	26.79	73.91	40.18	52.04
Gemini-2.5-Flash (Comanici et al., 2025)	Gen	53.07	53.86	45.16	58.75	38.35	58.70	56.81	52.10
Qwen2.5-VL-7B (Bai et al., 2025)	Gen	47.77	51.18	34.68	50.00	35.28	46.74	45.60	44.46
Qwen2.5-VL-7B+EP	Gen	46.54	49.50	35.48	48.75	35.12	57.61	37.27	44.44
+ Scene-CoT		57.49	50.73	52.42	55.00	30.77	48.91	45.65	48.71
+ TCogMap		**62.79**	**65.79**	**50.81**	**60.00**	**46.26**	**69.57**	**57.95**	**62.18**
Qwen2.5-VL-32B (Bai et al., 2025)	Gen	60.56	53.25	47.58	56.25	**44.96**	60.87	**62.91**	55.20
Qwen2.5-VL-32B+EP	Gen	64.13	59.24	53.23	60.00	44.45	52.17	60.10	56.19
+ Scene-CoT		60.45	54.09	46.77	55.00	31.56	44.57	48.99	48.78
+ TCogMap		**71.85**	**68.59**	**58.87**	**65.00**	41.42	**78.26**	**62.91**	**63.84**
InternVL3-8B (Zhu et al., 2025)	Gen	57.26	59.85	53.23	50.00	35.42	52.17	45.64	50.51
InternVL3-8B+EP	Gen	53.57	57.28	54.84	47.50	37.58	46.74	62.96	51.50
+ Scene-CoT		**62.52**	54.31	51.61	56.25	37.31	51.09	50.34	51.92
+ TCogMap		61.37	**66.69**	**57.26**	**66.25**	**47.08**	**67.39**	**64.10**	**61.45**
InternVL3-14B (Zhu et al., 2025)	Gen	56.87	53.70	52.42	56.25	43.53	59.78	43.90	52.35
InternVL3-14B+EP	Gen	60.41	59.80	52.42	55.00	46.78	52.17	**62.83**	55.63
+ Scene-CoT		62.06	53.42	51.61	56.25	36.61	51.09	51.64	51.81
+ TCogMap		**66.09**	**65.29**	**58.06**	**71.25**	**50.24**	**75.00**	59.00	**63.56**
InternVL3-38B (Zhu et al., 2025)	Gen	50.69	58.34	45.97	55.00	**45.41**	71.74	**68.27**	56.49
InternVL3-38B+EP	Gen	55.57	61.81	52.42	56.25	38.33	68.48	66.00	56.98
+ Scene-CoT		63.67	55.77	53.23	57.50	38.27	55.43	53.77	53.95
+ TCogMap		**67.01**	**68.25**	**60.48**	**72.50**	44.51	**80.43**	66.43	**65.66**
RoboTron (Huang et al., 2025)	Drv	43.63	58.17	29.84	41.25	31.00	65.22	28.92	42.58
RoboTron+EP	Drv	43.01	59.69	33.06	43.75	28.01	61.96	36.10	43.65
+ Scene-CoT		**58.68**	52.80	**52.42**	**57.50**	30.40	54.35	34.79	48.71
+ TCogMap		48.20	**62.77**	37.90	52.50	**40.15**	**79.35**	**40.58**	**51.64**
Cosmos-Reason (Azzolini et al., 2025)	Drv	50.46	49.38	39.52	52.50	16.39	64.13	16.06	41.21
Cosmos-Reason+EP	Drv	47.08	53.86	37.90	52.50	**37.42**	45.65	**46.70**	45.87
+ Scene-CoT		54.38	47.87	**55.65**	60.00	28.79	55.43	43.72	49.41
+ TCogMap		**58.26**	**64.61**	50.00	**67.50**	36.39	**67.39**	23.97	**52.59**

$0.2\,\text{m/s}$, $V_{\text{stopping}} = 1.0\,\text{m/s}$, $\Psi_{\text{turn}} = 10.0°$, $V_{y,\text{lc}} = 0.4\,\text{m/s}$, $V_{x,\text{lc}} = 1.0\,\text{m/s}$.

5.2. Chance and Human Performance

Human and chance performance were also computed on TAD. Chance performance was computed based on random selection for multiple-choice questions; a random subset of frames was used for temporal localization tasks. Human-level performance was computed using a randomly-sampled 10% subset of each question type. The first two rows in Table 4 show human-level and chance performance. Human performance is well above chance across the tasks. Performance on *Temporal Action Localization* is somewhat lower for humans, likely due to the similarity of some motion categories (e.g., stopped vs. stopping) and minor misalignments between the ground truth (in segments) and human responses (in frames).

5.3. Benchmarking Generalist VLMs

Experiments were performed using both closed- and open-source generalist VLMs not specifically trained for AD. For closed-source models, GPT-5-mini (OpenAI, 2025) and Gemini-2.5-Flash (Comanici et al., 2025) were tested. For open-source models, Qwen2.5-VL (Bai et al., 2025) and InternVL3 (Zhu et al., 2025) were selected, as these families have consistently provided top-tier performance and have the added benefit of released checkpoints across a range of sizes. Further parameter details are provided in the supplementary materials.

Open-Source Models. As shown in Table 4, for the baseline models, the average performance generally increases with the model size: Qwen2.5-VL improves by more than 10% from 7B to 32B. Scene-CoT, the proposed CoT-based method, provides modest gains over the baselines for smaller generalist models (e.g., Qwen2.5-VL-7B with Scene-CoT yields a 4.25% improvement). For larger models, Scene-CoT does not offer additional gains; this result is hypothesized to arise from larger VLMs' stronger inherent capabilities, which let them produce internal representations better suited for temporal QA than an explicit CoT-based description.

TCogMap consistently outperforms the *baseline + ego pose* variants in average performance by as much as 17.74% (Qwen2.5-VL-7B), and large models still show substantial gains (e.g., 8.68% with InternVL3-38B). Even though baseline + ego pose and TCogMap receive identical raw information, the VLM is unable to digest the raw pose data, illustrating the value of the structured temporal cognitive map.

Closed-Source Models. The average performance of GPT-5-mini and Gemini-2.5-Flash is comparable to the medium-sized InternVL3-14B. Interestingly, GPT-5-mini shows strong results on two of the "easier" tasks, *Video Temporal Ordering* and *Relative Temporal Localization*. Gemini-2.5-Flash performs worse on these two tasks, but achieves higher accuracies for *Temporal Action Localization* and *Temporal Object Localization*.

5.4. Benchmarking on AD-Specific Models

Two models trained on AD data, RoboTron (a.k.a. DriveMM) (Huang et al., 2025) (8B) and Cosmos-Reason (Azzolini et al., 2025) (7B), were selected as the specialist models, due to their recent SoTA results and demonstrated ability to perform a range of AD tasks. As shown in Table 4, the baselines for these models are on par with Qwen2.5-VL-7B. However, Cosmos-Reason is significantly outperformed by RoboTron on the *Temporal Action Localization* and *Temporal Object Localization* tasks, likely due to its input design: Cosmos-Reason receives entire videos rather than frame sequences, and thus lacks the explicit frame-index awareness required for these temporally-grounded questions.

Table 5. Average accuracies on ego vs. non-ego questions.

Method	Ego Qs	Non-Ego Qs
InternVL3-8B	55.15	43.24
+ Scene-CoT	54.63	48.62
+ TCogMap	68.17	47.81

These results indicate that in-domain training data is not a panacea for improving temporal understanding. Beyond the influence of training data, mechanisms that explicitly capture temporal structure and causal dependencies across frames are required to enhance understanding. Incorporating the proposed Scene-CoT and TCogMap as training-free extensions on top of the AD-specific models results in average accuracy gains of up to 11.38% on TAD, a directly actionable improvement for any agentic driving system that uses these VLMs as components.

5.5. Ablation Studies

Ablations use InternVL3-8B as the base model unless noted; further ablations (LLM choice, inference time) are in the supplementary materials.

Ego vs. Non-ego Performance Gains. As mentioned in Section 3, the questions related to the ego and non-ego vehicles are relatively balanced; however, it is crucial to investigate whether the performance varies between these two question categories. Table 5 shows the average accuracies for ego and non-ego questions. The baseline VLM achieves higher accuracy for ego-related questions. Ego-related cues are implicitly encoded in every frame through camera motion, whereas non-ego agents appear intermittently at lower resolution, giving the model a longer, more stable signal for ego questions, yielding better accuracies.

Furthermore, while Scene-CoT drops accuracy slightly for ego questions (0.52%) compared to the baseline (see Table 5), it improved performance on non-ego questions by more than 5%. Finally, results for TCogMap show boosts of 13.02% and 4.57% on ego and non-ego questions, respectively. These results are encouraging as they show that the additional context of the ego's motion offers the VLM with hints about other objects' actions and generally improves the model's temporal understanding of the scene. Further analysis with other models and detailed task accuracies for ego vs. non-ego are given in the supplementary materials.

