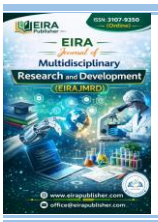

EIRA Journal of Multidisciplinary Research and Development (EIRAJMRD)
Journal homepage: https://eirapublisher.com/eirajmrd/
Email: office.eirapublisher@gmail.com
ISSN: 3107-9350 (Online)
Volume 2, Issue 4, (July-August) 2026
DOI: https://doi.org/10.5281/zenodo.21609762

Lightweight Vision-Language Incident QA for Smart-City Traffic Monitoring with Uncertainty-Calibrated Explanations

Elena Zhou

Computer Science, University of California, Irvine, Irvine, CA, USA

Article history:
Received: 17/05/2026
Accepted: 04/07/2026
Published: 26/07/2026

Keywords: *smart-city traffic monitoring; active transportation; incident question answering; uncertainty calibration; anomaly detection*

***Corresponding Author**:
Elena Zhou

Abstract **Original Research Article**

Smart-city traffic monitoring increasingly requires systems that answer operational questions about unusual road-user behavior while exposing uncertainty. This paper presents LVLIQA, a lightweight incident question-answering framework for active-transportation monitoring. Rather than retaining raw camera frames, the system consumes privacy-preserving sensor-hour slices containing location, direction, mode, count, and seasonal context. A robust seasonal residual encoder converts each slice into compact visual-state tokens; a question-answering layer maps the state to normal, traffic-volume surge, or traffic-volume drop; and isotonic calibration converts residual margins into probabilities used in concise, evidence-linked explanations. The evaluation uses complete 2025 California Department of Transportation pedestrian and bicycle hourly count streams. After duplicate resolution, 1,735,171 sensor-hour slices were divided chronologically into January–August training, September–October validation, and November–December testing. On 346,253 test slices, LVLIQA achieved 0.868 binary F1, 0.997 AUROC, 0.911 AUPRC, and 0.0024 expected calibration error. Event-type question answering reached 0.9916 exact-match accuracy and 0.919 macro-F1. A compact decision tree produced the highest binary F1 (0.927), whereas LVLIQA supplied the most direct calibrated answer-and-evidence pathway. Because event labels are defined from the same seasonal-residual family used by the detectors, these results measure reproducible, rule-consistent anomaly screening rather than independently verified crash detection. LVLIQA should therefore be interpreted as an auditable triage layer that prioritizes count anomalies for subsequent operational review.

How to cite this article: Zhou, E. (2026). Lightweight vision-language incident QA for smart-city traffic monitoring with uncertainty-calibrated explanations. EIRA Journal of Multidisciplinary Research and Development, 2(4). 27-39.

Introduction

City traffic cameras and sensors increasingly support not only counting but also explanation-oriented safety analysis. The AI City Challenge established a city-scale benchmarking tradition for traffic video analytics and later expanded toward multimodal tracking and traffic-safety description (Naphade et al., 2017; Wang et al., 2024). At the richer end of this sensing stack, object-level LiDAR–camera fusion can preserve multi-target state under clutter, illustrating what sensor-intensive monitoring can provide when raw multimodal data are available (Xin, 2026a). Vision-language models such as CLIP and BLIP align visual representations with text (Li et al., 2022; Radford et al., 2021), while VQA and GQA made grounded question answering a standard evaluation task (Antol et al., 2015; Hudson & Manning, 2019). Many transportation agencies, however, cannot retain raw video continuously because of bandwidth, privacy, and governance constraints. This study therefore asks whether compact outputs from cameras or automated counters can support auditable incident-like question answering without access to the original frames.

The core operational question is whether a monitored hour looks normal for its place, direction, mode, and time. Raw magnitude alone is insufficient: a bicycle count of five may be ordinary at a busy corridor and exceptional at a quiet detector, while a pedestrian count of zero may be expected at night but suspicious during a historically busy afternoon. LVLIQA therefore grounds each answer in a location-, direction-, and hour-of-week-specific seasonal baseline. The resulting residual is treated as a compact visual-state token that represents what the monitored scene is doing relative to its learned rhythm. The language layer renders this state as a

normal, surge, or drop answer together with the observed count, expected median, residual magnitude, and confidence.

Uncertainty calibration is central to this design. Accurate classifiers may still produce probabilities that do not match observed event frequencies (Guo et al., 2017; Pakdaman Naeini et al., 2015). Post-hoc calibration can improve probabilistic reliability, although distribution shift remains a separate concern (Hendrycks & Gimpel, 2017; Kuleshov et al., 2018). Related transport-perception research likewise treats confidence calibration as part of the fusion design rather than as a cosmetic label (Xin, 2025). LVLIQA calibrates residual margins on the validation period with isotonic regression, freezes an operating threshold selected on validation F1, and reports Brier score, negative log-likelihood, and expected calibration error alongside discrimination metrics.

The empirical contribution is a complete evaluation on the 2025 Caltrans active-transportation count files, an official open dataset that documents hourly pedestrian and bicycle traffic on selected California state-highway locations (California Department of Transportation, 2026). The cleaned data contain 1,735,171 sensor-hour slices across seven districts, 99 pedestrian sites, and 81 bicycle sites. The November–December test period contains 346,253 slices, of which 11,118 satisfy the study's data-derived anomaly rule. This imbalance makes F1, AUPRC, calibration, and workload-sensitive measures more informative than accuracy alone. The experiments compare LVLIQA with global and seasonal score rules, isolation forest, balanced logistic regression, and a compact decision tree.

Active-transportation counts support exposure estimation, facility planning, and safety analysis, but their interpretation depends on temporal sampling, site placement, and counter quality (Nordback et al., 2013, 2018). Probabilistic bike-sharing research further shows that demand changes across weather and seasonal regimes, supporting the present decision to estimate local temporal rhythm before interpreting an hourly deviation (Xin, 2026b). The method does not equate a count anomaly with a verified crash or police-reported incident. Instead, an incident-like slice is a statistically exceptional change that is answerable, explainable, and rankable for review.

The study addresses the space between heavy video-language systems and simple threshold dashboards. The former can describe rich scenes but often require frame storage, accelerated inference, and prompt control; the latter are inexpensive but rarely explain why an alert occurred or how reliable it is. Urban-mobility digital-twin research illustrates that aggregated observations can still support operational analysis when data limitations and policy assumptions are explicit (Zhang et al., 2025). LVLIQA follows the same transparency principle while retaining simple, maintainable components, a concern that is central to production machine-learning systems (Sculley et al., 2015).

