iCLUEDO: CLUE-DRIVEN AGENTIC INVESTIGATION FOR EVIDENCE-GROUNDED TRAFFIC-VIOLATION REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Traffic-violation analysis from dashcam video is a hypothesis-driven, sparse-evidence reasoning problem. The decisive observations may be brief, temporally dispersed, and uncertain, yet a reliable judgment must establish *who* acted, *where* and *when* the behavior occurred, *what* happened, and which visual *evidence* supports the conclusion. We first introduce **CVCV**, a benchmark of 4.65K real-world dashcam videos covering 20 fine-grained behaviors across six California Vehicle Code (CVC) violation families. Each video is represented as a case file containing its violation labels, responsible road user, decisive timestamp, accident context, human-reviewed evidence, and grounding to the corresponding CVC section. We also introduce **iCluedo**, a clue driven agentic framework that addresses this problem through a progression from *structured case memory*, to *competing hypotheses*, to *selective evidence acquisition*, and finally *map the violation to the law-grounded decision*. iCluedo first organize perception tools output into a persistent, ego-centric spatio-temporal case memory that represents road users, traffic controls, events, and their evolving relationships. Segment-level hypothesis agent use this memory to generate competing violation hypotheses, which a violation investigation agent verifies through a selective READ–ASK–INSPECT–DECIDE loop. It retrieves relevant structured evidence, queries specialized scouts to resolve uncertainty, revisits the original video only when a decision-critical clue is missing, contradictory, or unreliable, and selects the interpretation best supported by the accumulated evidence. A visual witness provides observable facts from targeted video intervals, while explicit uncertainty determines whether additional evidence must be acquired. The resulting hypothesis is checked for consistency across the actor, location, time, behavior, and supporting evidence, and is then mapped deterministically to a traffic-law violation. **iCluedo** achieves 50.6% accuracy, outperforming four of five direct VLM baselines.

1 INTRODUCTION

Safety-critical driving understanding requires more than detecting individual road users; it requires reasoning about how their spatial and temporal relationships evolve with roadway geometry, lanes, and traffic controls Malawade et al. (2021). Accident anticipation asks whether a collision may occur, whereas post-hoc incident analysis asks what led to the event Chan et al. (2016); Suzuki et al. (2018); Moura et al. (2025). Traffic-violation analysis asks a stricter question: ***who broke which rule, when, and what evidence proves it?*** Answering it means establishing that a rule applied, that a specific road user's behavior contradicted it, and that no higher-priority obligation excused it. Because decisive observations may be brief, temporally dispersed, or uncertain, this is a sparse-evidence event-reconstruction problem rather than frame-level recognition.

Existing benchmarks do not measure this ability. Accident datasets Chan et al. (2016); Yao et al. (2023); Moura et al. (2025) localize collisions or explain their causes, yet many violations, such as crossing a solid line, involve no collision, and none ties a behavior to the statute it breaks. Traffic-rule benchmarks such as IDKB Lu et al. (2025), DriveQA Wei et al. (2025), and DriveCombo Ma et al. (2026) test rule knowledge on static or simulated scenes: they ask what a driver *should* do, not whether a road user in real footage *did* break a rule. Recent violation work targets temporal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

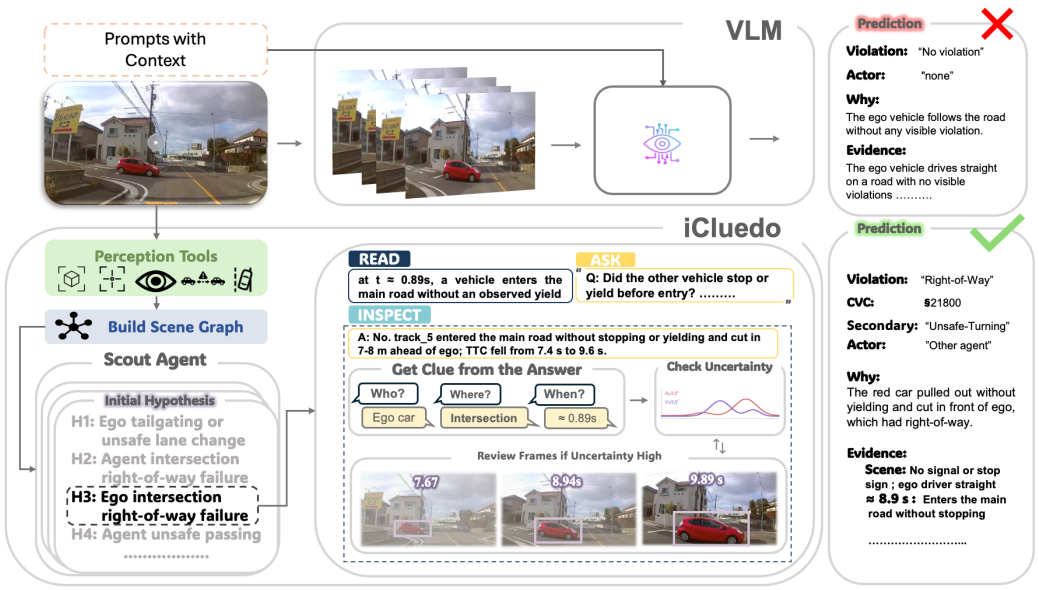


Figure 1: **Overview of iCluedo.**

grounding or fisheye traffic-camera footage, which has no ego vehicle Kim et al. (2026). We therefore introduce **CVCV**, a benchmark of 4.65K real-world dashcam videos spanning 20 fine-grained behaviors in six California Vehicle Code (CVC) violation families, plus no-violation and Unknown categories. We ground labels in the CVC because a behavior is a violation only relative to the rule it breaks, and the code defines each rule within an authoritative hierarchy (division, chapter, article, section), so a system's verdict maps to the same legal sections that insurers and police collision reports use to assign fault. Each video is annotated like a case file: the primary violation, a co-occurring secondary violation when present, the violation time, whether an accident followed and when, the responsible road user (ego or another), and a human-reviewed rationale with visual evidence. Every violation is traced to the exact statute it breaks (e.g., § 21650, driving on the wrong side of the road), so a model can be evaluated on the right violation, by the right actor, at the right moment, under the right law.