Blind VLM Test. This ablation examines two questions about the TAD benchmark: 1) How much a model can infer from the question alone. 2) Whether temporal cognitive

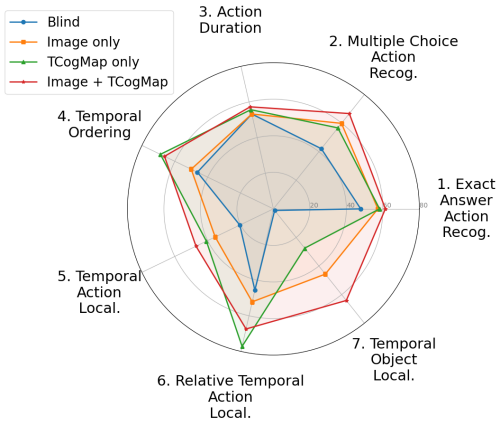

Figure 4. Blind test results for InternVL3-8B across the four configurations (Blind, Image Only, TCogMap Only, Image + TCogMap).

maps are sufficient for temporal reasoning in isolation.

As shown in Figure 4, four configurations were evaluated: **1) Blind**: Only the question is provided to the VLM. **2) Image Only:** Scene frames and question are given as input. **3) TCogMap Only:** Only the temporal cognitive map and question are provided, excluding frames. **4) Image + TCogMap:** All inputs (frames, TCogMap, and question) are provided together. Task accuracies in the **Blind** setting are the lowest among all configurations, often approaching chance-level. Comparing **Blind** with **Image Only** confirms that the benchmark questions cannot be answered without visual input. Interestingly, using the **TCogMap Only** setting yields higher accuracies in five tasks compared to **Image Only**. This result demonstrates that the proposed TCogMap effectively captures temporal dynamics even without direct visual input. Finally, combining both frames and TCogMap achieves best performance across most tasks, indicating that while TCogMap substantially enhances temporal reasoning, visual content remains essential for complete temporal understanding.

Scene-CoT and TCogMap were minimally adapted for STSBench to account for the shorter segment lengths, and the results are shown in Table 6. The result trends are similar to those seen in Table 4. In particular, Scene-CoT provides a modest average improvement over the base VLM, indicating the merits of extracting an intermediary CoT reasoning trace and allowing an LLM to perform final question-answering. TCogMap also yields an impressive boost of more than 10% over the base model. As expected, TCogMap's most significant gains are for ego vehicle questions (Ego and Ego-Agent tasks), but it still provides modest improvements for non-ego tasks. Similar to the TAD findings, this result reinforces that even if a question is not directly related to the ego vehicle, the context information provided by the ego's trajectory can still be instructive. Overall, the results for Scene-CoT and TCogMap show that they can generalize and provide accuracy increases on other datasets.

Table 6. Task and average accuracies on STSBench for the InternVL3-8B baseline and the proposed methods.

Method	Ego	Ego-Agent	Agent	Agent-Agent	Avg.
InternVL3-8B	34.36	51.41	46.69	39.52	43.00
+ Scene-CoT	40.05	50.83	37.51	**51.38**	44.94
+ TCogMap	**56.49**	**61.45**	**52.94**	42.53	**53.35**

Generalization to STSBench. Scene-CoT and TCogMap were also evaluated on STSBench (Fruhwirth et al., 2026), the most closely-related AD benchmark, which contains only short (3-second) event-recognition segments. With minor adaptations, Scene-CoT yields a modest average gain over the InternVL3-8B baseline, while TCogMap yields a 10.35% boost (43.00% → 53.35% average), with the largest gains on ego-centric tasks — evidence that the agentic-style augmentation transfers across temporal AD benchmarks. Per-task results are in the supplementary materials.

6. Conclusion and Future Work

This work introduced TAD, a benchmark for temporal understanding in autonomous driving, a domain that exemplifies the challenges of deploying agentic AI in the real world. Evaluation of generalist and AD-specific VLMs revealed a substantial gap between current SoTA models and human performance, directly limiting the reliability of any agent that uses such VLMs as components. To address this, two training-free solutions were proposed: **Scene-CoT**, a CoT-based reasoning approach, and **TCogMap**, which leverages a deterministic trajectory-analysis tool to feed the VLM an ego-centric temporal cognitive map. Both approaches follow a lightweight, tool-augmented agentic pattern that is increasingly central to making autonomous agents reliable in open-ended environments. Future work includes extending temporal cognitive maps to non-ego objects, integration into closed-loop planning agents, and creation of a fine-tuning dataset for VLM-based agentic AD systems.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically in the context of Vision-Language Models (VLMs) for autonomous driving (AD). By introducing the Temporal Understanding in Autonomous Driving (TAD) benchmark and proposing training-free methods to improve temporal reasoning, this work aims to contribute to the development of safer and more reliable autonomous driving systems. The societal implications of this work are predominantly positive: improved temporal understanding in VLM-based driving agents can directly contribute to safer autonomous vehicles, with downstream benefits for road safety and transportation efficiency. However, we acknowledge that advances in autonomous driving perception and reasoning carry broader responsibilities. Overestimating the capabilities of such systems based on benchmark performance—without accounting for real-world complexity, edge cases, and distribution shift—poses safety risks if results are prematurely translated into deployment. We therefore caution that TAD, while a rigorous testbed, should be regarded as one component of a broader validation framework rather than a definitive measure of deployment readiness. Regarding data, TAD is built upon publicly available datasets (e.g., NuScenes), and our human-annotated action labels were collected under standard annotation protocols. No personally identifiable information beyond what is already present in the source datasets is introduced. We encourage future users of TAD to adhere to the licensing terms of the underlying data sources. In summary, we believe the benefits of this work—providing a principled benchmark and effective enhancements for temporal reasoning in safety-critical agentic systems—outweigh its risks, provided that the limitations described herein and in the main paper are carefully observed by practitioners.

References

Alvar, S. R., Singh, G., Akbari, M., and Zhang, Y. DivPrune: Diversity-based visual token pruning for large multimodal models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9392–9401, 2025.

Azzolini, A., Bai, J., Brandon, H., Cao, J., Chattopadhyay, P., Chen, H., Chu, J., Cui, Y., Diamond, J., Ding, Y., et al. Cosmos-Reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Niebles, J. ActivityNet: A large-scale video benchmark for human activity understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 961–970, 2015.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 11621–11631, 2020.

Cai, M., Tan, R., Zhang, J., Zou, B., Zhang, K., Yao, F., Zhu, F., Gu, J., Zhong, Y., Shang, Y., et al. TemporalBench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Chen, K., Li, Y., Zhang, W., Liu, Y., Li, P., Gao, R., Hong, L., Tian, M., Zhao, X., Li, Z., et al. Automated evaluation of large vision-language models on self-driving corner cases. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7817–7826. IEEE, 2025.

Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., and Chang, B. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Eur. Conf. Comput. Vis.*, pp. 19–35. Springer, 2024a.

Chen, S., Lan, X., Yuan, Y., Jie, Z., and Ma, L. TimeMarker: A versatile video-LLM for long and short video understanding with superior temporal localization ability. In *arXiv preprint arXiv:2411.18211*, 2024b.

Chen, Y., Ding, Z.-h., Wang, Z., Wang, Y., Zhang, L., and Liu, S. Asynchronous large language model enhanced planner for autonomous driving. In *Eur. Conf. Comput. Vis.*, pp. 22–38. Springer, 2024c.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Contributors, L. LMDeploy: A toolkit for compressing, deploying, and serving LLM. https://github.com/InternLM/lmdeploy, 2023.

Deng, C., Chen, S., Chen, D., He, Y., and Wu, Q. Sketch, ground, and refine: Top-down dense video captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 234–243, 2021.

Ding, X., Han, J., Xu, H., Liang, X., Zhang, W., and Li, X. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13668–13677, 2024.

Egiazarian, V., Panferov, A., Kuznedelev, D., Frantar, E., Babenko, A., and Alistarh, D. Extreme compression of large language models via additive quantization. *arXiv preprint arXiv:2401.06118*, 2024.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *Int. Conf. Learn. Represent.*, 2022.

Fruhwirth, C., Malić, D., Lin, W., Schinagl, D., Schulter, S., and Possegger, H. STSBench: A spatio-temporal scenario benchmark for multi-modal large language models in autonomous driving. In *Adv. Neural Inform. Process. Syst.*, 2026.

Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 24108–24118, 2025.

Gholami, M., Akbari, M., Cannons, K., and Zhang, Y. CASP: Compression of large multimodal models based on attention sparsity. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9372–9381, 2025.

Gholami, M., Rezaei, A., Weimin, Z., Mao, S., Zhou, S., Zhang, Y., and Akbari, M. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *Int. Conf. Learn. Represent.*, 2026.

Gou, Y., Chen, K., Liu, Z., Hong, L., Jin, X., Li, Z., Kwok, J. T., and Zhang, Y. Perceptual decoupling for scalable multi-modal reasoning via reward-optimized captioning. *arXiv e-prints*, pp. arXiv–2506, 2025.

Guo, X., Zhang, R., Duan, Y., He, Y., Nie, D., Huang, W., Zhang, C., Liu, S., Zhao, H., and Chen, L. SURDS: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. 2025a.

Guo, Z., Liu, M., Ji, Z., Bai, J., Zhang, L., and Zuo, W. Decoupled visual interpretation and linguistic reasoning for math problem solving. *arXiv e-prints*, pp. arXiv–2505, 2025b.

Huang, D.-A., Liao, S., Radhakrishnan, S., Yin, H., Molchanov, P., Yu, Z., and Kautz, J. LITA: Language instructed temporal-localization assistant. In *Eur. Conf. Comput. Vis.*, pp. 202–218. Springer, 2024.