The term vision-language is used here in a deliberately bounded sense. The visual input is not a frame embedding; it is the measured state of a monitored scene after the sensing layer has converted camera or counter evidence into an hourly record. The language component is an operational interface that converts this state into a direct answer and a human-readable reason. Explanation research emphasizes both faithfulness and rigorous evaluation rather than persuasive fluency alone (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016). Evidence-card work likewise suggests that confidence, rationale, and inspectable evidence should remain visibly distinct (Jin, 2025b). LVLIQA therefore emphasizes traceability from each answer to the observed count, baseline, residual, calibration curve, and fixed answer grammar.

Methodology

The input unit is a monitored sensor-hour slice $x = (\text{site}, \text{district}, \text{latitude}, \text{longitude}, \text{timestamp}, \text{mode}, \text{direction}, \text{count})$. The modes are Bicycle and Pedestrian. Exact duplicate records were resolved with a composite key containing mode, location, timestamp, direction, coordinates, district, and collection type; matching counts were summed so that a sensor-hour appears once in the QA dataset. The count_type field is used only to identify duplicate collection records and is not a model feature. Hour, day of week, month, weekend status, and cyclic time encodings were derived from the timestamp. To prevent temporal leakage, January 1–August 31, 2025 formed training, September 1–October 31 validation, and November 1–December 31 testing.

Event labels are deterministic and use training-period seasonal statistics. For each mode-location-direction-hour-of-week stream, the training median m and median absolute deviation (MAD) are computed; sparse stream-hours back off to the mode-hour-of-week median and MAD. The robust residual is $z = (c - m) / \max(1.4826\ \text{MAD}, f_{mode})$, where c is the observed count and f_{mode} is 3 for pedestrians and 1 for bicycles. A slice is labeled surge when $z \geq 3.5$ and its count exceeds the baseline and a mode-specific minimum by at least 20 pedestrians or 5 bicycles. It is labeled drop when $z \leq -2.5$, the baseline is at least 15 pedestrians or 3 bicycles, and the observed count is at most 35% of the baseline. All other slices are normal. These rules define a reproducible proxy target; they are not external incident ground truth.

The QA records are generated from the labeled slices without altering the measurements. Each slice answers two questions: Does this stream show an incident-like anomaly? and What kind of event is present? The answer set is intentionally restricted to no incident, traffic-volume surge, and traffic-volume drop, making exact-match evaluation unambiguous. A third question—Why was this answer produced?—is answered by rendering the evidence fields used by the residual encoder. Classification and explanation consistency can therefore be evaluated without introducing manually authored causal narratives.

LVLIQA has three components. First, the residual encoder converts the structured slice into visual-state tokens: log count, log baseline, clipped robust residual, absolute residual, relative change, calendar encodings, coordinates, district, mode, direction, and site. Second, the QA layer assigns a normal, surge, or drop answer from the calibrated incident probability and residual sign. Third, the explanation generator fills a fixed template with the observed count, expected median, residual, event type, and confidence. Deterministic provenance research similarly argues that generated claims should remain traceable to supporting records (Jin, 2025a). LVLIQA applies that principle without external retrieval: the template may use only fields present in the sensor slice or computed seasonal baseline.

The LVLIQA risk score combines positive and negative residual margins. The surge margin is $z - 3.5$ plus a small logarithmic excess-count term; the drop margin is $-z - 2.5$ plus a collapse-ratio term that rewards observations far below the expected median. A small peak-hour term is included for ranking. Isotonic regression fitted on validation labels maps the raw score to a probability, and the incident threshold is selected by validation F1 and then frozen. The same state representation thus yields a calibrated probability, binary decision, three-class answer, and evidence statement.

Five comparison models use the same chronological protocol. The global count percentile detector uses a mode-specific 99th-percentile score and ignores local seasonality. The seasonal mean-z detector replaces robust medians with stream-hour means and standard deviations. Isolation forest provides an unsupervised isolation baseline (Liu et al., 2008), complemented conceptually by local-density outlier detection (Breunig et al., 2000). A class-weighted logistic model tests linear separation, and a compact decision tree tests a shallow nonlinear partition (Quinlan, 1986). All thresholds and calibrators are selected on validation data, and every reported metric uses the complete test split.

The comparisons separate three sources of performance. Global percentile scoring tests whether count magnitude alone is sufficient; seasonal mean-z scoring tests local context without robust residuals; and isolation forest tests a generic anomaly mechanism. The linear and tree baselines reveal how strongly the engineered residual features encode the deterministic label rule. LVLIQA differs primarily by making the calibrated probability and evidence template first-class outputs. Consequently, comparisons against the tree must be read as rule-reproduction comparisons, not as independent validation against real-world crashes.

Evaluation spans three levels. Binary incident recognition uses precision, recall, F1, accuracy, balanced accuracy, Matthews correlation, AUROC, and AUPRC. Calibration and operational cost use Brier score, negative log-likelihood, 10-bin expected calibration error, prediction time per 1,000 slices, and serialized model size. Event QA uses exact-match accuracy and per-class F1 for normal, surge, and drop. Explanation checks record evidence-field coverage, generation failures, and accuracy among predictions with confidence at least 0.80. The last measure characterizes a routed subset; it is not assumed to exceed full-stream accuracy.

Preprocessing preserves the count-stream measurement semantics. The date field is parsed as an hourly timestamp; mode and direction remain categorical descriptors; and duplicates are aggregated only when every identifying field, including collection type, matches. Because count_type is preprocessing metadata rather than a predictive input, Table 1 distinguishes its duplicate-resolution role. Neither future observations nor balanced sampling are used to alter the test distribution, so the reported November–December metrics retain the observed operational class imbalance.

The seasonal median–MAD baseline is more resistant to transient spikes and outages than a moving average. A recent surge can pull a moving average upward, while an outage can pull it downward; either effect may make the next abnormal hour appear ordinary. The robust baseline limits this contamination for sparse bicycle streams and high-volume pedestrian streams. Backoff is used only when a location-direction-hour combination lacks training support, preserving local behavior wherever possible while allowing every validation and test slice to receive an answer.

The answer grammar is fixed before evaluation. A slice below the incident threshold answers no incident; a slice above the threshold with a positive residual answers traffic-volume surge; and a slice above the threshold with a negative residual answers traffic-volume drop. Every answer reports the same evidence fields, preventing the language layer from introducing a label or causal claim absent from the decision layer. This synchronization supports table-driven review, dashboard search, and downstream audit.

Implementation uses a fixed random state for subset selection and model fitting. The full training split is used for robust baselines, feature construction, validation calibration, and the balanced linear model. Deterministic stratified subsets are used only for expensive nonlinear baselines, but their scores are evaluated on the complete validation and test periods. Saved outputs include thresholds, predictions, calibration bins, QA samples, and all tabular summaries reported below.