We first applied several vision-language models (VLMs) Bai et al. (2025b); NVIDIA (2026) directly to CVCV, with multiple prompting strategies designed to recover the violation and the rest of the case file. None produced reliable predictions for most violation labels (Sec. 5), suggesting that the bottleneck lies not in the prompt but in fixed-context inference. Dense video can obscure the few moments that determine a violation, whereas sparse sampling may omit them entirely Wu et al. (2024). More importantly, a fixed-context model cannot actively resolve missing or contradictory facts, such as which signal governed a vehicle, whether it stopped before the intersection, or whether detections in different frames are the same road user. A plausible label can thus remain unsupported by the video.

Violation analysis is therefore better treated as an investigation than a single-pass judgment. We formulate it as **hypothesis-driven sparse-evidence reasoning** and introduce **iCluedo**, a clue-driven agentic framework for monocular driving video. Perception tools build an ego-centric spatio-temporal case memory of road users, tracks, roadway structure, traffic controls, and ego motion. Segment-level scouts propose competing violation hypotheses, which a violation expert investigates through an adaptive READ–ASK–INSPECT–DECIDE loop, revisiting decision-critical video intervals when visual facts are uncertain. The central novelty is not generic tool use or additional frame retrieval: each evidence-acquisition action is conditioned on an explicit missing, contradictory, or unreliable clue within a candidate hypothesis. Typed uncertainty lets the system request more evidence instead of forcing a conclusion. On CVCV, we were able to achieve better results in detecting

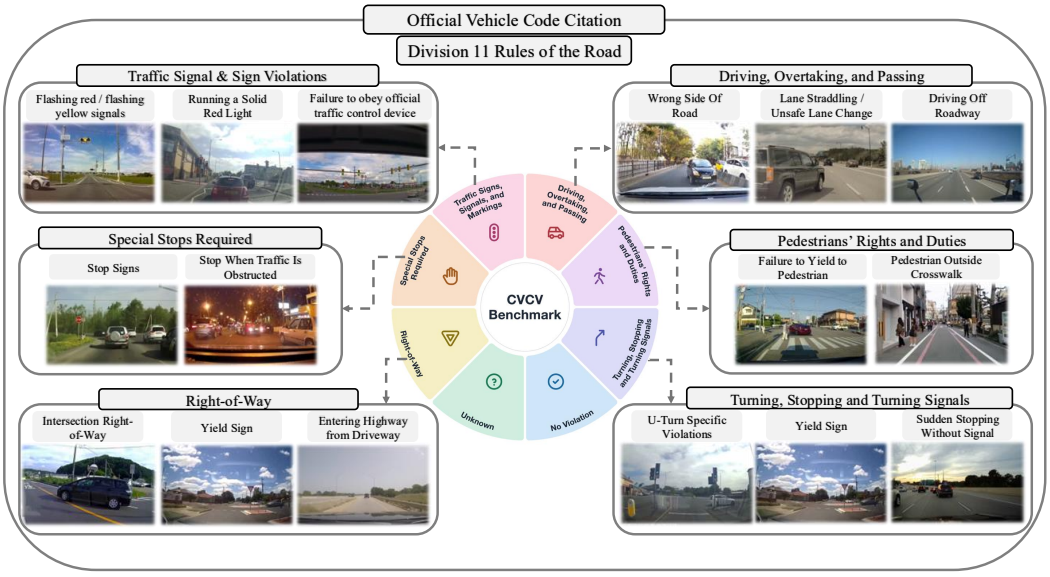


Figure 2: **Overview of the CVCV taxonomy.**

the violation using our agentic framework, iCluedo, over other VLM baseline and also improves actor attribution and temporal localization. Our main contributions are:

- **CVCV benchmark.** We introduce CVCV,, a 4.65K-video real-world dashcam benchmark and, to our knowledge, the first to ground traffic violations in statute. Each video annotations provide a case file and California Vehicle Code (CVC) violation.
- **Clue-driven agentic investigation.** We introduce iCluedo, which treats each video as an investigation rather than a single-pass prediction: it builds a spatio-temporal case memory, weighs competing violation hypotheses, and gathers new evidence only when a specific clue is missing, contradictory, or unreliable, before grounding its verdict in traffic law.
- **Case-level evaluation.** We evaluate VLM baselines and iCluedo on the full case (violation, actor, and violation time) rather than the label alone, exposing errors that label accuracy hides and showing that clue-driven investigation outperforms fixed-context inference.

2 Related Work

Traffic benchmarks. Accident and anomaly datasets such as CCD Bao et al. (2020), DADA-2000 Fang et al. (2019), COOOL AlShami et al. (2024) and DoTA Yao et al. (2023) annotate whether and when a collision or anomaly occurs, but not the rule behind it, while traffic-video QA benchmarks Xu et al. (2021); Li et al. (2025); Zhou et al. (2025) probe causal and interaction reasoning rather than legal compliance. Closest to our setting, recent benchmarks address rules and responsibility: DriveCombo Ma et al. (2026) grounds compositional rule questions in national traffic codes but in CARLA simulation, DecaTARA Zhou & Zhang (2026) allocates accident responsibility over 67.9K real videos, CAViAR Garg et al. (2026) labels at-fault agents and apparent, explicitly non-legal violation families in crash videos, and C-TRAIL Cui et al. (2026) links about 1K accident videos to Chinese traffic-law provisions. Table 1 positions CVCV among related benchmarks. Their rule and responsibility labels, however, are tied to crashes or simulation. CVCV targets violations whether or not a crash follows, recording each as a case file with violation and accident times, the responsible actor, human-reviewed evidence, and the exact CVC section it breaks, alongside no-violation and Unknown cases.

Table 1: Comparison of CVCV with representative benchmarks for accident understanding, traffic-video reasoning, and traffic-rule reasoning. ✓ denotes explicit supervision, while ✗ denotes that explicit supervision is absent or that the property is not a central task. **Law** indicates explicit linkage to an official traffic rule or jurisdictional legal source, while **CVC Code** indicates explicit grounding to a California Vehicle Code section. **Neg./Unk.** indicates explicit normal, no-violation, non-accident, ambiguous, or open-set cases. Dataset sizes follow the primary unit reported by each benchmark and are therefore not directly comparable across clips, videos, and scenes.