Huang, Z., Feng, C., Yan, F., Xiao, B., Jie, Z., Zhong, Y., Liang, X., and Ma, L. RoboTron-Drive: All-in-one large multimodal model for autonomous driving. In *Int. Conf. Comput. Vis.*, pp. 8011–8021, 2025.

Ishaq, A., Lahoud, J., More, K., Thawakar, O., Thawkar, R., Dissanayake, D., Ahsan, N., Li, Y., Khan, F. S., Cholakkal, H., et al. DriveLMM-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2025.

Jin, B., Zheng, Y., Li, P., Li, W., Zheng, Y., Hu, S., Liu, X., Zhu, J., Yan, Z., Sun, H., et al. TOD3Cap: Towards 3D dense captioning in outdoor scenes. In *Eur. Conf. Comput. Vis.*, pp. 367–384. Springer, 2024.

Jocher, G. and Qiu, J. Ultralytics yolo11, 2024. URL https://github.com/ultralytics/ultralytics.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Akata, Z. Textual explanations for self-driving vehicles. In *Eur. Conf. Comput. Vis.*, pp. 563–578, 2018.

Kim, M., Kim, H. B., Moon, J., Choi, J., and Kim, S. T. Do you remember? Dense video captioning with cross-modal memory retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 13894–13904, 2024.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Carlos Niebles, J. Dense-captioning events in videos. In *Int. Conf. Comput. Vis.*, pp. 706–715, 2017.

Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., and Shan, Y. SEED-Bench: Benchmarking multimodal large language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 13299–13308, 2024a.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 22195–22206, 2024b.

Li, Y., Tian, M., Lin, Z., Zhu, J., Zhu, D., Liu, H., Zhang, Y., Xiong, Z., and Zhao, X. Fine-grained evaluation of large vision-language models in autonomous driving. In *Int. Conf. Comput. Vis.*, pp. 9431–9442, 2025a.

Li, Y., Zhang, Y., Lin, T., Liu, X., Cai, W., Liu, Z., and Zhao, B. STI-Bench: Are MLLMs ready for precise spatial-temporal world understanding? In *Int. Conf. Comput. Vis.*, pp. 5622–5632, 2025b.

Lin, Z., Lin, M., Lin, L., and Ji, R. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *AAAI*, volume 39, pp. 5334–5342, 2025.

Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., and Hou, L. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL*, 2024.

Ma, Y., Cao, Y., Sun, J., Pavone, M., and Xiao, C. Dolphins: Multimodal language model for driving. In *Eur. Conf. Comput. Vis.*, pp. 403–420. Springer, 2024.

Malla, S., Choi, C., Dwivedi, I., Choi, J. H., and Li, J. DRAMA: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1043–1052, 2023.

Marcu, A.-M., Chen, L., Hünermann, J., Karnsund, A., Hanotte, B., Chidananda, P., Nair, S., Badrinarayanan, V., Kendall, A., Shotton, J., et al. LingoQA: Visual question answering for autonomous driving. In *Eur. Conf. Comput. Vis.*, pp. 252–269. Springer, 2024.

Nie, M., Peng, R., Wang, C., Cai, X., Han, J., Xu, H., and Zhang, L. Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving. In *Eur. Conf. Comput. Vis.*, pp. 292–308. Springer, 2024.

OpenAI. Gpt-5. https://openai.com/gpt-5/, 2025.

Pan, C., Yaman, B., Nesti, T., Mallik, A., Allievi, A. G., Velipasalar, S., and Ren, L. VLP: Vision language planning for autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 14760–14769, 2024.

Qian, T., Chen, J., Zhuo, L., Jiao, Y., and Jiang, Y.-G. NuScenes-QA: A multi-modal visual question answering benchmark for autonomous driving scenario. In *AAAI*, volume 38, pp. 4542–4550, 2024.

Rezaei, A., Gholami, M., Alvar, S. R., Cannons, K., Hossain, M. A., Weimin, Z., Zhang, Y., and Akbari, M. CPPO: Contrastive perception policy optimization for VLM agents. In *Int. Conf. Mach. Learn.*, 2026.

Sachdeva, E., Agarwal, N., Chundi, S., Roelofs, S., Li, J., Kochenderfer, M., Choi, C., and Dariush, B. Rank2Tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7513–7522, 2024.

Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S. L., Liu, Y., and Li, H. LMDrive: Closed-loop end-to-end driving with large language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 15120–15130, 2024.

Shu, Y., Liu, Z., Zhang, P., Qin, M., Zhou, J., Liang, Z., Huang, T., and Zhao, B. Video-XL: Extra-long vision language model for hour-scale video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 26160–26169, 2025.

Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., and Li, H. DriveLM: Driving with graph visual question answering. In *Eur. Conf. Comput. Vis.*, pp. 256–274. Springer, 2024.

Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., and Zhao, H. DriveVLM: The convergence of autonomous driving and large vision-language models. 2024.

Tseng, A., Chee, J., Sun, Q., Kuleshov, V., and De Sa, C. QuIP#: Even better LLM quantization with Hadamard incoherence and lattice codebooks. *Int. Conf. Mach. Learn.*, 2024.

Wang, H., Xu, Z., Cheng, Y., Diao, S., Zhou, Y., Cao, Y., Wang, Q., Ge, W., and Huang, L. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. 2024.

Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., and Alvarez, J. M. OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 22442–22452, 2025a.

Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., and Luo, P. End-to-end dense video captioning with parallel decoding. In *Int. Conf. Comput. Vis.*, pp. 6847–6857, 2021.

Wang, W., Duan, C., Peng, Z., Liu, Y., and Zhou, B. Embodied scene understanding for vision language models via MetaVQA. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 22453–22464, 2025b.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inform. Process. Syst.*, 35:24824–24837, 2022.

Wu, D., Han, W., Liu, Y., Wang, T., Xu, C.-z., Zhang, X., and Shen, J. Language prompt for autonomous driving. In *AAAI*, volume 39, pp. 8359–8367, 2025a.

Wu, Y., Hu, X., Sun, Y., Zhou, Y., Zhu, W., Rao, F., Schiele, B., and Yang, X. Number it: Temporal grounding videos like flipping manga. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 13754–13765, 2025b.

Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q. A., Liu, Z., and Pan, L. Are VLMs ready for autonomous driving? An empirical study from the reliability, data and metric perspectives. In *Int. Conf. Comput. Vis.*, pp. 6585–6597, October 2025.

Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.-Y. K., Li, Z., and Zhao, H. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024.

Xu, Z., Bai, Y., Zhang, Y., Li, Z., Xia, F., Wong, K.-Y. K., Wang, J., and Zhao, H. DriveGPT4-V2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 17261–17270, 2025.

Yang, A., Nagrani, A., Seo, P. H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., and Schmid, C. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10714–10726, 2023.

Yang, A., Yu, B., Li, C., Liu, D., Huang, F., Huang, H., Jiang, J., Tu, J., Zhang, J., Zhou, J., et al. Qwen2.5-1M technical report. *arXiv preprint arXiv:2501.15383*, 2025a.

Yang, J., Yang, S., Gupta, A. W., Han, R., Fei-Fei, L., and Xie, S. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10632–10643, 2025b.

Zhou, L., Louis, N., and Corso, J. J. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *Brit. Mach. Vis. Conf.*, 2018a.

Zhou, L., Xu, C., and Corso, J. Towards automatic learning of procedures from web instructional videos. In *AAAI*, volume 32, 2018b.

Zhou, X., Arnab, A., Buch, S., Yan, S., Myers, A., Xiong, X., Nagrani, A., and Schmid, C. Streaming dense video captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18243–18252, 2024a.

Zhou, Y., Huang, L., Bu, Q., Zeng, J., Li, T., Qiu, H., Zhu, H., Guo, M., Qiao, Y., and Li, H. Embodied understanding of driving scenarios. *Eur. Conf. Comput. Vis.*, 2024b.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A. TAD Benchmark: Details and Additional Statistics

In the main body of the paper, fundamental statistics regarding the benchmark, such as the number of questions per task, are provided. Further analyses on the statistics of the TAD benchmark are provided in this section.

A.1. Vehicle Action Statistics in TAD

To determine the distribution of vehicle actions in the benchmark, an analysis was performed on the *Exact Answer Action Recognition* question type, the results of which are shown in Figure 5. The data is presented in terms of ego, non-ego, as well as the combined total. The two most dominant action categories are "stopped" and "straight, constant speed". Notably, the ego vehicle is most commonly found to be traveling "straight, constant speed"; whereas, non-ego vehicles are most often stopped. This result is quite intuitive, as the NuScenes data generally shows the ego vehicle driving throughout cities with a significant number of stationary vehicles parked in lots or along streets. Thus, these statistics reflect the underlying NuScenes data distribution. Note that in the current work, action labels were also assigned to the video segments for the first time. A total of 4,481 vehicle actions were labeled during the human annotation procedure.