Table 1. Dataset fields used to construct traffic-event slices.

Field	Role	Type
loc_id	site identifier	int64

Field	Role	Type
district	Caltrans district	int64
latitude	site coordinate	float64
longitude	site coordinate	float64
date	hour timestamp	datetime64[ns]
mode	travel mode	object
direction	observed approach or crosswalk direction	object
count	hourly count	int64
count_type	duplicate-resolution metadata	object

Table 2. Feature groups used by the residual encoder and QA layer.

Feature group	Features	Role
Count evidence	count, log_count, baseline_median, delta	Raw traffic evidence
Seasonal context	hour-of-week median, MAD, robust_z	Expected local rhythm
Calendar context	hour, day-of-week, month, weekend, cyclic encodings	Temporal generalization
Spatial context	latitude, longitude, district, loc_id	Spatial conditioning
Mode and direction	mode_code, direction_code	Mode/direction conditioning
Uncertainty outputs	calibrated probability, residual margin, confidence threshold	Risk communication

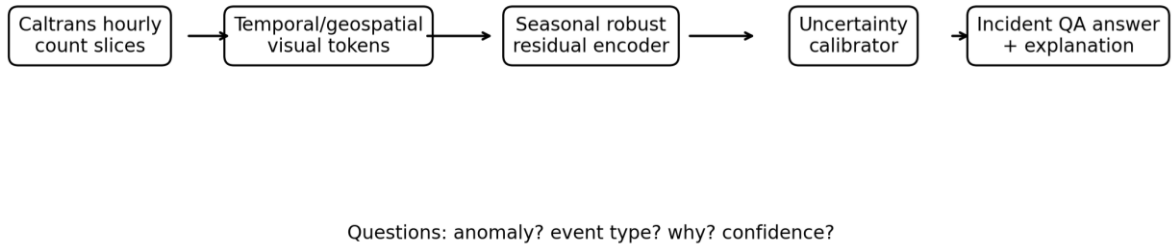

Fig. 1. LVLIQA pipeline from hourly count slices to calibrated answers and evidence.

Results and Discussion

The cleaned dataset preserves a full year of hourly monitoring. Table 3 shows pronounced mode imbalance: pedestrian slices average 16.89 counts with a 99th percentile of 334, whereas bicycle slices average 0.98 with a 99th percentile of 10. Mode-specific gates are therefore necessary; a single threshold would miss bicycle surges or over-alert on pedestrian corridors. Table 4 also shows that observed site coverage increases from 63 training sites to 65 validation sites and 69 test sites. The higher test incident rate should thus be interpreted alongside both temporal change and possible sensor onboarding.

Table 5 details the proxy-label distribution. Test bicycle slices contain 2,932 surges and 194 drops, while pedestrian slices contain 512 surges and 7,480 drops. This difference follows the empirical count structure: bicycle streams contain many zeros and low counts, making positive deviations prominent, whereas pedestrian corridors in the test months show more strong drops relative to their baselines. Table 6 adds district-level variation. These patterns identify where rule-defined deviations are frequent; they do not by themselves establish elevated safety risk.

Figures 2–4 check whether the slice representation matches the observed data. Monthly pedestrian volume rises sharply in late spring and summer and remains substantially larger

than bicycle volume; the diurnal profile peaks from late morning through early evening. The site map spans several California regions and includes overlapping mode locations. These figures support the use of month, hour, location, mode, and direction tokens, but they also expose a coverage limitation: changes in site participation may contribute to apparent shifts in total volume or event rate.

The chronological split gives the evaluation a forward-transfer interpretation. Training months establish expected behavior, validation months select thresholds and calibration, and the final two months test whether both remain useful as seasonal context changes. This is stricter than a random split because repeated hourly observations from the same sites are highly correlated. It also aligns with evidence that active-transportation demand changes across seasonal regimes (Xin, 2026b). Nevertheless, the split cannot isolate seasonality from site onboarding or unobserved weather because those covariates are absent.

The volume summaries also clarify why the task is imbalanced. Most sensor-hours are ordinary, and even a busy site can contain long low-activity periods. The incident label therefore represents an exceptional deviation from learned rhythm rather than a frequent target class. F1 and AUPRC are treated as primary comparison measures, while AUROC is reported for ranking completeness. In a stream with only a few percent positive slices, precision–recall behavior more directly reflects the size and usefulness of an alert queue.

The same numerical count has different meaning across modes and corridors. A 99th-percentile bicycle count is only 10, while pedestrian counts can be far larger at high-activity locations. The proxy definition therefore combines a residual threshold with a minimum absolute change. This prevents tiny bicycle fluctuations from being overinterpreted while retaining clear low-volume surges, and it prevents pedestrian drops from being called when the learned baseline is already near zero.

Table 3. Mode-level data coverage after exact duplicate sensor-hour aggregation.

Mode	Slices	Sites	Districts	Total count	Mean	Median	P95	P99	Max
Bicycle	853,266	81	7	837,917	0.98	0.00	4.00	10.00	236
Pedestrian	881,905	99	7	14,896,433	16.89	1.00	75.00	334.00	1,962

Table 4. Chronological train-validation-test split used for all experiments.

Split	Date range	Slices	Sites	Total count	Incident rate
train	2025-01-01 to 2025-08-31	1,087,719	63	10,617,890	0.56%
validation	2025-09-01 to 2025-10-31	301,199	65	2,992,207	1.63%
test	2025-11-01 to 2025-12-31	346,253	69	2,124,253	3.21%

Table 5. Data-derived event-label distribution by split and mode.

Split	Mode	Normal	Surge	Drop	Incident %
test	Bicycle	172,491	2,932	194	1.78%
test	Pedestrian	162,644	512	7,480	4.68%
train	Bicycle	519,444	2,332	225	0.49%
train	Pedestrian	562,167	2,164	1,387	0.63%
validation	Bicycle	152,866	2,733	49	1.79%
validation	Pedestrian	143,419	628	1,504	1.46%

Table 6. District-level coverage and incident-rate summary.