Dataset	Setting	Scale	Accid.	Time	Viola.	Time	Agent	Rule	Law	CVC Code	Why / Evidence	Neg. / Unk.
Accident and anomaly understanding												
CCD Bao et al. (2020)	Dashcam	4.5K clips	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓
DADA-2000 Fang et al. (2019)	Dashcam	2.0K videos	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
DoTA Yao et al. (2023)	Dashcam	4.7K videos	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗
Traffic-video and interaction reasoning												
SUTD-TrafficQA Xu et al. (2021)	Mixed view	10.1K videos	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
WOMD-Reasoning Li et al. (2025)	Motion scenes	63K scenes	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗
TUMTraf VideoQA Zhou et al. (2025)	Roadside	1.0K videos	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗
Traffic-rule and responsibility reasoning												
DriveCombo Ma et al. (2026)	CARLA	∼70K MCQ	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
DecaTARA Zhou & Zhang (2026)	Real World	67.9K videos	✓	✓	✗	✗	✓	✗	✗	✗	✓	✓
CAViAR Garg et al. (2026)	Dashcam	2.25K videos	✓	✗	✗	✗	✓	✓	✗	✗	✓	✗
Evidence-grounded traffic-violation reasoning												
CVCV (ours)	**Dashcam**	**4.65K videos**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Agentic video reasoning. Agentic video systems replace fixed dense or sparse video processing with iterative evidence acquisition. VideoAgent Wang et al. (2024) uses an LLM agent to search for and aggregate task-relevant visual observations, while the memory-augmented Agent combines temporal event memory Fan et al. (2025), persistent object memory, and interactive retrieval tools. In autonomous driving, DriveAgent-R1 Zheng et al. (2026) actively invokes visual tools when the initial context is insufficient for planning. iCluedo instead ties every evidence request to a specific missing, contradictory, or unreliable clue: scouts propose competing violation hypotheses from a spatio-temporal case memory, and a violation expert returns to decision-critical video intervals to resolve them before grounding the verified verdict in the CVC.

3 CVCV Benchmark

CVCV (*California Vehicle Code Violations*) contains 4.65K real-world dashcam videos drawn from Nexar Moura et al. (2025), RoadSocial Parikh et al. (2025), and the Car Crash Dataset (CCD) Bao et al. (2020). It is designed so that systems, models, pipelines can be judged not only on the violation it names, but on whether it can reconstruct the event behind that decision: who acted, when, and on what visual evidence. Table 1 contrasts its annotations with prior benchmarks.

Taxonomy. CVCV defines 20 fine-grained behaviors derived from Division 11 (Rules of the Road) of the California Vehicle Code. Each behavior is mapped to the code section it violates and grouped into six chapter-level families: *Traffic Signs, Signals, and Markings*; *Driving, Overtaking, and Passing*; *Right-of-Way*; *Pedestrians' Rights and Duties*; *Turning, Stopping, and Turning Signals*; and *Special Stops Required*. With *no violation*, these form seven closed-set categories. An auxiliary *Unknown* category holds novel or ambiguous scenarios that cannot be reliably assigned to the taxonomy; these clips are excluded from closed-set evaluation (Fig. 2).

Case-file annotation. Each clip i is annotated as a case file,

$$\mathbf{y}_i = \Big(\underbrace{y_i^V, b_i^{(1)}, b_i^{(2)}, c_i;}_{\text{recognition}} \ \underbrace{a_i, t_i^V, y_i^A, t_i^A;}_{\text{event grounding}} \ \underbrace{r_i, e_i}_{\text{evidence grounding}} \Big), \tag{1}$$

where y_i^V and y_i^A indicate whether a violation and an accident occur, $b_i^{(1)}$ and $b_i^{(2)}$ are the primary and optional secondary behaviors, c_i is the CVC section violated by $b_i^{(1)}$, a_i identifies the responsible

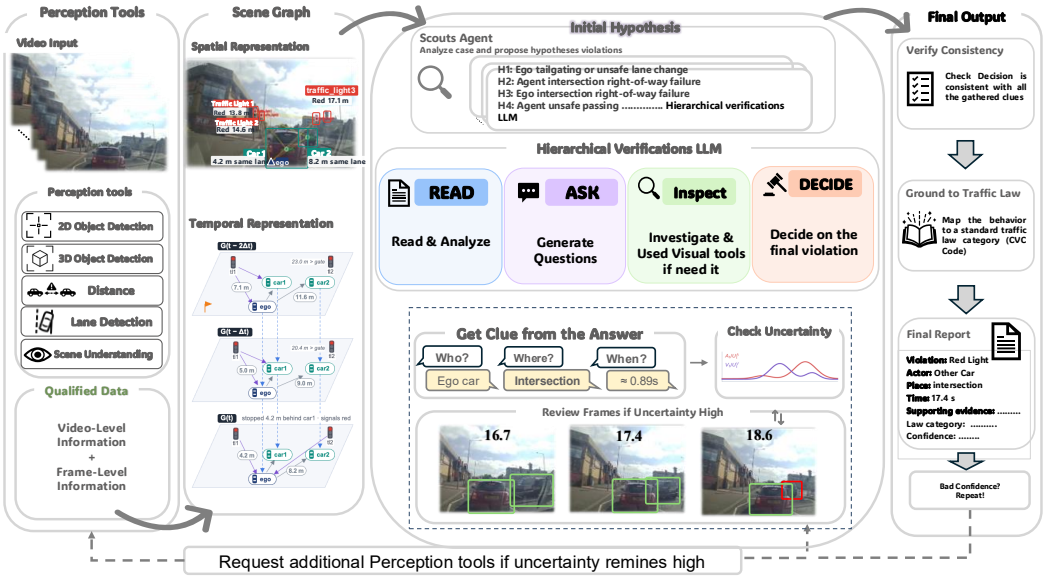


Figure 3: Overview of iCluedo. Perception tools are organized into a persistent spatio-temporal case memory from which scouts generate competing traffic-violation hypotheses. The violation expert identifies missing or unreliable decision-critical clues and selectively retrieves structured evidence, queries scouts, or inspects the original video. The READ, ASK, INSPECT, and DECIDE operations instantiate evidence retrieval, hypothesis refinement, targeted visual verification, and provisional hypothesis selection, respectively. The resulting decision is then checked for evidential consistency and grounded in the traffic-law taxonomy.

road user (ego or another), and t_i^V and t_i^A are the decisive violation and accident times. Human annotators provide all behavior, actor, accident, and temporal labels; a vision-language model drafts the rationale r_i and visual evidence e_i, which annotators then review.

Evaluation levels. The three groups in Eq. equation 1 define three levels of evaluation. **Violation recognition** tests whether a model predicts the correct behaviors and CVC section. **Event grounding** tests whether it attributes the violation to the right road user and localizes the decisive violation and accident times. **Evidence grounding** tests whether its explanation is supported by the visual evidence in the corresponding segment rather than merely plausible. Together, these levels check that a model's verdict rests on the same evidence a human would cite to justify it.