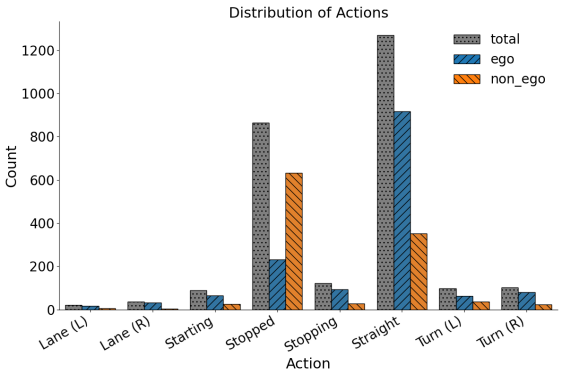

Figure 5. Distribution of the actions in the *Exact Answer Action Recognition* question type. Shorthand notations: Lane (L): change lane to left, Lane (R): change lane to right, Turn (L): Turn left, Turn (R): Turn right, and Straight: Straight, constant speed

A.2. Vehicle Action Definition Clarification

To avoid confusion, an additional note should be made regarding the definition of the actions used in TAD. In particular, the actions are defined **with respect to the vehicle that is performing the action.** As an example, imagine a scenario where a truck is traveling in the opposite direction to the ego vehicle and subsequently, the truck turns right onto a side street. In TAD, the truck's action would be labeled as **Turn Right**, since the driver of the truck has literally turned the wheel of his/her vehicle to make a right hand turn. The labels in TAD are **not** based on how the ego vehicle observes vehicles moving from its own perspective (i.e., in this example, within the ego vehicle's front camera field of view, the truck would be seen as traveling further to the left during its turn). Defining vehicle actions in this manner presents an additional challenge for VLMs, but models that can differentiate between such actions exhibit a greater understanding of the environment, road layout, and 3D scene.

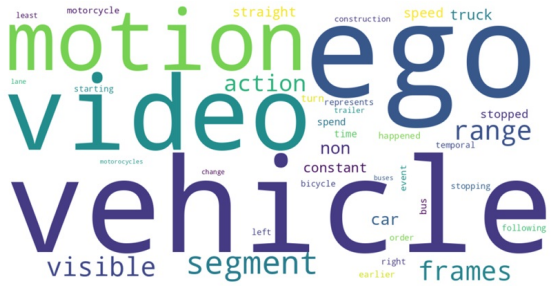

Figure 6. Word cloud for the questions in the TAD benchmark (excluding the instructions)

A.3. Word Cloud of Questions in the TAD Benchmark

The word cloud in Figure 6 summarizes the distribution of question terms in the benchmark after removing instructional tokens related to the answer format. The most prominent words are vehicle, motion, ego, video, segment, and action, which highlight the benchmark's emphasis on reasoning about temporal dynamics. The dominance of these words reflects the centrality of motion cues, ego-vehicle context, and temporal segmentation in the tasks, underscoring the benchmark's focus on temporal understanding in autonomous driving.

B. Parameters for the Baselines

In this section, the specific parameters and settings related to the baselines in the main body are introduced.

- **GPT-5-mini:** The model is called using its default API configuration, with reasoning enabled by default and set to "medium". `max_new_token` is set to 2048.
- **Gemini-2.5-Flash:** The model is called using its default API configuration, with thinking enabled by default and the thinking budget set to 1024 tokens. Gemini-2.5-Flash supports video inputs. Hence, the sampled frames in the NuScenes dataset are stored as videos with FPS=2. Then, the stored videos are used for inference. `max_new_token` is set to 2048.
- **InternVL3 family:** The LMDeploy inference frame-

work (Contributors, 2023) is used to get the inference results for the InternVL3 family of models. To keep the number of visual tokens produced by long sequences of frames within the model's maximum context length, `max_dynamic_patch=1` is used. `max_new_token` = 1024 is used for this model.

- **Qwen-2.5-VL family:** The LMDeploy inference framework (Contributors, 2023) is used to get the inference results for Qwen-2.5-VL family of models. To keep the number of visual tokens obtained from long sequences of frames within the model's maximum context length, `min_pixels`=256 $\times$ 28 $\times$ 28 and `max_pixels`=512 $\times$ 28 $\times$ 28 are used. `max_new_token` = 1024 is used for this model.
- **Robotron:** The parameters and prompts provided in their official examples are used [1].
- **Cosmo-reason:** The parameters provided in their official code [2] are used. Since the model receives video as input, the sampled frames in the NuScenes dataset are converted to videos with FPS=2. The stored videos are then used for inference.

B.1. Task Details and Examples

In the main body of the paper, Fig. 1 provides an overview of the TAD benchmark, including a listing of the seven different question types, as well as the frames and labels for several video segments. In this section, additional details and examples from TAD are given.

Table 7 provides a more detailed description for each of the seven question types, along with an explicit example (including both the question and answer) from the benchmark. In the description, some of the motivations that inspired the inclusion of question types are also listed.

An additional pictorial example showing the format of questions in TAD is provided in Figure 7. This figure is meant for visualization purposes only to understand the format of the benchmark questions. The figure highlights the different question types and the breakdown of the video into segments. The seven question types are shown in grey bubbles. Dark blue text corresponds to the questions, while red text shows the available answer options (note, not all of the options are **correct**) and the correct answers are **bolded**. Three video segments are displayed (Segment 1, Segment 2, and Segment 10), represented by three frames from the respective segment. Cyan text and arrows are used to highlight questions related to the actions of the bus; whereas orange text and arrows draw attention to the bicycle. The frame numbers displayed on the image in red (top-left corner) are also included only for visualization. In practice, the red numbers are not superimposed on the images that are used as input to the VLMs.

[1] https://github.com/zhijian11/RoboTron-Drive/blob/main/scripts/inference_demo/demo_video.py

[2] https://github.com/nvidia-cosmos/cosmos-reason1/blob/main/scripts/inference_sample.py

C. Scene-CoT: Additional Details

In the main body of the paper, the critical details of the proposed Scene-CoT method were provided. Indeed, the inspiration of the Scene-CoT method was to design a general baseline that divides the QA problem into a VLM-based video perception module and an LLM-based reasoner. There is related ongoing research that shows the benefits of such perception and reasoning decoupling for solving complex math and science problems (Gou et al., 2025; Guo et al., 2025b) as well as for RL-based finetuning (Rezaei et al., 2026). Here, more details of the Scene-CoT approach, including the overall system diagram (shown in Figure 8), are provided. Of note, the VLM CoT reasoning block shows the information flow during the series of CoT VLM calls. As discussed in the main paper, general scene descriptions are first generated using a uniformly sampled sub-set of frames in each segment. Using these scene descriptions as additional context, the VLM is prompted to describe the ego vehicle motion and subsequently the nearby vehicles' motion. The VLM also receives the additional context of the ego vehicle motion descriptions, as well as the scene descriptions, when it is prompted to describe the motion of the nearby vehicles. The final step in the CoT generates a JSON-formatted output summary using the ego and nearby motion descriptions as contextual information. Finally, to perform question-answering, the LLM is provided with three inputs: 1) The original question. 2) The general Scene Description. 3) The JSON-style motion summary.

A concrete example from TAD showing the series of CoT prompts, along with the VLM's outputs is shown in Figure 9.

D. Additional Ablation Studies

In the main body of the paper, ablations were presented that showed accuracy breakdowns for ego and non-ego questions, a Blind VLM test, as well as performance generalization on STSBench (Fruhwirth et al., 2026). In this section, results for five additional ablation experiments are presented. The ablation experiments are as follows: 1) **Ego vs. Non-Ego Results.** Additional details and results further analyzing model accuracies on ego and non-ego questions. 2) **Captioning Style.** Ablations on the prompting style used in Scene-CoT, for creating the segment descriptions. 3) **QA LLM Selection.** An ablation experiment considering the effect of the specific LLM used in Scene-CoT. 4) **Object**

Table 7. TAD Question Types and Examples

Task	Description	Example
Exact Answer Action Recognition	Evaluates if a model can perform vehicle action recognition, on a short, temporally cropped video segment. This segment-level question type isolates the VLM's ability to perform fine-grained action recognition without concerning longer term temporal understanding in a full-length video. The model is expected to produce the exact textual answer from a list of action options.	**Question:** What best describes the motion of the bus visible in this video segment? Respond with exactly one full phrase from the following list: 'Starting', 'Stopping', 'Turn left', 'Turn right', 'Change lane to the left', 'Change lane to the right', 'Straight, constant speed', 'Stopped' **Answer:** *Turn left*
Multiple Choice Action Recognition	Evaluates if a model can understand the action of a vehicle in a short, temporally cropped video segment. This segment-level question type isolates the VLM's ability to perform fine-grained action recognition without concerning longer term temporal understanding in a full-length video.	**Question:** What is the motion of the ego vehicle in this video segment? Respond with exactly one letter corresponding to the correct option. A. Turn left. B. Straight, constant speed. C. Stopped. D. Starting. **Answer:** *D*
Action Duration	Assesses if a model can understand the durations of various events that transpire in a video. The question asks which event lasts the longest/shortest in a video and is presented with several options in a multiple choice setting. The actions for this question type are all related to the ego vehicle.	**Question:** Which action did the ego vehicle spend the most time doing? Respond with exactly one letter corresponding to the correct option. A. Straight, constant speed. B. Starting. C. Stopping. D. Stopped. **Answer:** *A*
Temporal Ordering	Determines if a model can properly identify the temporal sequence of actions performed by a specific vehicle in the scene. Various action sequences are presented as multiple choice options.	**Question:** Which of the following represents the correct temporal order of motions for the ego vehicle? Respond with exactly one letter corresponding to the correct option. A. Starting; Straight, constant speed; Stopping. B. Stopping; Starting; Straight, constant speed; C. Starting, Stopping; Straight, constant speed; D. Stopping; Straight, constant speed; Starting. **Answer:** *C*
Temporal Action Localization	Evaluates if a model can identify the frames where a specific action is being performed by a vehicle. The expected output is a list of the frames that correspond to the specified actions.	**Question:** In which frames is the ego vehicle doing action 'stopped'? Respond with an explicit list of all frame numbers, enumerating each value individually, without summarizing as a range. **Answer:** *[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]*
Relative Temporal Action Localization	Assesses if a model can identify which of two vehicle actions happened earlier/later in the video. Whereas temporal localization exactly localizes a single action, this task requires localization of two actions and determining their relative temporal ordering. The action answers are presented as multiple choice options.	**Question:** Which ego vehicle event happened earlier in the video? Respond with exactly one letter corresponding to the correct option. A. Stopping. B. Starting. **Answer:** *B*
Temporal Object Localization	Evaluates if a model can identify the frame ranges where a specific object is present. This task disentangles the challenges of action recognition and temporal localization (i.e., to determine if the VLM can identify the frames where a **non-action** event occurs in a video). The expected output is a list of the frames that correspond to the question criteria.	**Question:** In which frames are there any trucks visible to the ego vehicle? Respond with an explicit list of all frame numbers, enumerating each value individually, without summarizing as a range. **Answer:** *[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]*