District	Mode	Slices	Sites	Total count	Incidents	Incident rate
3	Bicycle	9,463	4	2,337	26	0.27%
3	Pedestrian	99,192	19	68,579	294	0.30%
4	Bicycle	331,086	13	165,901	1,054	0.32%
4	Pedestrian	296,699	16	599,973	1,007	0.34%

District	Mode	Slices	Sites	Total count	Incidents	Incident rate
5	Bicycle	46,806	6	250,568	486	1.04%
5	Pedestrian	36,692	6	300,963	396	1.08%
6	Bicycle	144,408	5	51,659	221	0.15%
6	Pedestrian	122,096	5	131,740	198	0.16%
8	Bicycle	35,020	1	45,314	406	1.16%
8	Pedestrian	35,017	1	128,851	61	0.17%
11	Bicycle	768	1	9	0	0.00%
11	Pedestrian	1,152	1	11	0	0.00%
12	Bicycle	285,715	51	322,129	6,272	2.20%
12	Pedestrian	291,057	51	13,666,316	11,719	4.03%

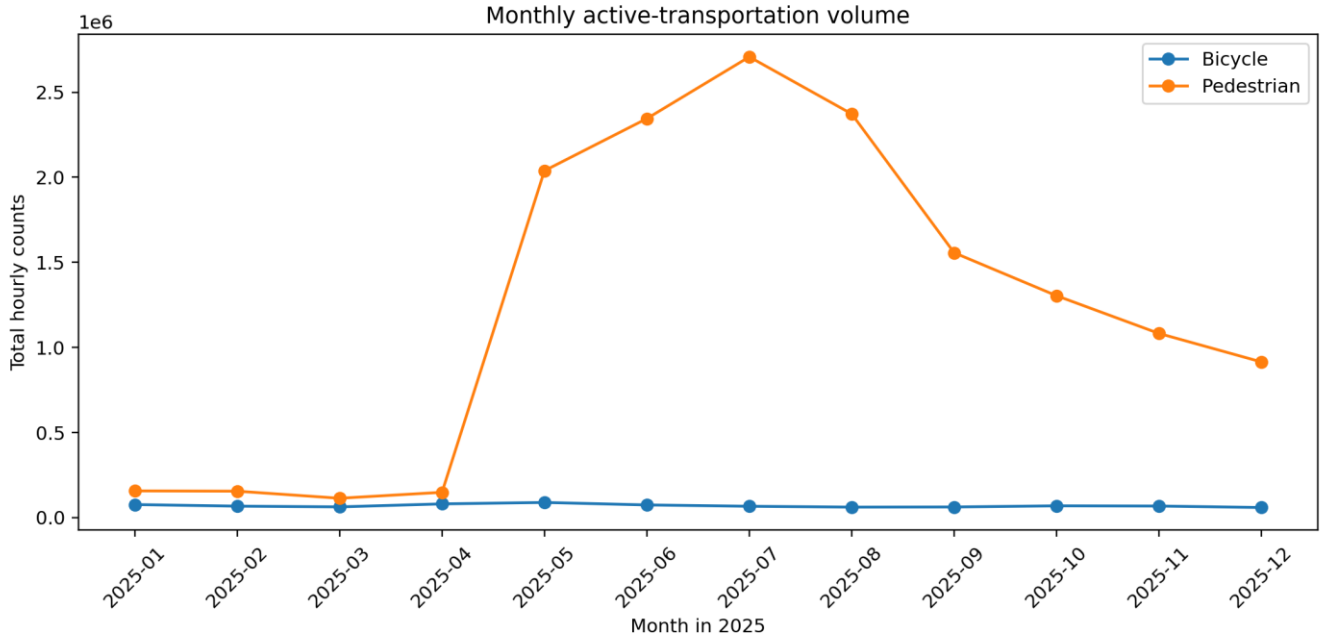

Fig. 2. Monthly pedestrian and bicycle volumes in the cleaned 2025 data.

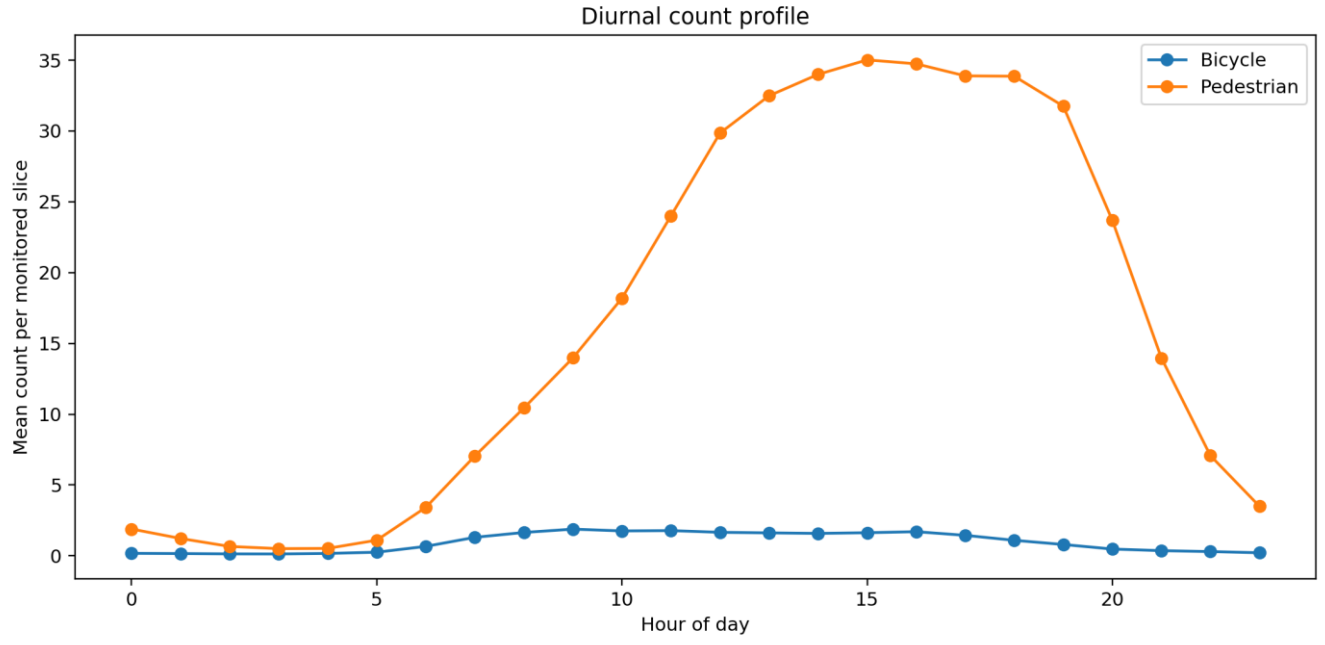

Fig. 3. Mean hourly count profile by mode.