4 ICLUEDO

iCluedo separates *evidence organization* from *evidence acquisition*. As illustrated in Fig. 3, a spatio-temporal scene graph compresses dense perception outputs into a persistent case memory of the actors, events, relations, and evidence already known. An agentic investigation then reasons over competing hypotheses derived from this memory and returns to the original video only when the existing evidence cannot distinguish among them. The system therefore does not process all frames uniformly; it spends additional perception only on the clues that matter most for the final traffic-law decision.

Given a driving video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, the overall pipeline is

$$\mathcal{R} = \mathcal{P}(\mathcal{V}), \qquad (\mathcal{G}, \mathcal{S}, \mathcal{U}) = \Gamma(\mathcal{R}), \qquad h^* = \mathcal{A}(\Phi(\mathcal{G}, \mathcal{S}, \mathcal{U}), \mathcal{V}), \tag{2}$$

where $\mathcal{P}$ denotes the perception tools, Γ constructs the scene graph $\mathcal{G}$ together with per-frame violation-evidence scores $\mathcal{S} = \{S_t\}$ and uncertainties $\mathcal{U} = \{U_t\}$, Φ prepares case memories, and $\mathcal{A}$ conducts the investigation and returns the selected hypothesis h^*. $\mathcal{A}$ also receives $\mathcal{V}$ because its

INSPECT action returns to the raw frames. After verification, h^* yields the output

$$\hat{\mathbf{y}} = \Big(\hat{y}^V, \hat{b}^{(1)}, \hat{b}^{(2)}, \hat{c};\ \hat{a}, \hat{t}^V;\ \hat{r}, \hat{e}\Big), \tag{3}$$

where each field predicts its counterpart in the CVCV annotation (Eq. 1); the accident fields are not predicted.

4.1 PERCEPTION TOOLS

We use the perception toolchain of iFinder Yao et al. (2025). All models are pretrained, publicly available, and frozen; none is trained or fine-tuned on CVCV. OWL-V2 Minderer et al. (2023) detects road users and traffic controls, ByteTrack Zhang et al. (2022) assigns persistent identities, and CenterTrack Zhou et al. (2020) estimates object yaw. Each object's distance to the ego vehicle is read from a Metric3D v2 Hu et al. (2024) depth map inside its SAM Kirillov et al. (2023) mask; OMR Jin & Kim (2024) detects lanes, GeoCalib Veicht et al. (2024) estimates camera intrinsics for undistortion, and DROID-SLAM Teed & Deng (2021) recovers ego motion. InternVL Chen et al. (2024) labels scene and object attributes, including traffic-signal state, and VideoLLaMA2 Cheng et al. (2024) provides a video-level description. Table 4 lists checkpoints, training data, thresholds, frame rate, and runtime. The outputs form a perception record

$$\mathcal{R} = \left(\mathbf{x}^{\text{video}}, \{\mathbf{x}_t\}_{t=1}^{T}\right), \tag{4}$$

where $\mathbf{x}^{\text{video}}$ holds scene attributes, ego motion, and the video description, and $\mathbf{x}_t$ lists each object's class, track ID, 2D box, yaw, distance, lane position, attributes, and motion state in frame I_t.

4.2 SPATIO-TEMPORAL CLUE GRAPH

Γ converts $\mathcal{R}$ into an ego-centric spatio-temporal scene graph

$$\mathcal{G} = \left(\{G_t\}_{t=1}^{T}, \Pi\right), \qquad G_t = (\mathcal{N}_t, \mathcal{E}_t), \tag{5}$$

where $\mathcal{N}_t$ contains the ego vehicle, surrounding road users, traffic controls, and roadway elements, and Π links persistent road-user identities across time. The relations in $\mathcal{E}_t$ (lane membership, following, approaching, crossing, and traffic-control applicability) are derived by deterministic rules over lane assignments, distances, headings, and their temporal changes (Table 5); no part of graph construction is learned. A traffic control is linked to the ego vehicle only within 20 m, which prevents distant signals from producing spurious violations.

Changes in graph state are converted into temporally grounded clues

$$\kappa_k = (a_k, \ell_k, t_k, p_k, e_k), \tag{6}$$

where a_k is the road user involved, ℓ_k its location, t_k the time interval, p_k the event predicate (e.g., a lane transition, entering an intersection under a red signal, failing to stop or yield, wrong-way motion, or a conflict between road users), and e_k the frames and perception fields that support it.

Evidence strength and uncertainty. For each frame t in t_k, clue κ_k carries a severity $s_{k,t} \in [0,1]$, measuring how strongly it indicates a violation, and a reliability $\rho_{k,t} \in [0,1]$, measuring how trustworthy its supporting observations are; both are computed deterministically from $\mathcal{R}$ before any language model is involved. Because gap, closing speed, and time-to-collision may describe the same interaction, clues about the same pair of road users, or of a road user and a traffic control, form a group g_k, and only the strongest corridor-weighted clue of each group is retained:

$$\bar{s}_{g,t} = \max_{k:\, g_k = g} w_{k,t}\, s_{k,t}, \qquad \bar{\rho}_{g,t} = \rho_{k^*,t}, \tag{7}$$

where k^* is the maximizing clue and $w_{k,t} \in [0,1]$ down-weights distant or roadway-irrelevant observations. With $\mathcal{Q}_t$ the set of groups active at frame t and all sums running over $g \in \mathcal{Q}_t$,

$$\begin{aligned} S_t &= 1 - \prod_{g \in \mathcal{Q}_t} (1 - \bar{s}_{g,t}), \\ U_t &= \begin{cases} 1 - \dfrac{\sum_g \bar{s}_{g,t}\, \bar{\rho}_{g,t}}{\sum_g \bar{s}_{g,t}}, & \text{if } \sum_g \bar{s}_{g,t} > 0, \\ 1, & \text{otherwise.} \end{cases} \end{aligned} \tag{8}$$

Independent groups thus reinforce one another without counting the same trajectory twice, and strong but unreliable evidence keeps a high score together with high uncertainty. S_t and U_t are evidence summaries, not calibrated probabilities.