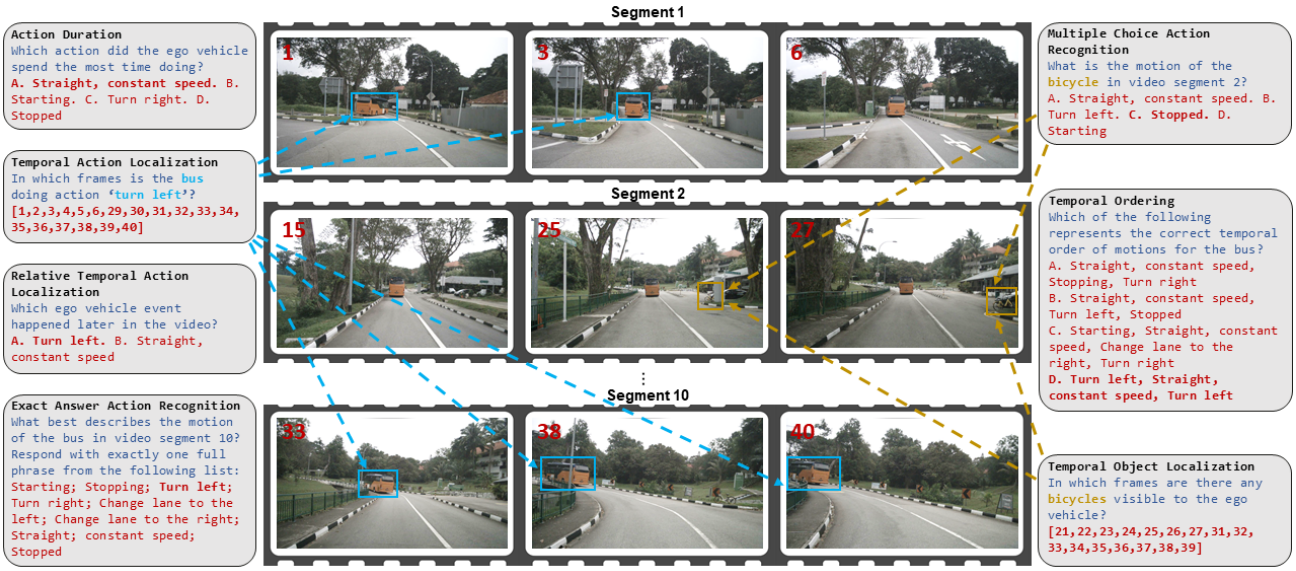

Figure 7. Pictorial visualization example of the question types and format in the TAD benchmark.

detections as context. An ablation that considers using the output of an object detector to improve captioning quality. 5) **Inference Time.** An analysis of inference time for the methods.

D.1. Ego. vs Non-Ego Performance Gains: Additional Results and Analysis

As mentioned in the main body of the paper, the questions related to the ego and non-ego vehicles are relatively balanced; however, it is still important to consider whether the performance varies between these two question categories.

In the main body of the paper, Tab. 5 shows a summarized version of the results for the ego vs. non-ego vehicle ques-

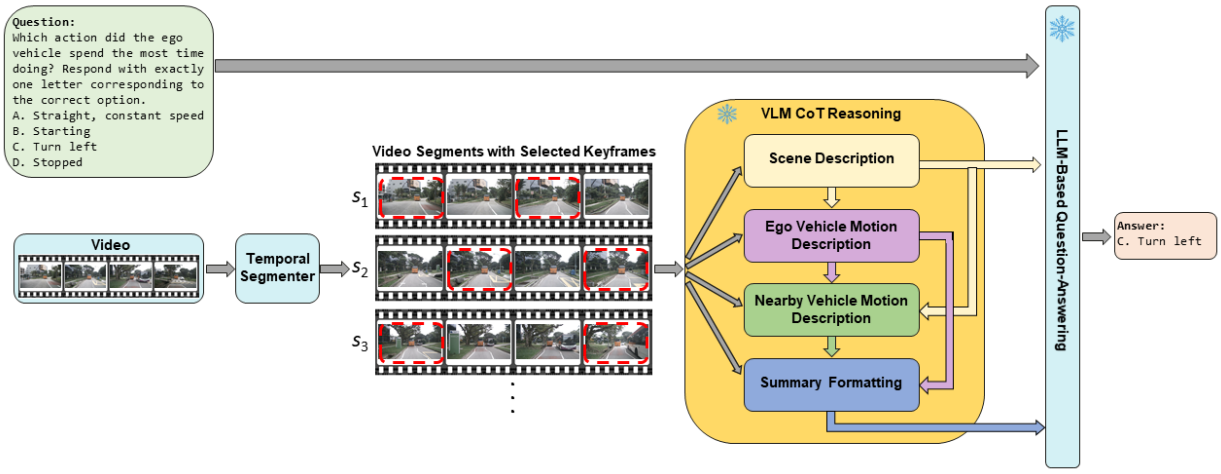

Figure 8. Overview of the proposed Scene-CoT method.

tions by presenting the **average** accuracies across all question types for the InternVL3-8B model. Here, these results are significantly expanded upon along two dimensions: 1) **Additional models.** The ego vs. non-ego breakdown is provided for six additional models (InternVL3-14B, InternVL3-38B, Qwen2.5-VL-7B, Qwen2.5-VL-32B, RoboTron, and Cosmos-Reason). This set of models forms the complete set of open-source generalist and specialist models that were considered in Tab 4 in the main body. 2) **Task-Wise Accuracies.** The ego vs. non-ego question accuracies are provided on a per-type basis, permitting a more fine-grained results analysis. The detailed ego vs. non-ego vehicle accuracies are presented in Table 8 and Table 9. The format of these two tables matches that of the main results table in the body of the paper, but each table only evaluates a sub-set of the questions (i.e., ego questions in Table 8 and non-ego question in Table 9). Note that "N/A" in a column indicates that the corresponding task does not include either ego-related or non-ego-related questions.

A comparison of the results in Table 8 and Table 9 shows that the impact of Scene-CoT varies notably across models. While some models, such as Qwen2.5-VL-7B, exhibit substantially larger gains on non-ego vehicle queries (+8.11% vs. +3.38% for ego), others display the opposite pattern. For instance, RoboTron benefits most on ego-centric questions (+10.81%), suggesting that certain models are more capable of exploiting Scene-CoT's structured reasoning for inferring ego-motion cues.

With TCogMap, Table 8 shows significant accuracy improvements on ego-related questions (typically around 15%) and interestingly, more modest increases are still generally observed in Table 9 for non-ego questions. These results indicate that incorporating the ego vehicle's motion offers useful contextual cues about the behavior of surrounding objects, leading to more reliable temporal reasoning by the VLM.

D.2. Additional Experiments on CoT Descriptions for Scene-CoT

In the main body of the paper, details of the four-step CoT reasoning process were described, providing motivation for dividing the problem into sub-tasks that focus on general scene descriptions, ego vehicle motion, motion of nearby vehicles and summarizing/formatting. However, one may question the following: "Why was this four step process selected?" and "How sensitive are the benchmark results to the specific CoT prompts that are used?" These questions are answered with this ablation study.

Table 10 compares different styles of descriptions generated during CoT reasoning for Scene-CoT. Three configurations were considered. With the Scene Description configuration, only the first step from the four-step CoT reasoning process was used to describe the video segments. In other words, for each video segment, only a high-level scene description was used for the LLM-based QA. This scenario can essentially be thought of as a CoT-free approach, where a general video description or caption is produced. In the second configuration, CoT Output Summary, only the final step of CoT reasoning was used for the segment descriptions during QA. The output from Step 4 is a formatted summary derived from the first three steps, so it is expected to be relatively complete. Finally, the "Scene Description + CoT Output Summary" configuration concatenates the output of Step 1 and Step 4 to form the segment descriptions. This is the configuration used for the results reported in the main body of the paper.