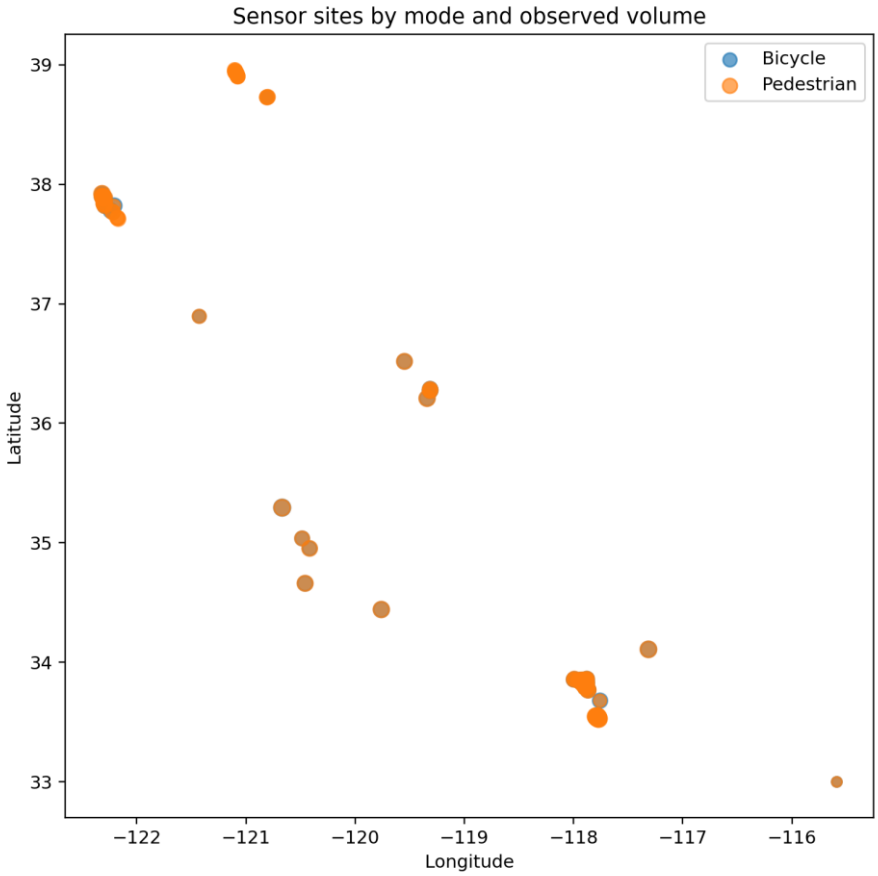

Fig. 4. Sensor-site coordinates by mode, with point size proportional to observed volume.

Table 7 reports binary proxy-label recognition. The compact decision tree obtains the highest F1 (0.927), with 0.991 precision and 0.871 recall. LVLIQA reaches 0.868 F1, 0.876 precision, 0.859 recall, 0.997 AUROC, and 0.911 AUPRC. The balanced logistic model attains 0.992 recall but only 0.675 precision, while generic and global-score baselines are weaker, especially in AUPRC. The tree's strong result is expected because both its inputs and the labels depend on the engineered residual family. These numbers demonstrate faithful reproduction of the declared anomaly rule, not independent accuracy on verified incidents.

The tree result does not eliminate the distinct interface role of LVLIQA. A tree produces a compact partition whose explanation must be reconstructed from feature paths; LVLIQA directly returns the evidence statement required by the QA task, and its confidence is tied to a monotone residual score calibrated on validation data. In a dashboard, the tree could serve as a detector cross-check while LVLIQA supplies the answer and inspectable evidence. This division should be validated with operators before deployment rather than inferred from automated metrics alone.

Table 8 compares calibration and operational cost. The decision tree has the smallest Brier score and ECE in this run; LVLIQA remains close, with Brier 0.0068 and ECE 0.0024. LVLIQA serializes to 1.3 KB and predicts 1,000 slices in approximately 0.045 ms in the recorded environment. Isolation forest is much larger and slower. Figure 6 shows validation-calibrated confidence tracking observed proxy-label rates across bins, and Figure 7 shows stronger precision–recall ranking than the seasonal mean-z baseline. These calibration results apply to the proxy target and the stated time split, not to real crash probabilities.

Table 9 and Figure 5 show that most event-type errors occur at the normal–drop boundary. LVLIQA correctly answers 333,783 of 335,135 normal slices, 3,303 of 3,444 surge slices, and 6,250 of 7,674 drop slices. Surge recall is 0.959 and drop recall is 0.814; overall exact-match accuracy is 0.9916 and macro-F1 is 0.919. Macro-F1 is the more informative class-balanced summary. Surge and drop are not confused with each other because the residual sign determines their labels; errors are instead missed or over-called events near the incident threshold.

Table 10 reports a residual-score ablation family rather than the complete LVLIQA configuration in Table 7. Within that restricted family, calibration yields F1 0.787 and ECE 0.0066, whereas removing calibration yields F1 0.779 and ECE 0.110. Removing the seasonal baseline collapses F1 to 0.345; using only the surge branch yields 0.455; and removing mode gates yields 0.632. Renaming the first row as residual-only score plus calibration makes its difference from the full Table 7 model explicit. The ablation supports the importance of local context, two-sided residual reasoning, mode-aware gates, and calibration without implying that its first row reproduces the full system.

Table 11 summarizes structural explanation checks. Every template contains numeric evidence, mode–time–direction context, and confidence, and no template-generation failures occurred. These 100% coverage values are guarantees of deterministic field completion, not evidence that people find the explanations understandable or trustworthy. At confidence at least 0.80, selective accuracy is 0.934 over 2.14% of test slices. Because this value is below full-stream exact-match accuracy, it should be treated as a description of a small routed subset rather than an accuracy improvement.

The model ranking shows why multiple metrics are necessary. The global percentile detector has high nominal accuracy because normal slices dominate, but its AUPRC of 0.177 indicates weak positive-class ranking. Seasonal mean-z raises AUROC to 0.961 but reaches only 0.375 F1, while isolation forest reaches 0.450 F1 at substantially greater storage and prediction cost. These comparisons support a simple traffic-specific residual mechanism for this proxy task, while leaving open whether the same ranking would hold against independently labeled operational incidents.

The linear and tree baselines confirm that the engineered tokens contain strong rule-aligned signal. Logistic regression captures nearly every positive slice but produces many false positives; the compact tree closely approximates the deterministic threshold structure and therefore attains high precision. This is also the central validity caution: features and labels share residual ingredients. The comparison quantifies implementation fidelity and interface trade-offs, but it cannot establish general predictive superiority without an external event label source.

The confusion pattern follows the proxy definition. Surge and drop are never confused because the sign of the residual separates them. Errors occur when a slice is near the decision threshold or when calibration suppresses a residual that satisfies only part of the rule. Drop recall is lower when the expected median satisfies the minimum-baseline condition but the observed count is not far enough below expectation after calibration. This mechanistic explanation should not be extended to causal statements about road users or sensor failures.