The graph is divided into overlapping temporal intervals $\{\mathcal{I}_m\}_{m=1}^{M}$, each serialized into a compact case memory

$$\mathcal{M}_m = \Phi\left(\mathcal{G}_m, \{S_t, U_t\}_{t\in\mathcal{I}_m}\right), \tag{9}$$

where $\mathcal{G}_m$ is the graph restricted to $\mathcal{I}_m$. Each memory keeps persistent actors, the ego-motion timeline, chronological events, and evidence trajectories, with global timestamps.

4.3 Hypothesis-Driven Evidence-Seeking Investigation

The agentic component verifies a traffic-violation hypothesis with the minimum decision-relevant evidence rather than producing an unconstrained reasoning trace. From the case memories $\{\mathcal{M}_m\}_{m=1}^{M}$, segment-level scouts propose competing hypotheses

$$\mathbb{H}_0 = \{h_j = (b_j, a_j, \ell_j, t_j, e_j, \eta_j)\}_{j=1}^{J}, \tag{10}$$

where b_j is the candidate behavior, a_j the actor, ℓ_j the location, t_j the decisive time, e_j the current supporting evidence, and η_j the strength assigned by the scout. Each hypothesis is examined along five clue dimensions, $\mathcal{D} = \{\textsc{Who}, \textsc{Where}, \textsc{When}, \textsc{What}, \textsc{Evidence}\}$. At investigation step n, let $\mathbb{H}_n$ be the current hypotheses, $\Omega_n(h_j) \subseteq \mathcal{D}$ the dimensions of h_j that are still missing, contradictory, or unreliable, and $\mathcal{O}_n$ the evidence gathered so far. The expert selects the next action

$$\alpha_n = \pi\big(\mathbb{H}_n,\, \mathcal{O}_n,\, \{\Omega_n(h_j)\}_{j=1}^{J}\big), \tag{11}$$

with $\alpha_n \in \{\textsc{Read}, \textsc{Ask}, \textsc{Inspect}, \textsc{Decide}\}$, executes it, and records its outcome o_n, so that $\mathcal{O}_{n+1} = \mathcal{O}_n \cup \{(\alpha_n, o_n)\}$. Read retrieves structured evidence from a case memory, Ask queries a scout to refine or contrast hypotheses, Inspect examines a decision-critical interval of the original video with a vision-language model, and Decide commits to the hypothesis whose decisive dimensions are resolved. The loop ends at Decide or after B actions, of which at most B_{pix} may be Inspect.

4.4 Verification and Law Grounding

The expert's decision is provisional. A consistency check compares the actor, location, time, behavior, and evidence of the selected hypothesis h^* against the gathered evidence $\mathcal{O}_n$ and the inspected frames; if the evidence remains insufficient, the case returns to the investigation loop within the remaining budget. Once verified, the primary behavior is mapped deterministically to the CVC section it violates,

$$\hat{c} = \mathcal{L}\big(\hat{b}^{(1)}\big), \tag{12}$$

where $\mathcal{L}$ follows the taxonomy of Sec. 3. The final report is $\hat{\mathbf{y}}$ in Eq. 3.

5 Experiments

We evaluate two capabilities: *violation recognition*, whether a method identifies the correct violation family, and *event grounding*, whether it attributes the violation to the correct road user and localizes its decisive time.

5.1 Evaluation Protocol

We compare iCluedo with direct video-language models from the Qwen3-VL Bai et al. (2025a) and Cosmos NVIDIA (2026) families, which receive the raw clip and must output a CVCV violation family. For each model we evaluated several prompting strategies and report its best-performing configuration. These baselines instantiate the fixed-context setting discussed in the introduction: the model observes a predefined set of frames once and must decide from them. Every method contributes exactly one prediction per clip; for iCluedo, this is its verified top-ranked hypothesis, so the comparison involves no additional guesses. Evaluation is performed on the matched-coverage subset of clips for which every method returns a valid prediction. We report video-weighted accuracy together with macro- and support-weighted metrics; macro metrics weight every violation family equally and therefore expose failures on less frequent families that accuracy alone conceals.

Table 2: Traffic-violation chapter classification. Per-class values report F1 scores, with class supports shown below the header. ACC denotes video-weighted accuracy, while R_{mac}, P_{mac}, and $\mathrm{F1}_{\mathrm{mac}}$ denote macro-averaged recall, precision, and F1, respectively. $\mathrm{F1}_{\mathrm{wt}}$ denotes support-weighted F1. **Bold** and underlined values indicate the best and second-best results among the reported methods, respectively.

	Per-class F1					Overall Performance				
Method	**Signal**	**Overt.**	**RoW**	**Ped.**	**Turn**	**ACC**	R	P	$F1$	$F1_{\mathrm{wt}}$
#videos	221	949	665	104	1325	–	–	–	–	–
Qwen3-VL-2B	0.310	0.015	0.257	0.146	0.008	0.148	0.326	0.455	0.148	0.087
Qwen3-VL-8B	0.017	0.073	0.107	0.364	0.074	0.051	0.089	0.563	0.128	0.087
Cosmos-Reason2	0.177	0.304	0.184	0.109	0.275	0.217	0.312	0.441	0.212	0.255
Cosmos3-Nano	**0.338**	0.401	0.235	**0.394**	0.532	0.306	0.327	**0.570**	0.382	0.414
iCluedo (ours)	0.236	**0.634**	**0.304**	0.218	**0.725**	**0.506**	**0.454**	0.462	**0.423**	**0.561**

5.2 Violation Recognition

Direct VLMs are unreliable on CVCV. Table 2 confirms the observation that motivated iCluedo. The strongest direct model, Cosmos3-Nano, reaches only 0.306 accuracy and 0.382 macro F1, and both Qwen3-VL models remain below 0.15 accuracy. Larger models do not close the gap: Qwen3-VL-8B is less accurate than Qwen3-VL-2B. Since each baseline uses its best prompting strategy, these failures point to fixed-context inference, rather than prompt design, as the limiting factor.

Investigation substantially improves recognition. iCluedo achieves the best accuracy (0.506), macro recall (0.454), macro F1 (0.423), and weighted F1 (0.561). Relative to Cosmos3-Nano, it improves accuracy by 20.0 points, macro recall by 12.7, macro F1 by 4.1, and weighted F1 by 14.7. Cosmos3-Nano attains higher macro precision (0.570 vs. 0.462) but much lower macro recall (0.327 vs. 0.454): it predicts a narrow subset of families conservatively, whereas iCluedo covers the taxonomy more broadly.