The average accuracies for these three segment description methods are shown in Table 10. The descriptions that

Video Segment 1

CoT Prompts and Outputs from Scene-CoT on TAD

Step 1: Scene Description

Prompt 1:
You are an expert in autonomous driving scene understanding. You are given a sequence of frames from the front camera of a vehicle. Your first task is to provide a concise, high-level description of the overall scene including the nearby vehicles with distinguishable descriptions. Describe the scene in a few sentences. Focus only on what you can see.
Step 1 Output:
The scene depicts a quiet, tree-lined road with a gentle curve to the left. On the right side, there is a small building with a brown roof, surrounded by lush greenery. A car is parked near the building. The road is marked with white and black curbs, and there is a pedestrian shelter on the left side. The area appears calm and residential.

Step 2: Ego Vehicle Motion Descriptions

Prompt 2:
Based on the provided frames and your previous scene description:---<step1_output>---Now, focus *only* on the ego vehicle's motion. Analyze how the background and lane markings move across the frames. Is the ego vehicle moving forward, turning, changing lanes, starting from a stop, or stopping? Describe your reasoning in one or two sentences, and then conclude with the most likely motion type.**Your final conclusion must be one of these exact phrases**:- Stopped- Turn left- Turn right- Change lane to the left- Change lane to the right- Starting- Stopping- Straight, constant speed
Step 2 Output:
*The background and lane markings appear to move consistently towards the viewer, indicating the ego vehicle is moving forward at a constant speed.\n\n**Conclusion**: Straight, constant speed"*

Step 3: Nearby Vehicle Motion Descriptions

Prompt 3:
Excellent. Your analysis of the scene and the ego vehicle's motion was:---<step1_output><step2_output>---Now, analyze nearby vehicles (cars, trucks, buses, bicycles, motorcycles, trailer, construction vehicle) in front of or near the ego vehicle. For each distinct vehicle, provide a brief reasoning and conclude with its most likely motion type (including stopped) from the following list:- Stopped- Turn left- Turn right- Change lane to the left- Change lane to the right- Starting- Stopping- Straight, constant speed If there are no other vehicles or their motion is unclear, state that.
Step 3 Output:
*There are no other vehicles visible in front of or near the ego vehicle in the provided frames. The motion of any potential vehicles is unclear due to the lack of visible vehicles.\n\n**Conclusion**: No other vehicles detected"}*

Step 4: Summary Formatting

Prompt 4:
Based on all the previous analysis: ---Ego Vehicle Analysis Summary:<step2_output>---Nearby Vehicles Analysis Summary:<step3_output>---Your final task is to consolidate this information into a single JSON object. - First, extract the final motion classification for the ego vehicle and each nearby vehicle from your previous responses.- Second, map each classification to the corresponding descriptive phrase using the provided dictionary.- Finally, construct the JSON object.
Motion Phrases Dictionary:{json.dumps(<classification_to_phrase>, indent=2)}**Output Format**:{{
"ego_vehicle_motion": "<phrase from dictionary>", "nearby_vehicles_motion": [{{ "vehicle_id": "<a brief, unique description, e.g., 'white SUV in front'>", "motion": "<phrase from dictionary>" }}, ...]}}
Important: Respond with *only* the raw JSON object and nothing else. Do not wrap it in markdown or add any explanations."""
Step 4 Output:
Description of video keysegment 1:\n ```json\n{\n \"ego_vehicle_motion\": \"moving forward at a relatively constant speed.\",\n \"nearby_vehicles_motion\": []\n}\n ```

NOTE:
classification_to_phrase is defined as:

```
classification_to_phrase = {
"Stopped": "stopped.",
"Turn left": "turning to the left.",
"Turn right": "turning to the right.",
"Change lane to the left": "changing lane to the left.",
"Change lane to the right": "changing lane to the right.",
"Starting": "starting from a stop.",
"Stopping": "stopping.",
"Straight, constant speed": "moving forward at a relatively constant speed."  }
```

Figure 9. Exact prompts and sample output with Scene-CoT for a question in TAD.

consist of scene descriptions alone are relatively lower at 46.26%. General scene descriptions are likely too impoverished to properly answer the breadth of questions considered in TAD. General scene descriptions tend to focus on the overall environment along with a few prominent scene objects. However, the question set of TAD focuses on many aspects of the scene that may not appear in a general description: vehicle actions, less prominent vehicles in the background, and ego vehicle actions. Using the final output from the CoT procedure shows a significant improvement, which is

Table 8. Results of **ego-related** questions for open- and closed-source generalist VLMs as well as AD-specific VLMs. Bold and underline indicate best and second-best results when comparing the three system variants for each open-source VLM. Gen : generalist models, Drv : driving specialist models. **EA**: exact answer, **MC**: multiple choice.

Method	Type	EA Act. Recog.	MC Act. Recog.	Action Dur.	Temp. Ord.	Temp. Act. Local.	Rel. Temp. Local.	Temp. Obj. Local.	Avg
Qwen2.5-VL-7B	Gen	63.53	58.93	34.68	56.45	35.00	46.74	N/A	49.22
+ Scene-CoT		65.13	56.60	**52.42**	54.84	37.69	48.91	N/A	52.60
+ TCogMap		**79.67**	**79.32**	50.81	**62.90**	**48.87**	**69.57**	N/A	**65.19**
Qwen2.5-VL-32B	Gen	63.47	55.63	47.58	58.06	46.09	60.87	N/A	55.81
+ Scene-CoT		67.93	60.58	46.77	54.84	36.31	44.57	N/A	51.83
+ TCogMap		**80.53**	**78.35**	**58.87**	**74.19**	**46.23**	**78.26**	N/A	**69.83**
InternVL3-8B	Gen	71.60	62.33	53.23	51.61	39.98	52.17	N/A	55.15
+ Scene-CoT		67.60	58.93	51.61	56.45	42.08	51.09	N/A	54.63
+ TCogMap		**79.13**	**76.50**	**57.26**	**72.58**	**56.15**	**67.39**	N/A	**68.17**
InternVL3-14B	Gen	68.87	60.87	52.42	59.68	43.52	59.78	N/A	57.52
+ Scene-CoT		68.60	59.51	51.61	54.84	42.80	51.09	N/A	54.74
+ TCogMap		**80.53**	**77.67**	**58.06**	**80.65**	**55.83**	**75.00**	N/A	**71.29**
InternVL3-38B	Gen	55.67	65.73	45.97	53.23	46.15	71.74	N/A	56.42
+ Scene-CoT		70.00	62.23	53.23	56.45	44.48	55.43	N/A	56.97
+ TCogMap		**80.53**	**78.06**	**60.48**	**74.19**	**51.26**	**80.43**	N/A	**70.83**
RoboTron	Drv	58.27	71.26	4.84	43.55	29.78	65.22	N/A	45.49
+ Scene-CoT		**71.07**	64.17	**52.42**	**58.06**	37.70	54.35	N/A	56.30
+ TCogMap		65.73	**77.77**	37.90	**58.06**	**46.46**	**79.35**	N/A	**60.88**
Cosmos-Reason	Drv	66.73	61.55	39.52	54.84	16.66	64.13	N/A	50.57
+ Scene-CoT		65.87	57.96	**55.65**	59.68	34.78	55.43	N/A	54.90
+ TCogMap		**80.13**	**79.71**	50.00	**72.58**	**40.15**	**67.39**	N/A	**64.99**

derived from a richer summary description that includes details of both the ego vehicle's and other scene vehicles' motions. However, best results are achieved when both the scene description and final CoT output are concatenated. This trend indicates that explicit scene-level context complements the structured reasoning provided by CoT, yielding more comprehensive representations for downstream question answering.

D.3. Object detections as context

This ablation considers the possibility of leveraging knowledge from an "external expert" detector agent to improve the VLM performance in the proposed methods. In this case, ablation experiments were completed on a modified version of the Scene-CoT pipeline to explore the effect on reasoning trace quality when information from an object detector is provided.

As shown in Figure 8, the VLM CoT Reasoning module is used to generate a descriptive reasoning trace for each video segment, s_j, which are subsequently used by an LLM to perform question answering. In the Scene-CoT approach, for a specific video segment, four keyframes are provided to the VLM to perform the four steps of the CoT reasoning. In this ablation, one additional pre-processing step is added prior to VLM CoT Reasoning. Specifically, the Ultralytics YOLO11x (Jocher & Qiu, 2024) object detector is first run on each of the keyframes in a segment. The output of the detector is then post-processed to provide a textual summary denoting the count of the relevant vehicle categories that are visible within that keyframe. The relevant object categories are defined as those that overlap with the vehicle categories used in TAD, which are: bicycle, car, motorcycle, bus, and truck. As an explicit example, the post-processed detector output for one keyframe has the format 'Frame1: 5 cars, 2 bicycles'. This object detection summary string is computed for each of the keyframes in a video segment and then provided as additional context to the first step of the CoT Reasoning module that computes the summary Scene Description. The rationale behind this design is that the expert object detector can provide a prior to the VLM that can resolve ambiguities in challenging situations when the VLM is less confident about the presence of certain objects.

This ablation study was performed using four VLMs, Qwen2.5-VL-7B/32B and InternVL3-8B/14B, the results of which are shown in Table 11. It appears there may be some promise in adding external knowledge via an expert model, since the addition of the YOLO11x detection summary im-

Table 9. Results of **non-ego** questions for open- and closed-source generalist VLMs as well as AD-specific VLMs. Bold and underline indicate best and second-best results when comparing the 3 system variants for each open-source VLM. Gen : generalist models, Drv : driving specialist models. **EA**: exact answer, **MC**: multiple choice.