Calibration is operationally useful because a dashboard can sort alerts by a common probability-like score rather than by incomparable raw residuals. The low expected calibration error indicates alignment between predicted confidence bins and observed proxy-label frequencies on this test period. It does not demonstrate stability under later seasonal, weather, or coverage shifts. Threshold selection should therefore be treated as an adjustable workload policy and re-audited when the stream changes.

From a deployment perspective, latency and model size are as relevant as discrimination. The calibrated residual model can be embedded in a reporting pipeline using fields already present in the count stream, reducing integration risk relative to systems that require frame extraction, dense trajectories, or large multimodal encoders for every hour. The fallback baseline also lets a new or sparse site receive an answer, although such answers should be flagged because location-specific evidence is weaker.

Table 7. Binary incident recognition on the complete November-December test split.

Model	Precision	Recall	F1	Accuracy	AUROC	AUPRC
Global count percentile	0.396	0.305	0.345	0.963	0.772	0.177
Seasonal mean-z	0.504	0.299	0.375	0.968	0.961	0.362
Isolation forest	0.495	0.412	0.450	0.968	0.933	0.453
Balanced linear logistic	0.675	0.992	0.804	0.984	0.989	0.672
Compact decision tree	0.991	0.871	0.927	0.996	1.000	0.982
LVLIQA-calibrated (ours)	0.876	0.859	0.868	0.992	0.997	0.911

Table 8. Calibration and efficiency on the complete test split.

Model	Brier	ECE-10	NLL	ms/1k slices	Size KB
Global count percentile	0.0282	0.0197	0.1371	0.043	0.9
Seasonal mean-z	0.0245	0.0159	0.0816	0.174	1.3
Isolation forest	0.0227	0.0074	0.0862	1.359	12,860.9
Balanced linear logistic	0.0112	0.0056	0.0332	0.503	2.2
Compact decision tree	0.0025	0.0016	0.0077	0.017	5.2
LVLIQA-calibrated (ours)	0.0068	0.0024	0.0224	0.045	1.3

Table 9. Event-type incident QA performance for LVLIQA.

Class	Support	Precision	Recall	F1	Exact match	Weighted F1
normal	335,135	0.995	0.996	0.996		
surge	3,444	0.885	0.959	0.920		
drop	7,674	0.872	0.814	0.842		
overall	346,253			0.919	0.9916	0.991

Table 10. Residual-score ablations; the first row is not the complete LVLIQA model in Table 7.

Variant	Precision	Recall	F1	AUPRC	ECE-10
Residual-only score + calibration	0.860	0.725	0.787	0.864	0.007
No calibration	0.860	0.712	0.779	0.869	0.110
No seasonal baseline	0.396	0.305	0.345	0.177	0.020
Surge-only branch	0.877	0.307	0.455	0.319	0.017
No mode gates	0.887	0.491	0.632	0.760	0.011

Table 11. Explanation generation and high-confidence selective accuracy.

Measure	Value
Test slices	346,253
Numeric evidence coverage	1.000
Mode-time-direction coverage	1.000
Confidence coverage	1.000
Generation failures	0
Selective accuracy at confidence >= 0.80	0.934
Coverage at confidence >= 0.80	0.021

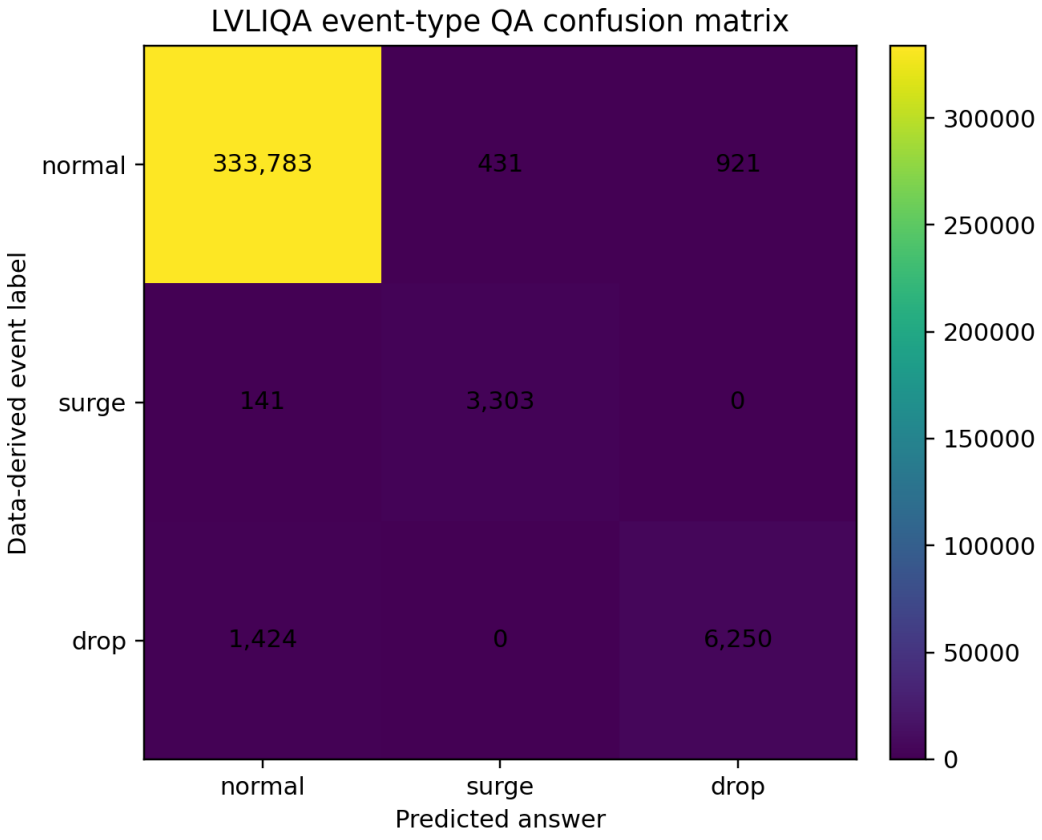

Fig. 5. Event-type QA confusion matrix for LVLIQA on the complete test split.