Where investigation helps. The largest gains occur for overtaking (0.634 vs. 0.401) and turning (0.725 vs. 0.532), the families whose decisions depend on how lane positions and relative motion evolve over time, which is precisely what the clue graph tracks and the investigation revisits. iCluedo also leads on right-of-way (0.304). Signal and pedestrian violations remain harder, and Cosmos3-Nano obtains the best F1 on both. A plausible cause is that small or briefly visible traffic lights and pedestrians are missed by upstream perception, so no clue prompts their inspection; improving perception coverage for these entities is a natural next step.

Multiple violations. A clip can contain more than one violation, and iCluedo retains a ranked alternative hypothesis for this reason. Scoring both retained hypotheses against the annotated violations raises accuracy from 0.506 to 0.646, recovering 28% of top-1 errors (Appendix A.4.2).

5.3 Actor Attribution and Temporal Grounding

Naming the violation is only part of the case: an event-reconstruction system must also determine *who* committed it and *when*. Table 3 evaluates iCluedo against the CVCV actor and violation-time annotations on violation clips. Actor attribution reaches a balanced accuracy of 0.550, with higher recall for ego violations (0.672) than for violations by other road users (0.428). The predicted decisive time lies within 5 s of the annotation for 63.0% of clips (median error 3.21 s), but only 20.9% fall within 1 s. Requiring the correct actor as well lowers these rates to 26.9% and 9.1%. Grounding the full case is therefore markedly harder than naming the violation, a gap that label-only benchmarks cannot reveal and that CVCV is designed to measure.

Actor attribution and temporal grounding. Beyond violation classification, we evaluate whether iCluedo identifies the relevant road user and localizes the decisive event in time. We re-

Table 3: Actor attribution and temporal grounding of iCluedo on CVCV. **Actor attribution** is binary (ego vehicle vs. other road user). Bal. ACC is the mean of the ego and other recalls. **Temporal grounding** compares the predicted decisive time with the annotated violation time on violation clips. MedAE is the median absolute error in seconds; R@ts is the fraction of clips whose predicted time lies within t seconds of the annotation. Joint@ts additionally requires the correct actor (four way label: ego / other / both / none).

	Actor Attribution			Temporal Grounding			
Method	**ACC↑**	**Ego R.↑**	**Other R.↑**	**MedAE (s)↓**	**R@1s↑**	**R@2s↑**	**R@5s↑**
iCluedo	0.550	0.672	0.428	3.21	0.209	0.366	0.630

port per class recall and balanced accuracy to account for class imbalance. Temporal grounding is evaluated on all dataset videos using the median absolute error between the predicted and reference event times, together with Recall@1s and Recall@2s, which measure the proportion of predictions localized within one and two seconds, respectively. We additionally report Joint@2s on the subset containing both annotations; a prediction is counted as correct only when the actor is correctly identified and the event time is localized within two seconds.

5.4 Actor Attribution and Temporal Grounding

Beyond identifying *which* violation occurred, an event-reconstruction system must determine *who* committed it and *when* it happened. We evaluate both capabilities against the CVCV annotations, which provide, for each clip, the responsible party and the time of the violation.

Limitations and future work. CVCV is limited by class imbalance and coverage across driving conditions. Future work will expand rare violation categories, explore novel and challenging violation scenarios, broaden coverage across environments and traffic-code jurisdictions, and add finer spatial annotations. iCluedo also relies on frozen perception modules, making improved detection of small or briefly visible entities an important direction.

6 Conclusion

We presented iCluedo, a clue-driven agentic framework that formulates traffic-violation understanding as a hypothesis-driven, sparse-evidence video-reasoning problem. Rather than processing every frame uniformly or directly predicting a violation from unstructured video, iCluedo constructs a persistent spatio-temporal case memory, maintains competing violation hypotheses, and selectively acquires the structured or visual evidence needed to distinguish among them. We also introduced CVCV, a real-world dashcam benchmark of 4.65K videos that supports three complementary levels of evaluation: violation recognition, event grounding, and evidence grounding. Experiments show that iCluedo improves traffic-violation recognition over direct VLM baselines and that retaining competing hypotheses substantially increases recovery of the correct interpretation, particularly for ambiguous right-of way and pedestrian scenarios. The actor and temporal grounding results further show that the framework associates its decisions with explicit road users and moments in the video, while also revealing remaining challenges in other-actor attribution and precise event localization. Future work will study stronger visual verification, better uncertainty calibration, and open-set reasoning over the novel and ambiguous scenarios retained in CVCV.

AI use statement

In this work, we used generative AI tools for refining the writing. We have not used generative AI tools for anything else. Additionally, we used generative AI tools for polishing writing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI

REFERENCES

Ali K AlShami, Ananya Kalita, Ryan Rabinowitz, Khang Lam, Rishabh Bezbarua, Terrance Boult, and Jugal Kalita. Coool: Challenge of out-of-label a novel benchmark for autonomous driving. *arXiv preprint arXiv:2412.05462*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2682–2690, 2020. doi: 10.1145/3394171.3413827.

Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *Computer Vision – ACCV 2016*, pp. 136–153. Springer, 2016. doi: 10.1007/978-3-319-54190-7_9.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.

Zhihong Cui, Haoran Tang, Tianyi Li, Yushuai Li, Peiyuan Guan, Amir Taherkordi, and Tor Skeie. C-trail: A commonsense world framework for trajectory planning in autonomous driving. *IEEE Transactions on Vehicular Technology*, 2026.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2025.

Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. Dada-2000: Can driving accident be predicted by driver attention? analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 4303–4309. IEEE, 2019. doi: 10.1109/ITSC.2019.8917218.

Sparsh Garg, Yi-Wen Chen, Vijay Kumar B G, and Abhishek Aich. Caviar: A causal video dataset for fine-grained accident reasoning in real-world scenarios. *arXiv preprint arXiv:2608.19380*, 2026.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. doi: 10.1109/TPAMI.2024.3444912.

Dongkwon Jin and Chang-Su Kim. OMR: Occlusion-aware memory-based refinement for video lane detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-73414-4_8.

Soo Yong Kim, Joonyoung Kim, Jaewon Lee, Seong Wook Lee, Tae-San Eom, and Jeonghun Chae. Trafficrag: Temporal grounding for traffic violations via retrieval-augmented generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 66–74, 2026. doi: 10.1109/WACVW68408.2026.00012.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Yiheng Li, Cunxin Fan, Chongjian Ge, Seth Z. Zhao, Chenran Li, Chenfeng Xu, Huaxiu Yao, Masayoshi Tomizuka, Bolei Zhou, Chen Tang, Mingyu Ding, and Wei Zhan. Womd-reasoning: A large-scale dataset for interaction reasoning in driving. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34288–34311. PMLR, 2025.

Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. Can lvlms obtain a driver's license? a benchmark towards reliable agi for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5838–5846, 2025.

Enhui Ma, Jiahuan Zhang, Guantian Zheng, Tao Tang, Shengbo Eben Li, Yuhang Lu, Xia Zhou, Xueyang Zhang, Yifei Zhan, Kun Zhan, Zhihui Hao, Xianpeng Lang, and Kaicheng Yu. Drivecombo: Benchmarking compositional traffic rule reasoning in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32113–32123, 2026.

Arnav V Malawade, Shih-Yuan Yu, Brandon Hsu, Deepan Muthirayan, Pramod P Khargonekar, and Mohammad A Al Faruque. Spatio-temporal scene-graph embedding for autonomous vehicle collision prediction. *arXiv preprint arXiv:2111.06123*, 2021.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Daniel Moura, Shizhan Zhu, and Orly Zvitia. Nexar dashcam collision prediction dataset and challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2608–2616, June 2025. doi: 10.1109/CVPRW67362.2025.00243.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Chirag Parikh, Deepti Rawat, Tathagata Ghosh, Ravi Kiran Sarvadevabhatla, et al. Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19002–19011. IEEE, 2025.

Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3521–3529, June 2018. doi: 10.1109/CVPR.2018.00371.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 16558–16569, 2021.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Learning single-image calibration with geometric optimization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-73661-2_1.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

Maolin Wei, Wanzhou Liu, and Eshed Ohn-Bar. Passing the driving knowledge test. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, 2024.

Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9878–9888, 2021.

Manyi Yao, Bingbing Zhuang, Sparsh Garg, Amit K. Roy-Chowdhury, Christian R. Shelton, Manmohan Chandraker, and Abhishek Aich. iFinder: Structured zero-shot vision-based LLM grounding for dash-cam video reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5541.

Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J. Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):444–459, 2023. doi: 10.1109/TPAMI.2022.3150763.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 1–21, 2022.

Weicheng Zheng, Xiaofei Mao, Nanfei Ye, Pengxiang Li, Kun Zhan, Xianpeng Lang, and Hang Zhao. Driveagent-r1: Advancing vlm-based autonomous driving with active perception and hybrid thinking. In *International Conference on Learning Representations*, volume 2026, pp. 125576–125610, 2026.

Xingcheng Zhou, Konstantinos Larintzakis, Hao Guo, Walter Zimmer, Mingyu Liu, Hu Cao, Jiajie Zhang, Venkatnarayanan Lakshminarasimhan, Leah Strand, and Alois Knoll. Tumtraf videoqa: Dataset and benchmark for unified spatio-temporal video understanding in traffic scenes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78786–78811. PMLR, 2025.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 474–490, 2020.

Zijin Zhou and Songan Zhang. Aitp: Traffic accident responsibility allocation via multimodal large language models. *arXiv preprint arXiv:2604.20878*, 2026.

A Appendix

A.1 Limitations and future work.

CVCV judges every clip under a single reference code, although part of the footage is recorded outside California; extending the annotations to other jurisdictions, and to cases where obligations conflict, would test whether legal grounding transfers across rule systems. iCluedo inherits the errors of its frozen perception tools, which explains much of its difficulty with signal and pedestrian violations, where the decisive entities are small or briefly visible. Its relations and reliability scores are rule-based and uncalibrated, its investigation policy is prompted rather than learned, and each decision requires several language-model calls, which precludes real-time use. Finally, the reference rationales are drafted by a vision-language model before human review and may carry its phrasing. Promising directions include using the strongest video-language models as investigation tools, calibrating clue reliability against verified outcomes, and learning when additional evidence is worth acquiring.

A.2 Perception Tools

The first table 4show the pretrained perception models that was used.

Table 4: Perception tools. All models are pretrained, publicly available, and frozen; none is trained or fine-tuned on CVCV.

Tool	**Checkpoint**	**Output**	**Training data**	**Setting**
OWL-V2	[ckpt]	2D boxes, 17 classes	WebLI pseudo-labels	conf. ≥ 0.25
ByteTrack	—	track IDs	— (association)	default
CenterTrack	[ckpt]	3D box, yaw	nuScenes	—
SAM	[ckpt]	object masks	SA-1B	—
Metric3D v2	[ckpt]	metric depth	[verify]	depth in mask
OMR	[ckpt]	lanes, lane index	—	—
GeoCalib	[ckpt]	intrinsics, distortion	—	OpenCV undistort
DROID-SLAM	[ckpt]	ego pose	TartanAir	turn $\geq 30^\circ$, g=2
InternVL	[ver.]	scene/object attributes	—	—
VideoLLaMA2	[ver.]	video description	—	—

Table 5 show the lane relation rules.

Table 5: Relation rules. d: ego–agent distance; λ: smoothed lane index; ψ: heading.

Relation	**Rule**	**Failure handling**
lane membership(i, λ)	majority lane label of i over w frames	no lane detected $\Rightarrow$ undetermined
following(i, j)	same λ, j ahead of i, $\lvert\psi_i - \psi_j\rvert < \theta_\psi$	lane undetermined $\Rightarrow$ undetermined
approaching(i, j)	$\Delta d_{ij}/\Delta t < -\delta_v$ over w frames	distance jump $> \delta_d \Rightarrow$ motion flag
crossing(i)	λ_i changes and persists $L_{\min}$ frames	flicker shorter than $L_{\min} \Rightarrow$ ignored
controls(σ, ego)	signal σ ahead of ego, $d \leq d_{\text{ctrl}}$=20 m	unknown state $\Rightarrow$ edge kept, state undetermined

A.3 Dataset Examples

Every clip in CVCV carries one *primary label* naming the violation it shows. Each violation label is grounded in a section of the California Vehicle Code (CVC), Division 11 (Rules of the Road), and belongs to one of six chapters named after CVC chapters: Traffic Signs, Signals, and Markings; Driving, Overtaking, and Passing; Right-of-Way; Pedestrians' Rights and Duties; Turning and Stopping and Turning Signals; and Special Stops Required. Two further labels mark clips with no violation and clips whose violation could not be assigned a specific label (*unknown*).