Method	Model Type	EA Act. Recog.	MC Act. Recog.	Action Dur.	Temp. Ord.	Temp. Act. Local.	Rel. Temp. Local.	Temp. Obj. Local.	Avg
Qwen2.5-VL-7B	Gen	26.36	40.61	N/A	27.78	35.50	N/A	45.60	35.17
+ Scene-CoT		**47.10**	42.72	N/A	**55.56**	25.35	N/A	45.65	43.28
+ TCogMap		39.86	**47.35**	N/A	50.00	**44.23**	N/A	**57.97**	**47.88**
Qwen2.5-VL-32B	Gen	56.61	50.00	N/A	50.00	**44.08**	N/A	**62.91**	**52.72**
+ Scene-CoT		50.27	45.24	N/A	**55.56**	27.84	N/A	48.99	45.58
+ TCogMap		**60.05**	**55.29**	N/A	33.33	37.65	N/A	**62.91**	49.85
InternVL3-8B	Gen	37.77	**56.48**	N/A	44.44	31.86	N/A	45.64	43.24
+ Scene-CoT		**55.62**	48.02	N/A	**55.56**	33.57	N/A	50.34	**48.62**
+ TCogMap		37.23	53.31	N/A	44.44	**39.98**	N/A	**64.10**	47.81
InternVL3-14B	Gen	40.58	43.92	N/A	44.44	43.53	N/A	43.90	43.27
+ Scene-CoT		**53.17**	45.11	N/A	**61.11**	31.76	NA	51.64	**48.56**
+ TCogMap		46.47	**48.41**	N/A	38.89	**45.87**	N/A	**59.00**	47.73
InternVL3-38B	Gen	43.93	48.28	N/A	61.11	**44.84**	N/A	**68.27**	53.29
+ Scene-CoT		**55.07**	46.96	N/A	61.11	33.42	N/A	53.77	50.07
+ TCogMap		48.64	**54.89**	N/A	**66.67**	39.23	N/A	66.43	**55.17**
RoboTron	Drv	23.73	40.34	N/A	33.33	31.95	N/A	28.92	31.65
+ Scene-CoT		**41.85**	37.30	N/A	**55.56**	24.68	N/A	34.79	**38.84**
+ TCogMap		24.37	**42.33**	N/A	33.33	**35.21**	N/A	**40.58**	35.16
Cosmos-Reason	Drv	28.35	32.80	N/A	44.44	16.18	N/A	16.06	27.57
+ Scene-CoT		**38.77**	34.13	N/A	**61.11**	24.10	N/A	**43.72**	**40.37**
+ TCogMap		28.53	**44.05**	N/A	50.00	**33.46**	N/A	23.97	36.00

Table 10. Effect of using different sections of the CoT outputs to form the segment descriptions.

Description Style	Average Accuracy
Scene Description	46.26
CoT Output Summary	51.13
Scene Description + CoT Output Summary	51.92

Table 11. Effect of adding object detector context as input to the VLM CoT Reasoning module. Numbers denote the average accuracy across all seven tasks in the TAD benchmark.

VLM Model	Scene-CoT	Scene-CoT + YOLO11x	Difference
Qwen2.5-VL-7B	48.71	49.17	0.46
Qwen2.5-VL-32B	48.78	49.56	0.78
InternVL3-8B	51.92	52.68	0.76
InternVL3-14B	51.81	51.93	0.12
Average	50.31	50.84	0.53

proved average accuracy across all four tested models. However, the gains are quite modest, with Qwen2.5-VL-32B seeing the biggest improvement of 0.78 and an average improvement of only 0.53 across the four models. Overall, this result suggests that although there is some promise, the inclusion of an additional detector agent may not be worth the extra system complexity.

D.4. Inference Time Analysis

While inference time was not a primary focus of our evaluation, we further analyzed the inference time for completeness. To measure the inference time, roughly 200 questions across three different question types (*Ego Action Duration*, *Relative Temporal Localization*, and *Temporal Ordering*) were used, the results for which are shown in Table 12

Four configurations were considered: the two baselines (with and without additional ego pose context) and the two proposed solutions (Scene-CoT and TCogMap). As the table shows, the baselines and TCogMap have quite comparable inference times. This result is due to the fact that the image tokens for all three methods are identical but the textual tokens are slightly different, depending on the specific contextual information that is added by the method. TCogMap has the additional overhead of computing the ego vehicle temporal cognitive map and classifying the ego vehicle actions, but these processing steps have a negligible effect on latency.

For Scene-CoT, on the other hand, a much higher inference time is observed. This large inference time is primarily due

Table 12. Average inference time for the baselines and proposed solutions.

Method	Inference time (s)
InternVL3-8B	2.20
InternVL3-8B + EP	2.72
+ Scene-CoT	47.10
+ TCogMap	2.21

to the increased number of VLM inference calls, since CoT descriptions are provided for each segment of the video. Specifically, for the experiments in this paper, the videos are divided into ten segments, each of which is described using four CoT processing steps, yielding a total of 40 VLM inference calls. This design results in an inference time of 42 seconds per question to generate the segment descriptions. In comparison, the actual LLM question-answering inference time is rather minor, at under 5 seconds per question.

Although this run-time of Scene-CoT can be viewed as a disadvantage, there is significant ongoing research on VLM inference optimization, including quantization (e.g., (Frantar et al., 2022; Tseng et al., 2024; Egiazarian et al., 2024; Gholami et al., 2025)) and pruning (Chen et al., 2024a; Lin et al., 2025; Alvar et al., 2025), that can provide significant speed improvements with limited accuracy losses. Moreover, the experiments in this paper were performed in a single inference setting, but batch inference could be used to parallelize the description generation across the segments, resulting in significant inference speed improvements (i.e., up to 10×, in the case of the experiments presented in this paper). Thus, there is significant scope to reduce the runtime of Scene-CoT.

D.5. Effect of LLM on Scene-CoT

In the main body of the paper, the results for Scene-CoT were generated using Qwen2.5-14B-Instruct-1M as the LLM for question answering. This LLM was selected through an empirical evaluation process. This procedure, as well as the effects of utilizing different LLMs, are described next.

In Table 13, the effect of changing the question answering LLM for Scene-CoT is shown. A clear accuracy improvement of 2.21% was observed when moving from the 7B Qwen2.5-7B-Instruct to the comparable 14B model, Qwen2.5-14B-Instruct. This observation motivated the selection of a 14B LLM that can offer greater reasoning power. Among the 14B models evaluated, Qwen2.5-14B-Instruct-1M achieves the highest average accuracy of 51.92%. The comparison between the standard and "1M" versions of the 14B model indicates that extended context capacity and stronger instruction tuning improve reasoning over temporally grounded inputs. Additionally, Qwen3-14B with its "thinking" mode enabled achieves comparable accuracy to Qwen2.5-14B-Instruct-1M; however, Qwen2.5-14B-Instruct-1M was ultimately selected due to its higher throughput.

Overall, the trends indicate that there is a 3-4% improvement when moving to the larger 14B LLMs and that amongst the 14B models, the accuracies are fairly comparable.

Table 13. Effect of using different LLMs for question answering in Scene-CoT. Average accuracy is computed over all seven tasks.

LLM QA Model	Average Accuracy
Qwen2.5-7B-Instruct	47.87
Qwen2.5-14B-Instruct	50.08
Qwen2.5-14B-Instruct-1M	**51.92**
Qwen3-14B (thinking enabled)	<u>51.05</u>

E. Data Annotation Interface

Figure 10 illustrates the annotation interface used to collect vehicle-action labels for TAD. Since labeling the actions of specific vehicles, including the ego vehicle, requires a level of granularity not supported by existing annotation tools, a custom interface is implemented. Each video segment is displayed alongside its corresponding bird's-eye-view (BEV) visualization, which shows the ego trajectory and the motion of nearby vehicles. Annotators view both modalities jointly to assess the dominant behavior of each target vehicle within the segment. A drop-down menu is provided for selecting the appropriate motion label for the target object, and the final annotation is saved through the interface. This setup ensures consistent access to both appearance and trajectory cues, facilitating reliable segment-level action labeling.