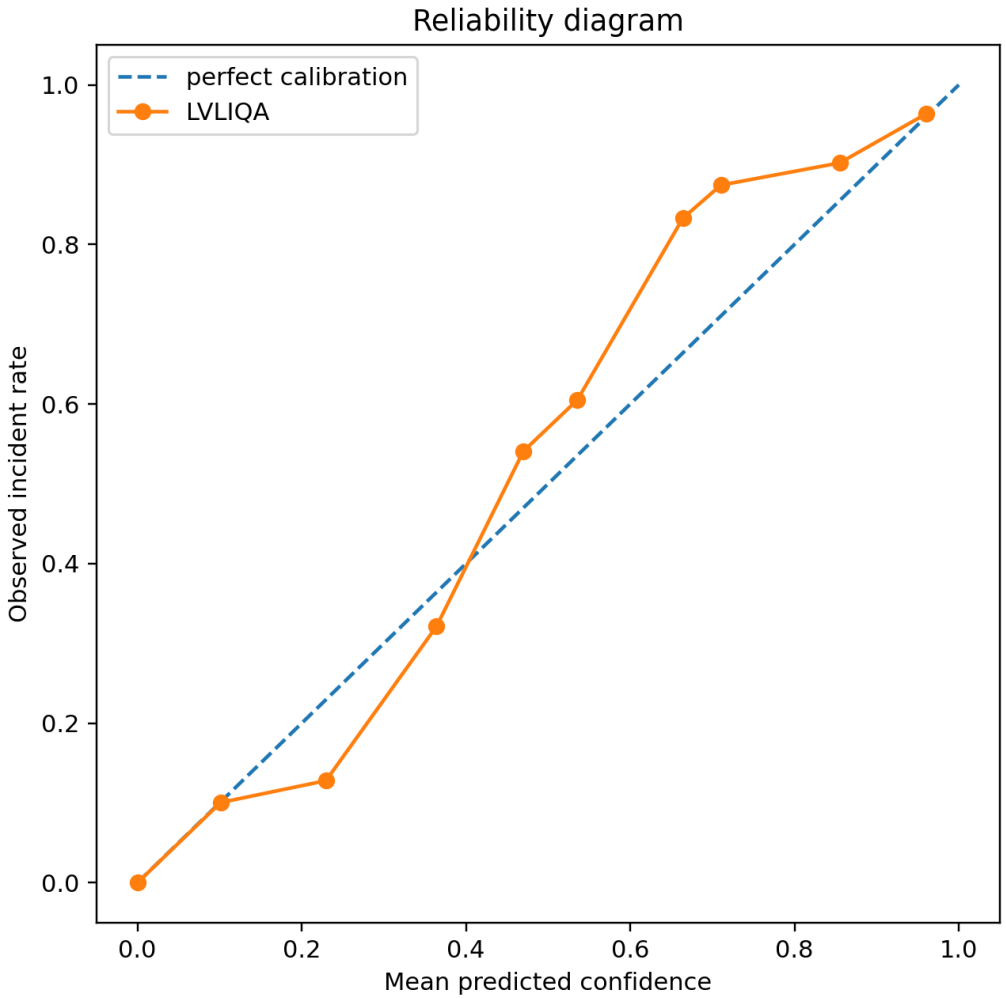

Fig. 6. Reliability diagram for LVLIQA calibrated incident probabilities.

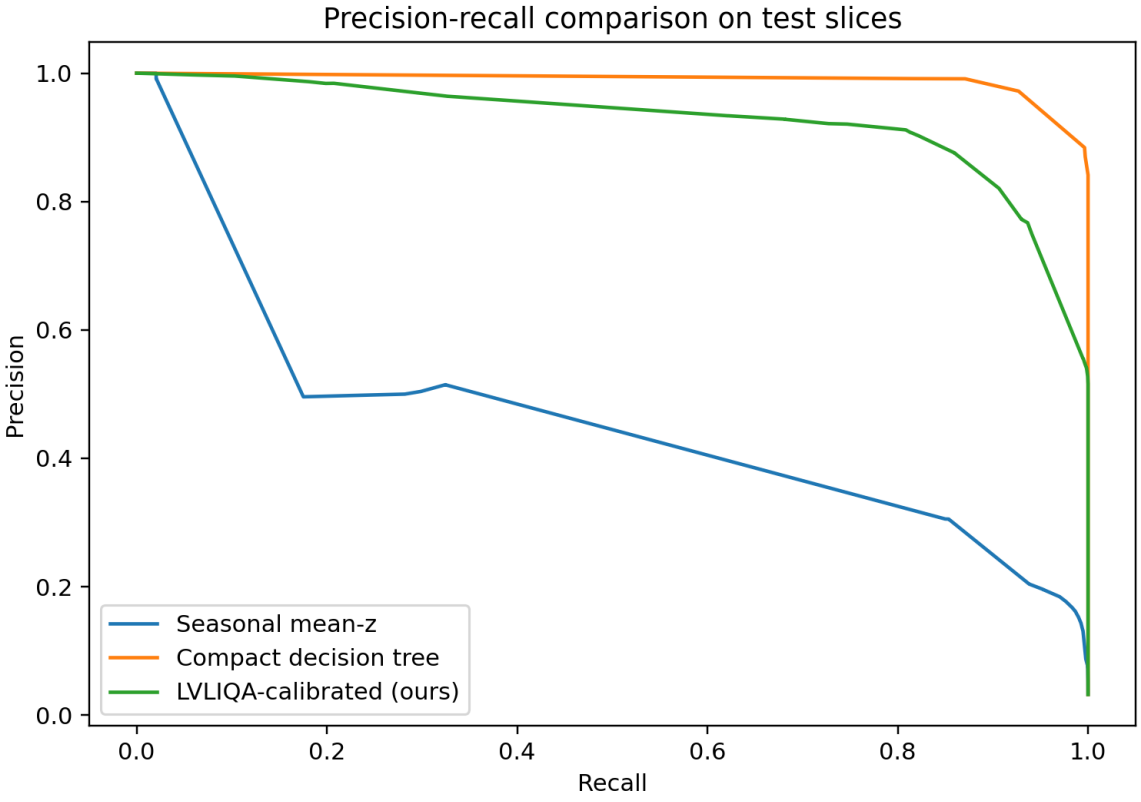

Fig. 7. Precision-recall comparison for representative models.

Limitations and practical implications. The study evaluates incident-like traffic-volume anomalies, not verified crash reports. A surge or drop is a statistically exceptional count event relative to a location-direction-hour baseline; it may reflect unusual flow, weather, a closure, a special event, or a detector problem. LVLIQA is therefore a triage and QA layer for monitoring streams, not a final incident-adjudication system.

The hourly count streams do not contain frame-level appearance, trajectories, gaze, or fine-grained pedestrian–vehicle interaction. The visual component is a count-derived state representation. This supports lightweight, privacy-preserving deployment but cannot replace video understanding when lane maneuvers, occlusion, or intent are required. Richer sensing could be attached to the same uncertainty and evidence interface in future work.