Table 6 shows ground-truth annotations for five common violations, taking the most frequent label in each of the five largest chapters. For each example we give the violated CVC section, the violator (*ego*: the camera vehicle; *other*: another road user), the annotated violation time (mm:ss.sss), and

the annotators' *Reason* and *Evidence* fields, quoted verbatim. In the quotes, "our vehicle" refers to the ego vehicle.

Table 6: Ground-truth annotations for common CVCV violations: the most frequent label in each of the five largest chapters. *Reason* and *Evidence* are quoted verbatim from the annotations.

Violation	Annotation
Running a solid red light CVC §21453(a) Violator: other Time: 00:02.652	*Reason:* "A black car ran a red light without stopping to give way." *Evidence:* "The video shows that a black car ran a red light, resulting in a collision with two black cars in the cross lane."
Driving on the wrong side of the road CVC §21650 Violator: ego Time: 00:04.017	*Reason:* "Our vehicle was driving in the opposite lane." *Evidence:* "The video shows that our vehicle entered the opposite lane, and the black vehicle traveling straight was unable to avoid in time, resulting in a collision between the two cars."
Failing to yield at an intersection CVC §21800 Violator: other Time: 00:03.170	*Reason:* "A truck at the intersection failed to yield to our vehicle, which had the right of way." *Evidence:* "The video shows that a truck, while making a right turn on a snow-covered road, failed to stop and yield to our vehicle, which had the right of way, resulting in a dangerous conflict."
Failing to yield to a pedestrian CVC §21950(a) Violator: other Time: 00:01.772	*Reason:* "A light-colored vehicle failed to yield to pedestrians at the crosswalk." *Evidence:* "The video shows that the light-colored vehicle did not slow down to yield to pedestrians crossing the street while passing through the intersection, which led to the accident."
Unsafe lane change CVC §22107 Violator: other Time: 00:01.500	*Reason:* "A light-colored semi-trailer truck suddenly changed lanes, causing a blue van to collide as it could not avoid it in time." *Evidence:* "The video shows the light-colored semi-trailer truck changing lanes abruptly, not leaving enough reaction space for the blue van on the right to avoid a collision safely."

A.4 Experiments

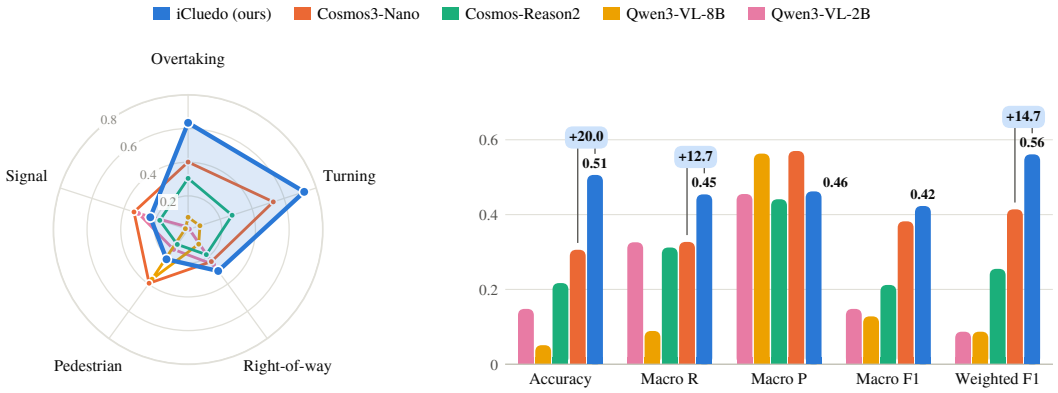
A.4.1 Results

Our clue driven argentic framework, iCluedo, Show better performance compare to other VLM as shown in Figure 4

A.4.2 Hypothesis Recovery and Multiple Violations

A single clip can contain more than one legally relevant behavior; for example, a driver may pass a cyclist unsafely while crossing a solid line. CVCV therefore records a secondary violation when one is present, and iCluedo keeps a ranked primary and alternative hypothesis throughout its investigation. Table 7 scores both. **Top-1** uses only the highest-ranked hypothesis, as in the main results; **Top-2** counts a clip as correct when either retained hypothesis matches an annotated violation. Top-2 raises accuracy from 0.506 to 0.646 and macro F1 from 0.423 to 0.558, and every family gains between 0.10 and 0.17 F1.

The gain reflects two effects: in clips with several violations, the alternative hypothesis recovers the second annotated behavior, and in other clips the correct interpretation was retained but ranked second. In total, 28% of top-1 errors are recovered under the top-2 criterion. The investigation therefore preserves the right answer among its competing hypotheses, and improving the final selection is a direct route to higher single-decision accuracy. Top-2 is a diagnostic of iCluedo and is not compared with the single-prediction baselines.



(a) Per-class F1

(b) Overall performance

Figure 4: Per-class F1 and overall performance of iCluedo compared with direct VLM baselines on CVCV. iCluedo shows stronger balanced performance across violation categories and improves key overall metrics.

Table 7: Top-1 versus top-2 traffic-violation classification performance of iCluedo. Top-1 evaluates only the highest-ranked prediction, whereas top-2 considers a prediction correct when the reference category appears among the two highest-ranked candidates. Panel (a) reports per-class F1, and Panel (b) reports overall performance. Δ denotes the absolute improvement from top-1 to top-2. ACC denotes accuracy; R_{mac}, P_{mac}, and P_{mic} denote macro recall, macro precision, and micro precision, respectively.

(a) Per-class F1

Prediction Setting	**Signal**	**Overtake**	**RoW**	**Ped.**	**Turn**
iCluedo (ours), top-1	0.236	0.634	0.304	0.218	0.725
iCluedo (ours), top-2	**0.367**	**0.746**	**0.469**	**0.382**	**0.826**
Δ (top-2 $-$ top-1)	**+0.131**	**+0.112**	**+0.165**	**+0.164**	**+0.101**

(b) Overall performance

Prediction Setting	**ACC**	R_{mac}	P_{mac}	P_{mic}	**F1**$_{\text{mac}}$	**F1**$_{\text{wt}}$	**AUC**
iCluedo (ours), top-1	0.5058	0.4540	0.4618	0.5716	0.4234	0.5612	0.6808
iCluedo (ours), top-2	**0.6463**	**0.6053**	**0.5771**	**0.7029**	**0.5581**	**0.6828**	**0.7688**
Δ (top-2 $-$ top-1)	**+0.1405**	**+0.1513**	**+0.1153**	**+0.1313**	**+0.1347**	**+0.1216**	**+0.0880**