F. Qualitative Examples

Figures 11–14 present qualitative examples corresponding to Tasks *Multiple Choice Action Recognition*, *Exact Answer Action Recognition*, *Action Duration*, and *Relative Temporal Action Localization*, respectively, complementing the quantitative results in the main table of the paper. Each example includes representative frames from the relevant segment or full scene, the question posed, the ground-truth answer, and the model (i.e, InternVL3-8B) outputs under different inference settings. The baseline corresponds to the model receiving only the frames and the question. The "+Scene-CoT" setting applies the proposed Scene-CoT reasoning procedure, while "+TCogMap" shows the outputs produced using the proposed TCogMap. Across tasks, the qualitative results show that applying Scene-CoT or TCogMap improves the model's predictions, often correcting cases where

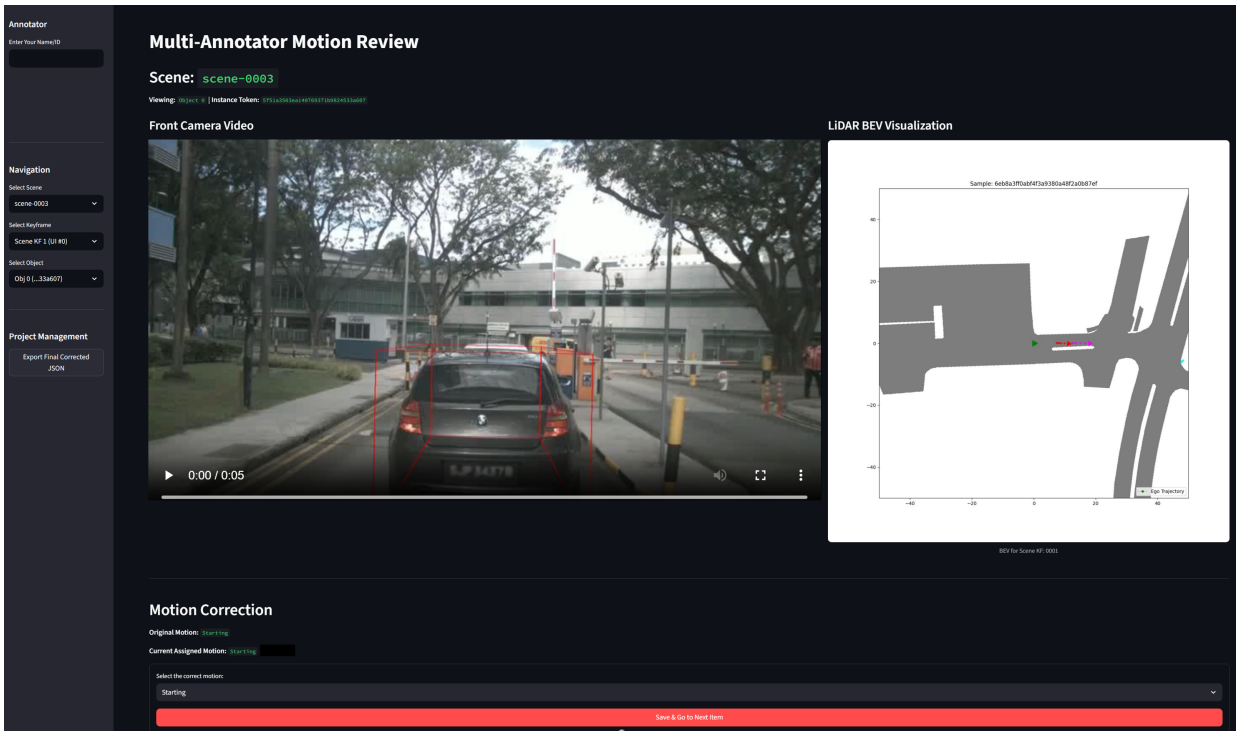

Figure 10. Data annotation pipeline.

the baseline fails, underscoring the value of both methods for strengthening temporal and scene-level reasoning.

In more detail, in Figure 11, the ego vehicle goes straight toward a barrier, which then lifts to allow the ego vehicle to pass. The baseline selects an incorrect answer, whereas both Scene-CoT and TCogMap produce the correct output. In Figure 12, the query concerns the motion of the car visible in the frames. Because the scene occurs at night and the rear lights of the vehicle are illuminated red, both the baseline and Scene-CoT incorrectly predict that the car is "Stopped." In contrast, TCogMap leverages the ego trajectory and scene dynamics to correctly identify the action of the vehicle ahead as "Straight, constant speed". In Figure 13, the video shows the ego vehicle first approaching a parked bus. The ego vehicle performs a long, gradual lane change to the right to avoid the bus, followed by a quick lane change back to the left. The baseline and Scene-CoT both predict incorrect answers, whereas TCogMap is able to differentiate between the left and right lane changes to provide the correct answer of "Change lane to the left". In Figure 14, the video first shows the ego vehicle driving straight down a city road. Thereafter, the ego vehicle performs a series of short, consecutive atomic actions: a lane change to the left, followed by a left hand turn around a corner, and finally a lane change to the right. The baseline incorrectly selects "Change lane to the right" as its answer. In contrast, Scene-CoT and TCogMap are both able to correctly assess that the ego vehicle drives straight at the beginning of the video, while the lane change to the right occurs near the end, thus allowing them both to provide the correct answer of "B".

A sample of the exact prompts used during the question answering step for the baseline and two proposed approaches is shown in Figure 15, depicting the same video scene that appears in Figure 14. Note that for the baseline and TCogMap approaches, question answering is performed via a VLM, whereas for Scene-CoT, the final step of answering the question is completed by an LLM.

As Figure 15 shows, the prompt for the baseline method is straightforward, consisting merely of the frame labels and corresponding image token tags, as well as the question text itself. TCogMap extends the baseline prompt by adding a motion summary, derived from the temporal cognitive map, for each segment of the video. As shown quantitatively in the main body of the paper, this additional context significantly improves the average accuracy on TAD for all of the VLMs that were considered. For Scene-CoT, since the prompt is being used with an LLM, the reference to frames and image tokens are removed. Instead, the prompt consists of the question itself along with textual descriptions for each video segment. These descriptions were formed using the proposed CoT procedure and include a scene description as well as a JSON-style summary. As shown in the main body of the paper, Scene-CoT, with its CoT reasoning process and LLM-based question answering, generally improves accuracies on the TAD benchmark for smaller models.

Video Frames	
Question	What is the motion of the ego vehicle in this video segment? Respond with exactly one letter corresponding to the correct option. A. Starting, B. Straight, constant speed, C. Stopping, D. Stopped
GT	B
Baseline	D. Stopped
+ Scene-CoT	B
+ TCogMap	B. Straight, constant speed

Figure 11. Qualitative example for *Multiple Choice Action Recognition* task using the baseline (InternVL3-8B) and the proposed methods. **Green** text indicates a correct answer, **red** denotes an incorrect answer.

Video Frames	
Question	What best describes the motion of the car visible in this video segment? Respond with exactly one full phrase from the following list: 'Starting', 'Stopping', 'Turn left', 'Turn right', 'Change lane to the left', 'Change lane to the right', 'Straight, constant speed', 'Stopped'.
GT	Straight, constant speed
Baseline	Stopped
+ Scene-CoT	Stopping
+ TCogMap	Straight, constant speed

Figure 12. Qualitative example for *Exact Answer Action Recognition* task using the baseline (InternVL3-8B) and proposed methods. **Green** text indicates a correct answer, **red** denotes an incorrect answer.

Video Frames	
Question	Which action did the ego vehicle spend the least time doing? Respond with exactly one letter corresponding to the correct option. A. Straight, constant speed, B. Change lane to the right, C. Change lane to the left
GT	C
Baseline	B
+ Scene-CoT	A
+ TCogMap	C

Figure 13. Qualitative example for *Action Duration* task using the baseline (InternVL3-8B) and proposed methods. **Green** text indicates a correct answer, **red** denotes an incorrect answer.

Video Frames	
Question	Which ego vehicle event happened earlier in the video? Respond with exactly one letter corresponding to the correct option. A. Change lane to the right, B. Straight, constant speed
GT	B
Baseline	A
+ Scene-CoT	B
+ TCogMap	B

Figure 14. Qualitative example for *Relative Temporal Action Localization* task using the baseline (InternVL3-8B) and proposed methods. **Green** text indicates a correct answer, **red** denotes an incorrect answer.

Video Scene

Question Answering Prompts

Baseline VLM Prompt:
Frame1: <IMAGE_TOKEN>
Frame2: <IMAGE_TOKEN>
…
Which ego vehicle event happened earlier in the video? Respond with exactly one letter corresponding to the correct option.
A. Change lane to the right
B. Straight, constant speed

Scene-CoT LLM Prompt:
Which ego vehicle event happened earlier in the video? Respond with exactly one letter corresponding to the correct option.
A. Change lane to the right
B. Straight, constant speed
Use the following descriptions of the video keysegments to assist in answering the question. Each keysegment corresponds to a short, five-second snippet of the original video. The original video is of a traffic scene from the perspective of an autonomous driving vehicle. In addition to using the video keysegment captions themselves, use logic, reasoning, and knowledge of traffic rules to answer the question. You may provide some justification, but please enclose the succint, final answer to the question between <answer></answer> tags.
The following information provides a description of video keysegment 1:
Scene Description:
The scene depicts a wide, multi-lane road leading towards a modern, multi-story building with a distinctive architectural design. On the right side of the road, there is a blue taxi and a green bus. Pedestrians are walking on the sidewalk to the left. The road is marked with directional arrows, and traffic lights are visible in the distance. The area is lined with trees and grassy patches, and the sky is overcast.
Summary Description:
```json
{
  \"ego_vehicle_motion\": \"moving forward at a relatively constant speed.\",
  \"nearby_vehicles_motion\": [
    {
      \"vehicle_id\": \"blue taxi on the right\",
      \"motion\": \"moving forward at a relatively constant speed.\"
    }
  ]
}
...

**TCogMap VLM Prompt:**
Frame1: <IMAGE_TOKEN>
Frame2: <IMAGE_TOKEN>
…
Motion summary for Frame1 to Frame8: The ego-vehicle is moving forward at a relatively constant speed.
Motion summary for Frame2 to Frame11: The ego-vehicle is moving forward at a relatively constant speed.
…
Which ego vehicle event happened earlier in the video? Respond with exactly one letter corresponding to the correct option.
A. Change lane to the right
B. Straight, constant speed",

*Figure 15.* Prompts used for question answering. Note that for the baseline and TCogMap, these prompts are provided to the VLM as input, whereas for Scene-CoT, this prompt is given to the LLM
```