The most important internal-validity limitation is target construction. The deterministic labels and the evaluated models share seasonal-residual ingredients, so high test scores partly measure agreement with the authored rule. No independently adjudicated incidents, human-rated explanations, confidence intervals, or repeated temporal splits are available. The reported metrics should not be interpreted as evidence of crash detection, causal understanding, or user trust.

Calibration can drift when travel patterns change. The adjacent validation period exposes the calibrator to recent behavior before November–December scoring, but a long-running deployment would need rolling reliability checks and threshold review. School calendars, tourism, construction, and weather can alter both residual magnitude and operational importance.

Coverage also changes across the split: 63 sites appear in training, 65 in validation, and 69 in testing, while the proxy incident rate rises from 0.56% to 3.21%. This forward evaluation is realistic, but it mixes temporal shift with possible sensor onboarding. Future studies should report fixed-site sensitivity analyses and stratified results for new versus established sites.

Count-derived explanations do not replace human judgment. Each answer is precise about the measurement—observed count, expected median, residual, and calibrated confidence—but does not infer cause, intent, fault, or injury. Likewise, zero generation failures and complete field coverage follow from the template design; user comprehension and decision quality require separate studies. Operators should combine the answer with maintenance notes, weather, field observations, or video review.

Finally, the answer vocabulary is intentionally narrow. It describes abnormal volume behavior rather than the full traffic scene. This restriction improves measurement alignment and auditability but limits semantic coverage. Broader causal or interaction questions require richer visual evidence and independently validated labels.

Conclusion

This paper reorganized active-transportation monitoring as an auditable question-answering task over compact sensor-hour slices. LVLIQA compares each slice with a robust seasonal baseline, calibrates its incident score on a validation period, and renders a fixed answer with numeric evidence. On 346,253 November–December 2025 Caltrans slices, it achieved 0.868 binary F1, 0.997 AUROC, 0.911 AUPRC, 0.0024 expected calibration error, 0.9916 event-type exact-match accuracy, and 0.919 macro-F1. A compact decision tree reached the best binary F1, whereas LVLIQA provided the more direct answer–evidence pathway.

The principal empirical lesson is that local seasonal context, two-sided surge/drop reasoning, mode-aware gates, and calibration are all important for reproducing the declared proxy target. The same result also limits the claim: because labels and features share residual structure, the experiment demonstrates rule-consistent screening rather than independent incident prediction. Calibration describes confidence in that proxy task and must be rechecked under temporal or coverage shift.

The broader contribution is a modular deployment pattern. Rich video or multimodal perception can be reserved for selected clips, while count-derived QA continuously screens the full stream using small, privacy-preserving inputs. Keeping confidence, evidence, and answer text synchronized makes the screening layer inspectable and suitable for human triage, provided that its narrow semantics are clearly communicated.

Future work should add weather, work-zone, event-calendar, maintenance, and manually adjudicated incident outcomes; compare fixed-site and newly onboarded sensors; quantify uncertainty over repeated time splits; and evaluate explanation comprehension with operators. These extensions would test whether the current residual tokens and evidence grammar generalize beyond rule reproduction to reliable operational decision support.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2425–2433). https://doi.org/10.1109/ICCV.2015.279
2. Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (pp. 93–104). https://doi.org/10.1145/335191.335388

3. California Department of Transportation. (2026). *Active transportation count dataset* [Data set]. California Open Data. Retrieved July 23, 2026, from https://data.ca.gov/dataset/at-count-dataset

4. Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1702.08608

5. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

6. Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*. https://openreview.net/forum?id=Hkg4TI9xl

7. Hudson, D. A., & Manning, C. D. (2019). GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6700–6709). https://doi.org/10.1109/CVPR.2019.00686

8. Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

9. Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

10. Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2796–2804). PMLR. https://proceedings.mlr.press/v80/kuleshov18a.html

11. Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 12888–12900). PMLR. https://proceedings.mlr.press/v162/li22n.html

12. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 2008 IEEE International Conference on Data Mining* (pp. 413–422). https://doi.org/10.1109/ICDM.2008.17

13. Naphade, M., Anastasiu, D. C., Sharma, A., Jagrlamudi, V., Jeon, H., Liu, K., Chang, M.-C., Lyu, S., & Gao, J. Z. (2017). The NVIDIA AI City Challenge. In *2017 IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI* (pp. 1–6). https://doi.org/10.1109/UIC-ATC.2017.8397673

14. Nordback, K., Marshall, W. E., Janson, B. N., & Stolz, E. (2013). Estimating annual average daily bicyclists: Error and accuracy. *Transportation Research Record, 2339*(1), 90–97. https://doi.org/10.3141/2339-10

15. Nordback, K., O'Brien, S., & Blank, K. (2018). *Bicycle and pedestrian count programs: Summary of practice and key resources*. Pedestrian and Bicycle Information Center. https://www.pedbikeinfo.org/downloads/PBIC_Infobrief_Counting.pdf

16. Pakdaman Naeini, M., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence, 29*(1), 2901–2907. https://doi.org/10.1609/aaai.v29i1.9602

17. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1007/BF00116251

18. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR. https://proceedings.mlr.press/v139/radford21a.html

19. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

20. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems, 28*, 2503–2511. https://proceedings.neurips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems

21. Wang, S., Anastasiu, D. C., Tang, Z., Chang, M.-C., Yao, Y., Zheng, L., Rahman, M. S., Arya, M. S., Sharma, A., Chakraborty, P., Prajapati, S., Kong, Q.,

Kobori, N., Gochoo, M., Otgonbold, M.-E., Alnajjar, F., Batnasan, G., Chen, P.-Y., Hsieh, J.-W., … Chellappa, R. (2024). The 8th AI City Challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 7261–7272). https://doi.org/10.1109/CVPRW63382.2024.00722

22. Xin, Q. (2025). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

23. Xin, Q. (2026a). LiDAR–camera object-level fusion for multi-target tracking using JPDA and EKF: A reproducible empirical study on a PandaSet-parameterised five-sequence dataset. *Journal of Technology Informatics and Engineering, 5*(1), 54–76. https://doi.org/10.51903/jtie.v5i1.486

24. Xin, Q. (2026b). Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models. *Findings*. https://doi.org/10.32866/001c.157499

25. Zhang, L., Ma, R., & Greg, P. (2025). Digital-twin dispatching for urban mobility via spatio-temporal transformers and offline reinforcement learning. *Journal of Technology Informatics and Engineering, 4*(2), 337–363. https://doi.org/10.51903/jtie.v4i2.501